Introduction to Mathematics for AI

Bayesian Estimation

Andres Mendez-Vazquez

May 28, 2020

Outline

Cinvestav

Outline

Cinvestav

The Basis of Bayesian Inference

A Basic setup

- Let $f(x|\theta)$ be a conditional distribution for X given the unknown parameter θ.

For the observed data, $X = x$, the function $\ell(\theta) = f(x|\theta)$

- It is called the likelihood function!!!

The name likelihood implies that, given x, the value of θ

- It is more likely to be the true parameter than θ', if

$$f(x|\theta) > f(x|\theta')$$

The Basis of Bayesian Inference

A Basic setup

- Let $f(x|\theta)$ be a conditional distribution for X given the unknown parameter θ.

For the observed data, $X = x$, the function $\ell(\theta) = f(x|\theta)$

- It is called the likelihood function!!!

The name likelihood implies that, given x, the value of θ

- It is more likely to be the true parameter than θ', if

$$f(x|\theta) > f(x|\theta')$$

The Basis of Bayesian Inference

A Basic setup

- Let $f(x|\theta)$ be a conditional distribution for X given the unknown parameter θ.

For the observed data, $X = x$, the function $\ell(\theta) = f(x|\theta)$

- It is called the likelihood function!!!

The name likelihood implies that, given x, the value of θ

- It is more likely to be the true parameter than θ', if

$$f\left(x|\theta\right) > f\left(x|\theta'\right)$$

Basically

We are talking about optimization functions

- Where optimal's are being looked upon...

Definition

- An optimal solution to an optimization problems is the feasible solution with the largest objective function value (for a maximization problem).

Basically

We are talking about optimization functions

- Where optimal's are being looked upon...

Definition

- An optimal solution to an optimization problems is the feasible solution with the largest objective function value (for a maximization problem).

Likelihood Principle

Remarks

- In the inference about θ, after x is observed, **all relevant experimental information is contained in the likelihood function** for the observed x.

There is an interesting example quoted by Lindley and Phillips in 1976 [1]

- Originally by Leonard Savage

Leonard Savage

- Leonard Jimmie Savage (born Leonard Ogashevitz; 20 November 1917 – 1 November 1971) was an American mathematician and statistician.
  - Economist Milton Friedman said Savage was "one of the few people I have met whom I would unhesitatingly call a genius.

Likelihood Principle

Remarks

- In the inference about θ, after x is observed, **all relevant experimental information is contained in the likelihood function** for the observed x.

There is an interesting example quoted by Lindley and Phillips in 1976 [1]

- Originally by Leonard Savage

Leonard Savage

- Leonard Jimmie Savage (born Leonard Ogashevitz; 20 November 1917 – 1 November 1971) was an American mathematician and statistician.
  - Economist Milton Friedman said Savage was "one of the few people I have met whom I would unhesitatingly call a genius.

Likelihood Principle

Remarks

- In the inference about θ, after x is observed, **all relevant experimental information is contained in the likelihood function** for the observed x.

There is an interesting example quoted by Lindley and Phillips in 1976 [1]

- Originally by Leonard Savage

Leonard Savage

- Leonard Jimmie Savage (born Leonard Ogashevitz; 20 November 1917 – 1 November 1971) was an American mathematician and statistician.
  - Economist Milton Friedman said Savage was "one of the few people I have met whom I would unhesitatingly call a genius.

History

Something Notable

- The likelihood principle was first identified by that name in print in 1962 (Barnard et al., Birnbaum, and Savage et al.),

However Fisher

- It was already using a version of it in 1920's.

However, versions of it can be tracked to

- To the mid-1700s
  - It seems to have become a commonplace among natural philosophers that problems of observational error were susceptible to mathematical description.

History

Something Notable

- The likelihood principle was first identified by that name in print in 1962 (Barnard et al., Birnbaum, and Savage et al.),

However Fisher

- It was already using a version of it in 1920's.

However, versions of it can be tracked to

- To the mid-1700s
  - It seems to have become a commonplace among natural philosophers that problems of observational error were susceptible to mathematical description.

Cinvestav

History

Something Notable

- The likelihood principle was first identified by that name in print in 1962 (Barnard et al., Birnbaum, and Savage et al.),

However Fisher

- It was already using a version of it in 1920's.

However, versions of it can be tracked to

- To the mid-1700s
  - It seems to have become a commonplace among natural philosophers that problems of observational error were susceptible to mathematical description.

Cinvestav

Outline

Cinvestav

Testing Fairness

Basic Setup

- Suppose we are interested in testing θ, the unknown probability of heads for possibly biased coin.

Suppose the following Hypothesis

$$H_0 : \theta = 1/2 \text{ v.s. } H_1 : \theta > 1/2$$

Then

- An experiment is conducted and 9 heads and 3 tails are observed.
  - Not enough information to fully specify $f(x|\theta)$

Testing Fairness

Basic Setup

- Suppose we are interested in testing θ, the unknown probability of heads for possibly biased coin.

Suppose the following Hypothesis

$$H_0 : \theta = 1/2 \text{ v.s. } H_1 : \theta > 1/2$$

Then

- An experiment is conducted and 9 heads and 3 tails are observed.
  - Not enough information to fully specify $f(x|\theta)$

Cinvestav

Testing Fairness

Basic Setup

- Suppose we are interested in testing θ, the unknown probability of heads for possibly biased coin.

Suppose the following Hypothesis

$$H_0 : \theta = 1/2 \text{ v.s. } H_1 : \theta > 1/2$$

Then

- An experiment is conducted and 9 heads and 3 tails are observed.
  - Not enough information to fully specify $f(x|\theta)$

Scenario 1

Based on rashomonian analysis

- The classic Akira Kurosawa film Rashomon has become a shorthand for the lie of objective truth—what you see, basically, depends on where you stand.

Number of flips, n = 12 is predetermined

- Then number of heads X is binomial $\mathcal{B}(n, \theta)$, with probability mass function:

$$P_\theta(X = x) = f(x|\theta) = \begin{pmatrix} n \\ x \end{pmatrix} \theta^x (1-\theta)^{n-x}$$

Scenario 1

Based on rashomonian analysis

- The classic Akira Kurosawa film Rashomon has become a shorthand for the lie of objective truth—what you see, basically, depends on where you stand.

Number of flips, n = 12 is predetermined

- Then number of heads X is binomial $\mathcal{B}(n, \theta)$, with probability mass function:

$$P_\theta\left(X = x\right) = f\left(x|\theta\right) = \left(\begin{array}{c} n \\ x \end{array}\right) \theta^x \left(1 - \theta\right)^{n-x}$$

Therefore

We have

$$P_\theta\left(X = x\right) = \left(\begin{array}{c} 12 \\ 9 \end{array}\right) \theta^9 \left(1-\theta\right)^3$$

Thus

- We can use the $p - value$ for testing the hypothesis.

Therefore

We have

$$P_\theta\left(X = x\right) = \left(\begin{array}{c} 12 \\ 9 \end{array}\right) \theta^9 \left(1 - \theta\right)^3$$

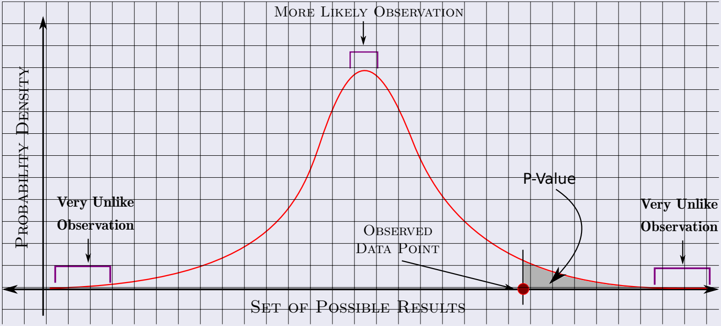
Thus

- We can use the $p - value$ for testing the hypothesis.

Then if we use the following $p - value$ analysis

Definition [2, 3]

- The p-value is defined as the probability, under the null hypothesis H_0 about the unknown distribution F of the random variable X.



Cinvestav

Therefore

For a frequentist, the $p - value$ of the test is

$$P\left(X \geq 9 | H_0\right) = \sum_{x=9}^{12} \begin{pmatrix} 12 \\ x \end{pmatrix} \left(\frac{1}{2}\right)^{x} \left(\frac{1}{2}\right)^{12-x} = 0.073$$

Given an $\alpha = 0.05$

- Then, H_0 is not rejected...

Therefore

For a frequentist, the $p - value$ of the test is

$$P\left(X \geq 9|H_0\right) = \sum_{x=9}^{12} \binom{12}{x} \left(\frac{1}{2}\right)^{x} \left(\frac{1}{2}\right)^{12-x} = 0.073$$

Given an $\alpha = 0.05$

- Then, H_0 is not rejected...

Cinvestav

Scenario 2

Number of tails (successes) 3 is predetermined

- i.e, the flipping is continued until 3 tails are observed.

Then you have a Negative Binomial with r the number of failures

$$f(x|\theta) = \begin{pmatrix} k+r-1 \\ k-1 \end{pmatrix} (1-\theta)^k \theta^r$$

Thus, we have

$$f(x|\theta) = \begin{pmatrix} 3+9-1 \\ 3-1 \end{pmatrix} (1-\theta)^3 \theta^9 = 55(1-\theta)^3 \theta^9$$

Cinvestav

Scenario 2

Number of tails (successes) 3 is predetermined

- i.e, the flipping is continued until 3 tails are observed.

Then you have a Negative Binomial with r the number of failures

$$f\left(x|\theta\right) = \begin{pmatrix} k+r-1 \\ k-1 \end{pmatrix} \left(1-\theta\right)^{k} \theta^{r}$$

Thus, we have

$$f\left(x|\theta\right) = \begin{pmatrix} 3+9-1 \\ 3-1 \end{pmatrix} \left(1-\theta\right)^{3} \theta^{9} = 55\left(1-\theta\right)^{3} \theta^{9}$$

Scenario 2

Number of tails (successes) 3 is predetermined

- i.e, the flipping is continued until 3 tails are observed.

Then you have a Negative Binomial with r the number of failures

$$f(x|\theta) = \binom{k+r-1}{k-1}(1-\theta)^k \theta^r$$

Thus, we have

$$f(x|\theta) = \binom{3+9-1}{3-1}(1-\theta)^3 \theta^9 = 55(1-\theta)^3 \theta^9$$

Cinvestav

In a similar way

We have

$$P\left(X \geq 9|H_0\right) = \sum_{x=9}^{\infty} \left(\begin{array}{c} 3+x-1 \\ 3-1 \end{array} \right) \left(\frac{1}{2}\right)^{x} \left(\frac{1}{2}\right)^{3} = 0.0327$$

Thus, the hypothesis H_0 is rejected

- But this change in decision is not caused by observations.

However, all relevant information is in the likelihood!!!

$$\ell\left(\theta\right) \propto \theta^{9}\left(1-\theta\right)^{3}$$

In a similar way

We have

$$P\left(X \geq 9 | H_0\right) = \sum_{x=9}^{\infty} \left(\begin{array}{c} 3+x-1 \\ 3-1 \end{array} \right) \left(\frac{1}{2}\right)^{x} \left(\frac{1}{2}\right)^{3} = 0.0327$$

Thus, the hypothesis H_0 is rejected

- But this change in decision is not caused by observations.

However, all relevant information is in the likelihood!!!

$$\ell\left(\theta\right) \propto \theta^{9}\left(1-\theta\right)^{3}$$

In a similar way

We have

$$P(X \geq 9|H_0) = \sum_{x=9}^{\infty} \begin{pmatrix} 3+x-1 \\ 3-1 \end{pmatrix} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^3 = 0.0327$$

Thus, the hypothesis H_0 is rejected

- But this change in decision is not caused by observations.

However, all relevant information is in the likelihood!!!

$$\ell(\theta) \propto \theta^9 (1-\theta)^3$$

Cinvestav

Remark

Edwards, Lindman, and Savage remarked

- The likelihood principle emphasized in Bayesian statistics implies, among other things, that the rules governing when data collection stops are irrelevant to data interpretation.

Therefore

- It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience.

Cinvestav

Remark

Edwards, Lindman, and Savage remarked

- The likelihood principle emphasized in Bayesian statistics implies, among other things, that the rules governing when data collection stops are irrelevant to data interpretation.

Therefore

- It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience.

Thus

Likelihood Principle [4]

- The Likelihood principle (LP) asserts that for inference on an unknown quantity θ, all of the evidence from any observation $X = x$ with distribution $X \sim f(x|\theta)$ lies in the likelihood function

$$L(\theta|x) \propto f(x|\theta), \theta \in \Theta$$

Thus

Something Notable

- The interpretation of LP hinges on the rather subtle point of allowing any observable X to draw conclusions about θ.

Therefore

- If there two ways to gather infromation about \theta, wither $X \sim f(x|\theta)$ or with $Y \sim g(x|\theta)$
  - with $X = x$ and $Y = y$ then

$$L(\theta|x) = \eta \times L(\theta|y), \forall \theta \in \Theta$$

Thus

Something Notable

- The interpretation of LP hinges on the rather subtle point of allowing any observable X to draw conclusions about θ.

Therefore

- If there two ways to gather infromation about \theta, wither $X \sim f(x|\theta)$ or with $Y \sim g(x|\theta)$
  - with $X = x$ and $Y = y$ then

$$L(\theta|x) = \eta \times L(\theta|y), \forall\theta \in \Theta$$

Outline

Cinvestav

In the case of Learning

Yes, we use the principle, but we add the idea of independence

- A trick to assume a set of samples $x_1, x_2, ..., x_N$ such that $x_i \sim f(X|\theta)$

Then, as we have seen it

$$\mathcal{L}(\theta) = f(x_1, x_2, ..., x_N|\theta) = \prod_{i=1}^{N} f(x_i|\theta)$$

In the case of Learning

Yes, we use the principle, but we add the idea of independence

- A trick to assume a set of samples $x_1, x_2, ..., x_N$ such that $x_i \sim f(X|\theta)$

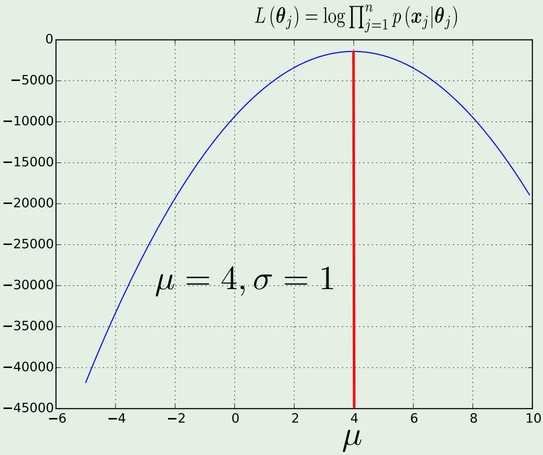
Then, as we have seen it

$$\mathcal{L}(\theta) = f(x_1, x_2, ..., x_N|\theta) = \prod_{i=1}^{N} f(x_i|\theta)$$

Cinvestav

Example, $p(\boldsymbol{x}|\omega_j) \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$

$L(\boldsymbol{\theta}_j) = \log \prod_{j=1}^{n} p(\boldsymbol{x}_j|\boldsymbol{\theta}_j)$



Outline

The Basics

Sufficiency Principle

- An **statistic** is sufficient with respect to a statistical model and its associated unknown parameter if
  - "no other statistic that can be calculated from the same sample provides any additional information as to the value of the parameter"[5]

However, as always

- We want a definition to build upon it... as always

The Basics

Sufficiency Principle

- An **statistic** is sufficient with respect to a statistical model and its associated unknown parameter if
  - "no other statistic that can be calculated from the same sample provides any additional information as to the value of the parameter"[5]

However, as always

- We want a definition to build upon it... as always

Cinvestav

A Basic Definition

Definition

- A statistic $t = T(X)$ is sufficient for underlying parameter θ precisely if the conditional probability distribution of the data X, given the statistic $t = T(X)$, does not depend on the parameter θ [6].

Something Notable

- This agreement is non-philosophical, it is rather a consequence of mathematics (**measure theoretic considerations**).

A Basic Definition

Definition

- A statistic $t = T(X)$ is sufficient for underlying parameter θ precisely if the conditional probability distribution of the data X, given the statistic $t = T(X)$, does not depend on the parameter θ [6].

Something Notable

- This agreement is non-philosophical, it is rather a consequence of mathematics (**measure theoretic considerations**).

Cinvestav

Outline

Cinvestav

Fisher's Factorization Theorem

Theorem

- Let $f(x|\theta)$ be the density or mass function for the random vector x, parametrized by the vector \theta. The statistic $t = T(x)$ is sufficient for θ if and only if there exist functions $a(x)$ (not depending on θ) and $b(t|\theta)$ such that ()

$$f(x|\theta) = a(x)\, b(t, \theta)$$

for all possible values of x.

Proof

First ⇒ (We will look only to the discrete case [7])

- Suppose $t = T(x)$ is sufficient for θ. Then, by definition

$$f(x|\theta, T(x) = t) \text{ is independient of } \theta$$

Let $f(x, t|\theta)$ denote the joint density function or mass function for $(X, T(X))$

- Observe $f(x|\theta) = f(x, t|\theta)$ then we have

Proof

First $\Rightarrow$ (We will look only to the discrete case [7])

- Suppose $t = T(x)$ is sufficient for θ. Then, by definition

$$f\left(x|\theta, T(x) = t\right) \text{ is independient of } \theta$$

Let $f(x, t|\theta)$ denote the joint density function or mass function for $(X, T(X))$

- Observe $f(x|\theta) = f(x, t|\theta)$ then we have

$$\begin{aligned} f(x|\theta) &= f(x, t|\theta) \\ &= f(x|\theta, t)\, f(t|\theta) \;\; \text{Bayesian} \\ &= \underbrace{a(x)}_{f(x|t)}\, \underbrace{b(t, \theta)}_{f(t|\theta)} \;\; \text{Independence} \end{aligned}$$

Proof

First $\Rightarrow$ (We will look only to the discrete case [7])

- Suppose $t = T(x)$ is sufficient for θ. Then, by definition

$$f(x|\theta, T(x) = t) \text{ is independient of } \theta$$

Let $f(x, t|\theta)$ denote the joint density function or mass function for $(X, T(X))$

- Observe $f(x|\theta) = f(x, t|\theta)$ then we have

$$\begin{aligned} f(x|\theta) &= f(x, t|\theta) \\ &= f(x|\theta, t)\, f(t|\theta) \quad \text{Bayesian} \\ &= \underbrace{a(x)\, b(t, \theta)}_{f(x|t) f(t|\theta)} \quad \text{Independence} \end{aligned}$$

Proof

First $\Rightarrow$ (We will look only to the discrete case [7])

- Suppose $t = T(x)$ is sufficient for θ. Then, by definition

$$f(x|\theta, T(x) = t) \text{ is independient of } \theta$$

Let $f(x, t|\theta)$ denote the joint density function or mass function for $(X, T(X))$

- Observe $f(x|\theta) = f(x, t|\theta)$ then we have

$$\begin{aligned} f(x|\theta) &= f(x, t|\theta) \\ &= f(x|\theta, t)\, f(t|\theta) \text{ Bayesian} \\ &= \underbrace{a(x)\; b(t, \theta)}_{f(x|t)\, f(t|\theta)} \text{ Independence} \end{aligned}$$

Proof

First $\Rightarrow$ (We will look only to the discrete case [7])

- Suppose $t = T(x)$ is sufficient for θ. Then, by definition

$$f(x|\theta, T(x) = t) \text{ is independient of } \theta$$

Let $f(x, t|\theta)$ denote the joint density function or mass function for $(X, T(X))$

- Observe $f(x|\theta) = f(x, t|\theta)$ then we have

$$\begin{aligned} f(x|\theta) &= f(x, t|\theta) \\ &= f(x|\theta, t)\, f(t|\theta) \text{ Bayesian} \\ &= \underbrace{a(x)\; b(t, \theta)}_{f(x|t) f(t|\theta)} \text{ Independence} \end{aligned}$$

Now, for the case $\Leftarrow$

Suppose the probability mass function for x can be written

$$f(x|\theta) = a(x)\, b(x|\theta) \text{ where } t = T(x)$$

The probability mass function for t is obtained by summing $f_\theta(x,t)$ over all x such that $T(x) = t$

$$\begin{aligned} f(t|\theta) &= \sum_{T(x)=t} f(x,t|\theta) \\ &= \sum_{T(x)=t} f(x|\theta) \longleftarrow \text{independence over } t \\ &= \sum_{T(x)=t} a(x)\, b_\theta(x) \end{aligned}$$

Now, for the case $\Leftarrow$

Suppose the probability mass function for x can be written

$$f(x|\theta) = a(x)\, b(x|\theta) \text{ where } t = T(x)$$

The probability mass function for t is obtained by summing $f_\theta(x,t)$ over all x such that $T(x) = t$

$$\begin{aligned} f(t|\theta) &= \sum_{T(x)=t} f(x,t|\theta) \\ &= \sum_{T(x)=t} f(x|\theta) \leftarrow \text{ independence over } t \\ &= \sum_{T(x)=t} a(x)\, b_\theta(x) \end{aligned}$$

Now, for the case $\Leftarrow$

Suppose the probability mass function for x can be written

$$f(x|\theta) = a(x)\, b(x|\theta) \text{ where } t = T(x)$$

The probability mass function for t is obtained by summing $f_\theta(x,t)$ over all x such that $T(x) = t$

$$f(t|\theta) = \sum_{T(x)=t} f(x,t|\theta)$$

$$= \sum_{T(x)=t} f(x|\theta) \leftarrow \text{independence over } t$$

$$= \sum_{T(x)=t} a(x)\, b_\theta(x)$$

Cinvestav

Now, for the case $\Leftarrow$

Suppose the probability mass function for x can be written

$$f(x|\theta) = a(x)\, b(x|\theta) \text{ where } t = T(x)$$

The probability mass function for t is obtained by summing $f_\theta(x,t)$ over all x such that $T(x) = t$

$$\begin{aligned} f(t|\theta) &= \sum_{T(x)=t} f(x,t|\theta) \\ &= \sum_{T(x)=t} f(x|\theta) \leftarrow \text{ independence over } t \\ &= \sum_{T(x)=t} a(x)\, b_\theta(x) \end{aligned}$$

Therefore, we have that

The conditional mass function of x given t

$$
\begin{aligned}
f(x|\theta, t) &= \frac{f(x, t|\theta)}{f(t|\theta)} \\
&= \frac{f(x|\theta)}{f(t|\theta)} \\
&= \frac{a(x)\, b_\theta(x)}{\sum_{T(x)=t} a(x)\, b_\theta(x)} = \frac{a(x)}{\sum_{T(x)=t} a(x)}
\end{aligned}
$$

This last expression does not depend on θ

- t is a sufficient statistic for θ.

Cinvestav

Therefore, we have that

The conditional mass function of x given t

$$
\begin{aligned}
f\left(x|\theta,t\right) &= \frac{f\left(x,t|\theta\right)}{f\left(t|\theta\right)}\\
&= \frac{f\left(x|\theta\right)}{f\left(t|\theta\right)}\\
&= \frac{a\left(x\right)b_{\theta}\left(x\right)}{\sum_{T(x)=t} a\left(x\right)b_{\theta}\left(x\right)} = \frac{a\left(x\right)}{\sum_{T(x)=t} a\left(x\right)}
\end{aligned}
$$

This last expression does not depend on θ

- t is a sufficient statistic for θ.

Cinvestav

Outline

Cinvestav

Using the Bernoulli Distribution

$x_n \sim$Bernoulli(θ) are i.d.d. $\forall n = 1, ..., N$

$$f\left(x_1, .., x_N|\theta\right) = \prod_{n=1}^{N} \theta^{x_n}\left(1-\theta\right)^{1-x_n}$$
$$= \theta^k\left(1-\theta\right)^{N-k}$$

- $k = \sum_{n=1}^{N} x_n$

Now, if we have the following choices

$$a(x) = 1 \text{ and } b_\theta(k) = \theta^k\left(1-\theta\right)^{N-k}$$

Cinvestav

Using the Bernoulli Distribution

$x_n \sim$Bernoulli(θ) are i.d.d. $\forall n = 1, ..., N$

$$f\left(x_1, .., x_N | \theta\right) = \prod_{n=1}^{N} \theta^{x_n} \left(1-\theta\right)^{1-x_n}$$
$$= \theta^k \left(1-\theta\right)^{N-k}$$

- $k = \sum_{n=1}^{N} x_n$

Now, if we have the following choices

$$a\left(x\right) = 1 \text{ and } b_\theta\left(k\right) = \theta^k \left(1-\theta\right)^{N-k}$$

Cinvestav

Therefore

Then choosing

- $T(x_1, .., x_N) = \sum_{n=1}^{N} x_n = k$

By the Fisher-Neyman Factorization Theorem

- k is sufficient for θ

Outline

Cinvestav

Something Quite Interesting

The Fisher-Neyman factorization lemma states

- The likelihood can be represented as

$$\ell\left(\theta\right) = f\left(x|\theta\right) = a\left(x\right) b_{\theta}\left(T\left(x\right)\right)$$

If the likelihood principle is adopted

All inference about θ should depend on sufficient statistics

Since $\ell(\theta) \propto b_\theta(T(x))$

Sufficiency Principle

- Let the two different observations x and y have the same values $T(x) = T(y)$, of a statistics sufficient for family $f(\cdot|\theta)$. Then the inferences about θ based on x and y should be the same.

Cinvestav

If the likelihood principle is adopted

All inference about θ should depend on sufficient statistics

Since $\ell(\theta) \propto b_\theta(T(x))$

Sufficiency Principle

- Let the two different observations x and y have the same values $T(x) = T(y)$, of a statistics sufficient for family $f(\cdot|\theta)$. Then the inferences about θ based on x and y should be the same.

Cinvestav

Outline

Cinvestav

Conditional Perspective

We have that

- **Conditional perspective** concerns reporting data specific measures of accuracy.

In contrast to the frequentist approach

- Performance of statistical procedures are judged looking at the observed data.

Cinvestav

Conditional Perspective

We have that

- **Conditional perspective** concerns reporting data specific measures of accuracy.

In contrast to the frequentist approach

- Performance of statistical procedures are judged looking at the observed data.

Cinvestav

Outline

Cinvestav

Example

Consider estimating θ in the model

$$P\left(X=\theta-1|\theta\right)=P\left(X=\theta+1|\theta\right) \text{ with } \theta\in\mathbb{R}$$

- on basis of two observations, X_1 and X_2 .

The procedure suggested is

$$\delta\left(X\right)=\begin{cases}\frac{X_1+X_2}{2} & \text{if } X_1\neq X_2\\ X_1-1 & \text{if } X_1=X_2\end{cases}$$

Example

Consider estimating θ in the model

$$P\left(X=\theta-1|\theta\right)=P\left(X=\theta+1|\theta\right) \text{ with } \theta\in\mathbb{R}$$

- on basis of two observations, X_1 and X_2 .

The procedure suggested is

$$\delta\left(X\right)=\begin{cases}\frac{X_1+X_2}{2} & \text{if } X_1\neq X_2\\ X_1-1 & \text{if } X_1=X_2\end{cases}$$

Therefore

To a frequentist, this procedure has confidence

- To a frequentist, this procedure has confidence of 75% for all θ, i.e., $P\left(\delta\left(X\right)=\theta\right)=0.75$.

The conditionalist would report the confidence

- 100% if observed data in hand are different
- 50% if the observations coincide

Therefore

To a frequentist, this procedure has confidence

- To a frequentist, this procedure has confidence of 75% for all θ, i.e., $P(\delta(X) = \theta) = 0.75$.

The conditionalist would report the confidence

- 100% if observed data in hand are different
- 50% if the observations coincide

Cinvestav

Then

Conditionality Principle

- If an experiment concerning the inference about θ is chosen from a collection of possible experiments, independently of θ, then any experiment not chosen is irrelevant to the inference.

Outline

Cinvestav

Not a good idea to integrate with respect to sample space

What?

- **A perfectly valid hypothesis can be rejected because the test failed to account for unlikely data that had not been observed...**

The Lindley Paradox

Suppose $\overline{y}|\theta \sim N\left(\theta, \frac{1}{n}\right)$

- We wish to test $H_0 : \theta = 0$ vs the two sided alternative.

Suppose a Bayesian puts the prior $P(\theta = 0) = P(\theta \neq 0) = \frac{1}{2}$

- The $\frac{1}{2}$ is uniformly spread over the interval $[-M/2, M/2]$.

Suppose $n = 40,000$ and $\overline{y} = 0.01$ are observed

- So, $\sqrt{n}\overline{y} = 2$

The Lindley Paradox

Suppose $\overline{y}|\theta \sim N\left(\theta, \frac{1}{n}\right)$

- We wish to test $H_0 : \theta = 0$ vs the two sided alternative.

Suppose a Bayesian puts the prior $P\left(\theta = 0\right) = P\left(\theta \neq 0\right) = \frac{1}{2}$

- The $\frac{1}{2}$ is uniformly spread over the interval $[-M/2, M/2]$.

Suppose $n = 40,000$ and $\overline{y} = 0.01$ are observed

- So, $\sqrt{n}\overline{y} = 2$

Cinvestav

The Lindley Paradox

Suppose $\overline{y}|\theta \sim N\left(\theta, \frac{1}{n}\right)$

- We wish to test $H_0 : \theta = 0$ vs the two sided alternative.

Suppose a Bayesian puts the prior $P\left(\theta = 0\right) = P\left(\theta \neq 0\right) = \frac{1}{2}$

- The $\frac{1}{2}$ is uniformly spread over the interval $[-M/2, M/2]$.

Suppose $n = 40,000$ and $\overline{y} = 0.01$ are observed

- So, $\sqrt{n}\overline{y} = 2$

Cinvestav

Therefore

Classical statistician

- She/he rejects H_0 at level $\alpha = 0.05$

Posterior odds in favor of H_0 are 11 if $M = 1$

- We will look at this... no worries, but Bayesian Statistician will choose H_0

Therefore

Classical statistician

- She/he rejects H_0 at level $\alpha = 0.05$

Posterior odds in favor of H_0 are 11 if $M = 1$

- We will look at this... no worries, but Bayesian Statistician will choose H_0

Cinvestav

Outline

Cinvestav

Using our likelihood

We have our function

$$\ell(\theta) = f(x|\theta)$$

Here

- The parameter θ is supported by the parameter space Θ and considered a random variable.
  - The random variable θ has a distribution $\pi(\theta)$ that is called the prior.

Using our likelihood

We have our function

$$\ell(\theta) = f(x|\theta)$$

Here

- The parameter θ is supported by the parameter space Θ and considered a random variable.
  - The random variable θ has a distribution $\pi(\theta)$ that is called the prior.

Cinvestav

Not only that

We have the following

- We can play a hierarchy game

$$\theta \sim \pi\left(\theta|\tau\right) \text{ where } \tau \text{ is called a hyperparameter}$$

This give us an idea about the marginals

$$m\left(x\right) = \int_{\Theta} f\left(x,\theta\right) = \int_{\Theta} f\left(x|\theta\right)\pi\left(\theta\right)d\theta$$

Not only that

We have the following

- We can play a hierarchy game

$$\theta \sim \pi\left(\theta|\tau\right) \text{ where } \tau \text{ is called a hyperparameter}$$

This give us an idea about the marginals

$$m\left(x\right)=\int_{\Theta} f\left(x,\theta\right)=\int_{\Theta} f\left(x|\theta\right)\pi\left(\theta\right)d\theta$$

Cinvestav

What about the posterior?

We have the following

$$f(\theta|x) = \frac{f(x,\theta)}{m(x)}$$

$$= \frac{f(x|\theta)\,\pi(\theta)}{m(x)}$$

$$= \frac{f(x|\theta)\,\pi(\theta)}{\int_{\Theta} f(x|\theta)\,\pi(\theta)\,d\theta}$$

What about the posterior?

We have the following

$$
\begin{aligned}
f(\theta|x) &= \frac{f(x,\theta)}{m(x)} \\
&= \frac{f(x|\theta)\,\pi(\theta)}{m(x)} \\
&= \frac{f(x|\theta)\,\pi(\theta)}{\int_{\Theta} f(x|\theta)\,\pi(\theta)\,d\theta}
\end{aligned}
$$

What about the posterior?

We have the following

$$
\begin{aligned}
f(\theta|x) &= \frac{f(x,\theta)}{m(x)} \\
&= \frac{f(x|\theta)\,\pi(\theta)}{m(x)} \\
&= \frac{f(x|\theta)\,\pi(\theta)}{\int_{\Theta} f(x|\theta)\,\pi(\theta)\,d\theta}
\end{aligned}
$$

Cinvestav

Outline

Cinvestav

An interesting case

Suppose that the observations are coming from $N\left(\theta, \sigma_1^2\right)$

- Assume prior on θ is $N\left(\sigma_2, \sigma_2\right)$

Then, under this setup

- the normal/normal model, the posterior is $f\left(\theta|X_1, ..., X_n\right) = f\left(\theta|\overline{X}\right)$

An interesting case

Suppose that the observations are coming from $N\left(\theta, \sigma_1^2\right)$

- Assume prior on θ is $N\left(\sigma_2, \sigma_2\right)$

Then, under this setup

- the normal/normal model, the posterior is $f\left(\theta|X_1, ..., X_n\right) = f\left(\theta|\overline{X}\right)$

Cinvestav

The connection

Lemma

- Suppose the sufficient statistics $T = T(X_1, ..., X_n)$ exist. Then $f(\theta|X_1, ..., X_n) = f(\theta|T)$.

Proof

Factorization theorem for sufficient statistics is

$$f\left(x|\theta\right) = b_{\theta}\left(t\right) a\left(x\right)$$

Where

- $t = T\left(x\right)$ and $a\left(x\right)$ do not depend on θ.

Proof

Factorization theorem for sufficient statistics is

$$f(x|\theta) = b_\theta(t)\, a(x)$$

Where

- $t = T(x)$ and $a(x)$ do not depend on θ.

Cinvestav

Furhtermore

Thus

$$\pi\left(\theta|x\right)=\frac{f\left(x|\theta\right)\pi\left(\theta\right)}{\int_{\Theta}f\left(x|\theta\right)\pi\left(\theta\right)d\theta}$$

$$=\frac{b_{\theta}\left(t\right)a\left(x\right)\pi\left(\theta\right)}{\int_{\Theta}b_{\theta}\left(t\right)a\left(x\right)\pi\left(\theta\right)d\theta}$$

$$=\frac{b_{\theta}\left(t\right)\pi\left(\theta\right)}{\int_{\Theta}b_{\theta}\left(t\right)\pi\left(\theta\right)d\theta}$$

Furhtermore

Thus

$$
\begin{aligned}
\pi(\theta|x) &= \frac{f(x|\theta)\,\pi(\theta)}{\int_{\Theta} f(x|\theta)\,\pi(\theta)\,d\theta} \\
&= \frac{b_\theta(t)\,a(x)\,\pi(\theta)}{\int_{\Theta} b_\theta(t)\,a(x)\,\pi(\theta)\,d\theta} \\
&= \frac{b_\theta(t)\,\pi(\theta)}{\int_{\Theta} b_\theta(t)\,\pi(\theta)\,d\theta}
\end{aligned}
$$

Cinvestav

Furhtermore

Thus

$$\pi\left(\theta|x\right) = \frac{f\left(x|\theta\right)\pi\left(\theta\right)}{\int_{\Theta} f\left(x|\theta\right)\pi\left(\theta\right)d\theta}$$
$$= \frac{b_{\theta}\left(t\right)a\left(x\right)\pi\left(\theta\right)}{\int_{\Theta} b_{\theta}\left(t\right)a\left(x\right)\pi\left(\theta\right)d\theta}$$
$$= \frac{b_{\theta}\left(t\right)\pi\left(\theta\right)}{\int_{\Theta} b_{\theta}\left(t\right)\pi\left(\theta\right)d\theta}$$

The

Multiply and divide by $\phi(t)$

$$= \frac{b_\theta(t)\,\pi(\theta)\,\phi(t)}{\int_\Theta b_\theta(t)\,\pi(\theta)\,\phi(t)\,d\theta}$$

$$= \frac{b_\theta(t)\,\pi(\theta)\,\phi(t)}{\int_\Theta b_\theta(t)\,\pi(\theta)\,\phi(t)\,d\theta}$$

$$= \frac{\pi(\theta)\,f(t|\theta)}{\int_\Theta \pi(\theta)\,f(t|\theta)\,d\theta} = \pi(\theta|t)$$

The

Multiply and divide by $\phi(t)$

$$
\begin{aligned}
&= \frac{b_\theta(t)\,\pi(\theta)\,\phi(t)}{\int_\Theta b_\theta(t)\,\pi(\theta)\,\phi(t)\,d\theta} \\
&= \frac{b_\theta(t)\,\pi(\theta)\,\phi(t)}{\int_\Theta b_\theta(t)\,\pi(\theta)\,\phi(t)\,d\theta} \\
&= \frac{\pi(\theta)\,f(t|\theta)}{\int_\Theta \pi(\theta)\,f(t|\theta)\,d\theta} = \pi(\theta|t)
\end{aligned}
$$

The

Multiply and divide by $\phi(t)$

$$
\begin{aligned}
&= \frac{b_\theta(t)\,\pi(\theta)\,\phi(t)}{\int_\Theta b_\theta(t)\,\pi(\theta)\,\phi(t)\,d\theta} \\
&= \frac{b_\theta(t)\,\pi(\theta)\,\phi(t)}{\int_\Theta b_\theta(t)\,\pi(\theta)\,\phi(t)\,d\theta} \\
&= \frac{\pi(\theta)\,f(t|\theta)}{\int_\Theta \pi(\theta)\,f(t|\theta)\,d\theta} = \pi(\theta|t)
\end{aligned}
$$

Cinvestav

Here, we have

The following equations

$$f(t|\theta) = \int_{x:T(x)=t} f(x|\theta)\, dx = \int_{x:T(x)=t} b_\theta(t)\, a(x)\, dx$$

Then

$$\int_{x:T(x)=t} b_\theta(t)\, a(x)\, dx = b_\theta(t) \int_{x:T(x)=t} a(x)\, dx = b_\theta(t)\, \phi(t)$$

Cinvestav

Here, we have

The following equations

$$f\left(t|\theta\right)=\int_{x:T(x)=t}f\left(x|\theta\right)dx=\int_{x:T(x)=t}b_{\theta}\left(t\right)a\left(x\right)dx$$

Then

$$\int_{x:T(x)=t}b_{\theta}\left(t\right)a\left(x\right)dx=b_{\theta}\left(t\right)\int_{x:T(x)=t}a\left(x\right)dx=b_{\theta}\left(t\right)\phi\left(t\right)$$

Cinvestav

Then

We have the following definition

Definition

- The statistics $T = T(X)$ is sufficient (in the Bayesian sense) if for any prior the resulting posterior satisfies

$$\pi\left(\theta|X\right) = \pi\left(\theta|T\right)$$

This is equivalent to the classic definition on sufficient statistics

Theorem

- T is sufficient in the Bayesian sense if and only if it is sufficient in the usual sense.

Cinvestav

Then

We have the following definition

Definition

- The statistics $T = T(X)$ is sufficient (in the Bayesian sense) if for any prior the resulting posterior satisfies

$$\pi(\theta|X) = \pi(\theta|T)$$

This is equivalent to the classic definition on sufficient statistics

Theorem

- T is sufficient in the Bayesian sense if and only if it is sufficient in the usual sense.

Outline

Cinvestav

Something quite important

Something Notable

- The posterior is the ultimate experimental summary for a Bayesian.

Not only that

- The location measures (especially the mean) of the posterior are of importance.

There is an important idea

- The posterior mode and median are also Bayes estimators under different loss functions!!!

Something quite important

Something Notable

- The posterior is the ultimate experimental summary for a Bayesian.

Not only that

- The location measures (especially the mean) of the posterior are of importance.

There is an important idea

- The posterior mode and median are also Bayes estimators under different loss functions!!!

Something quite important

Something Notable

- The posterior is the ultimate experimental summary for a Bayesian.

Not only that

- The location measures (especially the mean) of the posterior are of importance.

There is an important idea

- The posterior mode and median are also Bayes estimators under different loss functions!!!

Furthermore

Generalized Maximum Likelihood Estimator AKA MAP (Maximum Aposteriori)

- The generalized MLE is the largest mode of the $\pi(\theta|x)$.

Outline

Cinvestav

What can we do?

We can specify a distribution

Then, learn the parameters

Remember the Bayesian Rule

$$p(\Theta|\mathcal{X}) = \frac{p(\mathcal{X}|\Theta)\,p(\Theta)}{p(\mathcal{X})} \quad (1)$$

We seek that value for Θ, called $\hat{\Theta}_{MAP}$

It allows to maximize the posterior $p(\Theta|\mathcal{X})$

What can we do?

We can specify a distribution

Then, learn the parameters

Remember the Bayesian Rule

$$p(\Theta|\mathcal{X}) = \frac{p(\mathcal{X}|\Theta)\, p(\Theta)}{p(\mathcal{X})} \tag{1}$$

We seek that value for Θ, called $\hat{\Theta}_{MAP}$

It allows to maximize the posterior $p(\Theta|\mathcal{X})$

What can we do?

We can specify a distribution

Then, learn the parameters

Remember the Bayesian Rule

$$p\left(\Theta|\mathcal{X}\right)=\frac{p\left(\mathcal{X}|\Theta\right)p\left(\Theta\right)}{p\left(\mathcal{X}\right)} \quad (1)$$

We seek that value for Θ, called $\widehat{\Theta}_{MAP}$

It allows to maximize the posterior $p\left(\Theta|\mathcal{X}\right)$

Therefore

We can use this idea of maximizing the posterior

To obtain the distribution through the Maximum a Posteriori

Development of the solution

We look to maximize $\widehat{\Theta}_{MAP}$

$$\widehat{\Theta}_{MAP} = \underset{\Theta}{\text{argmax}} p\left(\Theta|\mathcal{X}\right)$$

$$= \underset{\Theta}{\text{argmax}} \frac{p\left(\mathcal{X}|\Theta\right) p\left(\Theta\right)}{P\left(\mathcal{X}\right)}$$

$$\approx \underset{\Theta}{\text{argmax}} p\left(\mathcal{X}|\Theta\right) p\left(\Theta\right)$$

$$= \underset{\Theta}{\text{argmax}} \prod_{x_i \in \mathcal{X}} p\left(x_i|\Theta\right) p\left(\Theta\right)$$

$P\left(\mathcal{X}\right)$ can be removed because it has no functional relation with Θ.

Development of the solution

We look to maximize $\widehat{\Theta}_{MAP}$

$$\begin{aligned}\widehat{\Theta}_{MAP} &= \underset{\Theta}{\text{argmax}}\, p\left(\Theta|\mathcal{X}\right)\\ &= \underset{\Theta}{\text{argmax}} \frac{p\left(\mathcal{X}|\Theta\right)p\left(\Theta\right)}{P\left(\mathcal{X}\right)}\\ &\approx \underset{\Theta}{\text{argmax}}\, p\left(\mathcal{X}|\Theta\right)p\left(\Theta\right)\\ &= \underset{\Theta}{\text{argmax}} \prod_{x_i \in \mathcal{X}} p\left(x_i|\Theta\right)p\left(\Theta\right)\end{aligned}$$

$P\left(\mathcal{X}\right)$ can be removed because it has no functional relation with Θ.

Development of the solution

We look to maximize $\widehat{\Theta}_{MAP}$

$$
\begin{aligned}
\widehat{\Theta}_{MAP} &= \underset{\Theta}{\text{argmax}}\, p\left(\Theta|\mathcal{X}\right) \\
&= \underset{\Theta}{\text{argmax}} \frac{p\left(\mathcal{X}|\Theta\right)p\left(\Theta\right)}{P\left(\mathcal{X}\right)} \\
&\approx \underset{\Theta}{\text{argmax}}\, p\left(\mathcal{X}|\Theta\right)p\left(\Theta\right) \\
&= \underset{\Theta}{\text{argmax}} \prod_{x_i \in \mathcal{X}} p\left(x_i|\Theta\right)p\left(\Theta\right)
\end{aligned}
$$

$P\left(\mathcal{X}\right)$ can be removed because it has no functional relation with Θ.

Cinvestav

Development of the solution

We look to maximize $\widehat{\Theta}_{MAP}$

$$
\begin{aligned}
\widehat{\Theta}_{MAP} &= \underset{\Theta}{\operatorname{argmax}} p\left(\Theta|\mathcal{X}\right) \\
&= \underset{\Theta}{\operatorname{argmax}} \frac{p\left(\mathcal{X}|\Theta\right) p\left(\Theta\right)}{P\left(\mathcal{X}\right)} \\
&\approx \underset{\Theta}{\operatorname{argmax}} p\left(\mathcal{X}|\Theta\right) p\left(\Theta\right) \\
&= \underset{\Theta}{\operatorname{argmax}} \prod_{x_i \in \mathcal{X}} p\left(x_i|\Theta\right) p\left(\Theta\right)
\end{aligned}
$$

$P\left(\mathcal{X}\right)$ can be removed because it has no functional relation with Θ.

Development of the solution

We look to maximize $\widehat{\Theta}_{MAP}$

$$\begin{aligned}\widehat{\Theta}_{MAP} &= \underset{\Theta}{\text{argmax}} p\left(\Theta|\mathcal{X}\right)\\ &= \underset{\Theta}{\text{argmax}} \frac{p\left(\mathcal{X}|\Theta\right)p\left(\Theta\right)}{P\left(\mathcal{X}\right)}\\ &\approx \underset{\Theta}{\text{argmax}} p\left(\mathcal{X}|\Theta\right)p\left(\Theta\right)\\ &= \underset{\Theta}{\text{argmax}} \prod_{x_i\in\mathcal{X}} p\left(x_i|\Theta\right)p\left(\Theta\right)\end{aligned}$$

$P\left(\mathcal{X}\right)$ can be removed because it has no functional relation with Θ.

We can make this easier

Use logarithms

$$\widehat{\Theta}_{MAP} = \underset{\Theta}{\text{argmax}} \left[\sum_{x_i \in \mathcal{X}} \log p\left(x_i | \Theta\right) + \log p\left(\Theta\right) \right] \quad (2)$$

What Does the MAP Estimate Get?

Something Notable

The MAP estimate allows us to inject into the estimation calculation our prior beliefs regarding the parameters values in Θ.

What Does the MAP Estimate Get?

Something Notable

The MAP estimate allows us to inject into the estimation calculation our prior beliefs regarding the parameters values in Θ.

For example

Let's conduct N independent trials of the following Bernoulli experiment with q parameter:

- We will ask each individual we run into in the hallway whether they will vote PRI or PAN in the next presidential election.

Cinvestav

What Does the MAP Estimate Get?

Something Notable

The MAP estimate allows us to inject into the estimation calculation our prior beliefs regarding the parameters values in Θ.

For example

Let's conduct N independent trials of the following Bernoulli experiment with q parameter:

- We will ask each individual we run into in the hallway whether they will vote PRI or PAN in the next presidential election.

With probability q to vote PRI

Where the values of x_i is either PRI or PAN.

What Does the MAP Estimate Get?

Something Notable

The MAP estimate allows us to inject into the estimation calculation our prior beliefs regarding the parameters values in Θ.

For example

Let's conduct N independent trials of the following Bernoulli experiment with q parameter:

- We will ask each individual we run into in the hallway whether they will vote PRI or PAN in the next presidential election.

With probability q to vote PRI

Where the values of x_i is either PRI or PAN.

Outline

Cinvestav

First the Maximum Likelihood Estimate

Samples

$$\mathcal{X} = \left\{ x_i = \begin{cases} PAN \\ PRI \end{cases} \quad i = 1, ..., N \right\} \tag{3}$$

The log likelihood function

First the Maximum Likelihood Estimate

Samples

$$\mathcal{X} = \left\{ x_i = \begin{cases} PAN \\ PRI \end{cases} \quad i = 1, ..., N \right\} \tag{3}$$

The log likelihood function

$$\begin{aligned} \log \ p(\mathcal{X}|q) = & \sum_{i=1}^{N} \log \ p(x_i|q) \\ = & \sum_{i} \log \ p(x_i = PRI|q) + ... \\ & \sum_{i} \log \ p(x_i = PAN|1-q) \\ = & n_{PRI} \log(q) + (N - n_{PRI}) \log(1-q) \end{aligned}$$

Where n_{PRI} are the numbers of individuals who are planning to vote PRI this fall

First the Maximum Likelihood Estimate

Samples

$$\mathcal{X} = \left\{ x_i = \begin{cases} PAN \\ PRI \end{cases} \quad i = 1, ..., N \right\} \tag{3}$$

The log likelihood function

$$\begin{aligned} \log\ p\left(\mathcal{X}|q\right) &= \sum_{i=1}^{N} \log\ p\left(x_i|q\right) \\ &= \sum_{i} \log\ p\left(x_i = PRI|q\right) + ... \\ & \sum_{i} \log\ p\left(x_i = PAN|1-q\right) \\ &= n_{PRI} \log\left(q\right) + \left(N - n_{PRI}\right) \log\left(1-q\right) \end{aligned}$$

Where n_{PRI} are the numbers of individuals who are planning to vote PRI this fall

First the Maximum Likelihood Estimate

Samples

$$\mathcal{X} = \left\{ x_i = \begin{cases} PAN \\ PRI \end{cases} \quad i = 1, ..., N \right\} \tag{3}$$

The log likelihood function

$$\begin{aligned} \log\ p\left(\mathcal{X}|q\right) =& \sum_{i=1}^{N} \log\ p\left(x_i|q\right) \\ =& \sum_{i} \log\ p\left(x_i = PRI|q\right) + ... \\ & \sum_{i} \log\ p\left(x_i = PAN|1-q\right) \\ =& n_{PRI} \log\left(q\right) + \left(N - n_{PRI}\right) \log\left(1-q\right) \end{aligned}$$

Where n_{PRI} are the numbers of individuals who are planning to vote PRI this fall

First the Maximum Likelihood Estimate

Samples

$$\mathcal{X} = \left\{ x_i = \begin{cases} PAN \\ PRI \end{cases} \quad i = 1, ..., N \right\} \tag{3}$$

The log likelihood function

$$\begin{aligned} \log \ p\left(\mathcal{X}|q\right) =& \sum_{i=1}^{N} \log \ p\left(x_i|q\right) \\ =& \sum_{i} \log \ p\left(x_i = PRI|q\right) + ... \\ & \sum_{i} \log \ p\left(x_i = PAN|1-q\right) \\ =& n_{PRI} \log\left(q\right) + \left(N - n_{PRI}\right) \log\left(1-q\right) \end{aligned}$$

Where n_{PRI} are the numbers of individuals who are planning to vote PRI this fall

We use our classic tricks

By setting

$$\mathcal{L} = \log \; p\left(\mathcal{X}|q\right) \tag{4}$$

We have that

$$\frac{\partial \mathcal{L}}{\partial q} = 0 \tag{5}$$

Thus

$$\frac{n_{PRI}}{q} - \frac{\left(N - n_{PRI}\right)}{\left(1 - q\right)} = 0 \tag{6}$$

We use our classic tricks

By setting

$$\mathcal{L} = \log \; p\left(\mathcal{X}|q\right) \tag{4}$$

We have that

$$\frac{\partial \mathcal{L}}{\partial q} = 0 \tag{5}$$

Thus

$$\frac{n_{PRI}}{q} - \frac{\left(N - n_{PRI}\right)}{\left(1 - q\right)} = 0 \tag{6}$$

We use our classic tricks

By setting

$$\mathcal{L} = \log \ p\left(\mathcal{X}|q\right) \tag{4}$$

We have that

$$\frac{\partial \mathcal{L}}{\partial q} = 0 \tag{5}$$

Thus

$$\frac{n_{PRI}}{q} - \frac{\left(N - n_{PRI}\right)}{\left(1 - q\right)} = 0 \tag{6}$$

Cinvestav

Final Solution of ML

We get

$$\widehat{q}_{PRI} = \frac{n_{PRI}}{N} \tag{7}$$

Thus

If we say that $N = 20$ and if 12 are going to vote PRI, we get $\widehat{q}_{PRI} = 0.6$.

Final Solution of ML

We get

$$\widehat{q}_{PRI} = \frac{n_{PRI}}{N} \tag{7}$$

Thus

If we say that $N = 20$ and if 12 are going to vote PRI, we get $\widehat{q}_{PRI} = 0.6$.

Building the MAP estimate

Obviously we need a prior belief distribution

We have the following constraints:

- The prior for q must be zero outside the $[0, 1]$ interval.
- Within the $[0, 1]$ interval, we are free to specify our beliefs in any way we wish.
- In most cases, we would want to choose a distribution for the prior beliefs that peaks somewhere in the $[0, 1]$ interval.

Cinvestav

Building the MAP estimate

Obviously we need a prior belief distribution

We have the following constraints:

- The prior for q must be zero outside the $[0,1]$ interval.
- Within the $[0,1]$ interval, we are free to specify our beliefs in any way we wish.
- In most cases, we would want to choose a distribution for the prior beliefs that peaks somewhere in the $[0,1]$ interval.

We assume the following

- The state of Colima has traditionally voted PRI in presidential elections.
- However, on account of the prevailing economic conditions, the voters are more likely to vote PAN in the election in question.

Building the MAP estimate

Obviously we need a prior belief distribution

We have the following constraints:

- The prior for q must be zero outside the $[0, 1]$ interval.
- Within the $[0, 1]$ interval, we are free to specify our beliefs in any way we wish.
- In most cases, we would want to choose a distribution for the prior beliefs that peaks somewhere in the $[0, 1]$ interval.

We assume the following

- The state of Colima has traditionally voted PRI in presidential elections.
- However, on account of the prevailing economic conditions, the voters are more likely to vote PAN in the election in question.

Building the MAP estimate

Obviously we need a prior belief distribution

We have the following constraints:

- The prior for q must be zero outside the $[0, 1]$ interval.
- Within the $[0, 1]$ interval, we are free to specify our beliefs in any way we wish.
- In most cases, we would want to choose a distribution for the prior beliefs that peaks somewhere in the $[0, 1]$ interval.

We assume the following

- The state of Colima has traditionally voted PRI in presidential elections.
- However, on account of the prevailing economic conditions, the voters are more likely to vote PAN in the election in question.

Building the MAP estimate

Obviously we need a prior belief distribution

We have the following constraints:

- The prior for q must be zero outside the $[0, 1]$ interval.
- Within the $[0, 1]$ interval, we are free to specify our beliefs in any way we wish.
- In most cases, we would want to choose a distribution for the prior beliefs that peaks somewhere in the $[0, 1]$ interval.

We assume the following

- The state of Colima has traditionally voted PRI in presidential elections.
- However, on account of the prevailing economic conditions, the voters are more likely to vote PAN in the election in question.

Building the MAP estimate

Obviously we need a prior belief distribution

We have the following constraints:

- The prior for q must be zero outside the $[0, 1]$ interval.
- Within the $[0, 1]$ interval, we are free to specify our beliefs in any way we wish.
- In most cases, we would want to choose a distribution for the prior beliefs that peaks somewhere in the $[0, 1]$ interval.

We assume the following

- The state of Colima has traditionally voted PRI in presidential elections.
- However, on account of the prevailing economic conditions, the voters are more likely to vote PAN in the election in question.

What prior distribution can we use?

We could use a Beta distribution being parametrized by two values α and β

$$p(q) = \frac{1}{B(\alpha, \beta)} q^{\alpha-1} (1-q)^{\beta-1}. \tag{8}$$

Where

We have $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the beta function where Γ is the generalization of the notion of factorial in the case of the real numbers.

Properties

When both the $\alpha, \beta > 0$ then the beta distribution has its mode (Maximum value) at

$$\frac{\alpha-1}{\alpha+\beta-2}. \tag{9}$$

What prior distribution can we use?

We could use a Beta distribution being parametrized by two values α and β

$$p(q) = \frac{1}{B(\alpha, \beta)} q^{\alpha-1} (1-q)^{\beta-1}. \tag{8}$$

Where

We have $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the beta function where Γ is the generalization of the notion of factorial in the case of the real numbers.

Properties

When both the $\alpha, \beta > 0$ then the beta distribution has its mode (Maximum value) at

$$\frac{\alpha-1}{\alpha+\beta-2}. \tag{9}$$

What prior distribution can we use?

We could use a Beta distribution being parametrized by two values α and β

$$p(q) = \frac{1}{B(\alpha, \beta)} q^{\alpha-1} (1-q)^{\beta-1}. \tag{8}$$

Where

We have $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the beta function where Γ is the generalization of the notion of factorial in the case of the real numbers.

Properties

When both the $\alpha, \beta > 0$ then the beta distribution has its mode (Maximum value) at

$$\frac{\alpha - 1}{\alpha + \beta - 2}. \tag{9}$$

We then do the following

We do the following

We can choose $\alpha = \beta$ so the beta prior peaks at 0.5.

As a further expression of our belief

We make the following choice $\alpha = \beta = 5$.

Why? Look at the variance of the beta distribution

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{10}$$

We then do the following

We do the following

We can choose $\alpha = \beta$ so the beta prior peaks at 0.5.

As a further expression of our belief

We make the following choice $\alpha = \beta = 5$.

Why? Look at the variance of the beta distribution

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{10}$$

We then do the following

We do the following

We can choose $\alpha = \beta$ so the beta prior peaks at 0.5.

As a further expression of our belief

We make the following choice $\alpha = \beta = 5$.

Why? Look at the variance of the beta distribution

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{10}$$

Cinvestav

Thus, we have the following nice properties

We have a variance with $\alpha = \beta = 5$

$Var\left(q\right) \approx 0.025$

Thus, the standard deviation

$sd \approx 0.16$ which is a nice dispersion at the peak point!!!

Thus, we have the following nice properties

We have a variance with $\alpha = \beta = 5$

$Var(q) \approx 0.025$

Thus, the standard deviation

$sd \approx 0.16$ which is a nice dispersion at the peak point!!!

Now, our MAP estimate for $\widehat{p}_{MAP}$...

We have then

$$\widehat{p}_{MAP} = \underset{\Theta}{\text{argmax}} \left[\sum_{x_i \in \mathcal{X}} \log p\left(x_i | q\right) + \log p\left(q\right) \right] \tag{11}$$

Plugging back the ML

$$\widehat{p}_{MAP} = \underset{\Theta}{\text{argmax}} \left[n_{PRI} \log q + \left(N - n_{PRI}\right) \log\left(1 - q\right) + \log p\left(q\right) \right] \tag{12}$$

Where

$$\log p\left(q\right) = \log \left(\frac{1}{B\left(\alpha, \beta\right)} q^{\alpha - 1} \left(1 - q\right)^{\beta - 1} \right) \tag{13}$$

Now, our MAP estimate for $\widehat{p}_{MAP}$...

We have then

$$\widehat{p}_{MAP} = \underset{\Theta}{\text{argmax}} \left[\sum_{x_i \in \mathcal{X}} \log p\left(x_i | q\right) + \log p\left(q\right) \right] \tag{11}$$

Plugging back the ML

$$\widehat{p}_{MAP} = \underset{\Theta}{\text{argmax}} \left[n_{PRI} \log q + \left(N - n_{PRI}\right) \log\left(1 - q\right) + \log p\left(q\right) \right] \tag{12}$$

Where

$$\log p\left(q\right) = \log\left(\frac{1}{B\left(\alpha, \beta\right)} q^{\alpha - 1} \left(1 - q\right)^{\beta - 1}\right) \tag{13}$$

Now, our MAP estimate for $\widehat{p}_{MAP}$...

We have then

$$\widehat{p}_{MAP} = \underset{\Theta}{\text{argmax}} \left[\sum_{x_i \in \mathcal{X}} \log p\left(x_i | q\right) + \log p\left(q\right) \right] \quad (11)$$

Plugging back the ML

$$\widehat{p}_{MAP} = \underset{\Theta}{\text{argmax}} \left[n_{PRI} \log q + \left(N - n_{PRI}\right) \log\left(1-q\right) + \log p\left(q\right) \right] \quad (12)$$

Where

$$\log p\left(q\right) = \log\left(\frac{1}{B\left(\alpha, \beta\right)} q^{\alpha-1} \left(1-q\right)^{\beta-1} \right) \quad (13)$$

The log of $p(q)$

We have that

$$\log p(q) = (\alpha - 1)\log q + (\beta - 1)\log(1 - q) - \log B(\alpha, \beta) \tag{14}$$

Now taking the derivative with respect to q, we get

$$\frac{n_{PRI}}{q} - \frac{(N - n_{PRI})}{(1 - q)} - \frac{\beta - 1}{1 - q} + \frac{\alpha - 1}{q} = 0 \tag{15}$$

Thus

$$\widehat{q}_{MAP} = \frac{n_{PRI} + \alpha - 1}{N + \alpha + \beta - 2} \tag{16}$$

The log of $p(q)$

We have that

$$\log p(q) = (\alpha - 1)\log q + (\beta - 1)\log(1 - q) - \log B(\alpha, \beta) \tag{14}$$

Now taking the derivative with respect to q, we get

$$\frac{n_{PRI}}{q} - \frac{(N - n_{PRI})}{(1 - q)} - \frac{\beta - 1}{1 - q} + \frac{\alpha - 1}{q} = 0 \tag{15}$$

Thus

$$\widehat{q}_{MAP} = \frac{n_{PRI} + \alpha - 1}{N + \alpha + \beta - 2} \tag{16}$$

The log of $p(q)$

We have that

$$\log p(q) = (\alpha - 1)\log q + (\beta - 1)\log(1 - q) - \log B(\alpha, \beta) \tag{14}$$

Now taking the derivative with respect to q, we get

$$\frac{n_{PRI}}{q} - \frac{(N - n_{PRI})}{(1 - q)} - \frac{\beta - 1}{1 - q} + \frac{\alpha - 1}{q} = 0 \tag{15}$$

Thus

$$\widehat{q}_{MAP} = \frac{n_{PRI} + \alpha - 1}{N + \alpha + \beta - 2} \tag{16}$$

Now

With $N = 20$ with $n_{PRI} = 12$ and $\alpha = \beta = 5$

$$\widehat{q}_{MAP} = 0.571$$

Another Example

Let $X_1, \ldots, X_n$ given θ are Poisson $\mathcal{P}(\theta)$ with probability $f(x_i|\theta) = \frac{\theta^{x_i}}{x_i!}e^{-\theta}$

- Assume $\theta \sim \Gamma(\alpha, \beta)$ given by $\pi(\theta) \propto \theta^{\alpha-1}e^{-\beta\theta}$

The MAP is equal to

$$\pi(\theta|X_1, X_2, \ldots, X_n) = \pi\left(\theta|\sum X_i\right) \propto \theta^{\sum X_i + \alpha - 1}e^{-(n+\beta)\theta}$$

- Basically $\Gamma\left(\sum X_i + \alpha - 1, n + \beta\right)$

The mean is

$$E[\theta|X] = \frac{\sum X_i + \alpha}{n + \beta}$$

Another Example

Let $X_1, ..., X_n$ given θ are Poisson $\mathcal{P}(\theta)$ with probability $f(x_i|\theta) = \frac{\theta^{x_i}}{x_i!}e^{-\theta}$

- Assume $\theta \sim \Gamma(\alpha, \beta)$ given by $\pi(\theta) \propto \theta^{\alpha-1}e^{-\beta\theta}$

The MAP is equal to

$$\pi(\theta|X_1, X_2, ..., X_n) = \pi\left(\theta|\sum X_i\right) \propto \theta^{\sum X_i+\alpha-1}e^{-(n+\beta)\theta}$$

- Basically $\Gamma\left(\sum X_i + \alpha - 1, n+\beta\right)$

The mean is

$$E[\theta|X] = \frac{\sum X_i + \alpha}{n+\beta}$$

Another Example

Let $X_1, ..., X_n$ given θ are Poisson $\mathcal{P}(\theta)$ with probability $f(x_i|\theta) = \frac{\theta^{x_i}}{x_i!}e^{-\theta}$

- Assume $\theta \sim \Gamma(\alpha, \beta)$ given by $\pi(\theta) \propto \theta^{\alpha-1}e^{-\beta\theta}$

The MAP is equal to

$$\pi(\theta|X_1, X_2, ..., X_n) = \pi\left(\theta|\sum X_i\right) \propto \theta^{\sum X_i+\alpha-1}e^{-(n+\beta)\theta}$$

- Basically $\Gamma(\sum X_i + \alpha - 1, n + \beta)$

The mean is

$$E[\theta|X] = \frac{\sum X_i + \alpha}{n + \beta}$$

Cinvestav

Now, given the mean of the Γ

We can rewrite the mean as

$$E\left[\theta|X\right] = \frac{n}{n+\beta} \times \frac{\sum \boldsymbol{X_i}}{\boldsymbol{n}} + \frac{\beta}{\beta+n} \times \frac{\boldsymbol{\alpha}}{\boldsymbol{\beta}}$$

Given that the means are

- Mean of MLE $\frac{\sum X_i}{n}$
- Mean of the prior $\frac{\alpha}{\beta}$

Now, given the mean of the Γ

We can rewrite the mean as

$$E\left[\theta|X\right] = \frac{n}{n+\beta} \times \frac{\sum \boldsymbol{X_i}}{\boldsymbol{n}} + \frac{\beta}{\beta+n} \times \frac{\boldsymbol{\alpha}}{\boldsymbol{\beta}}$$

Given that the means are

- Mean of MLE $\frac{\sum \boldsymbol{X_i}}{\boldsymbol{n}}$
- Mean of the prior $\frac{\boldsymbol{\alpha}}{\boldsymbol{\beta}}$

Remarks

Something Notable

- The standard MLE maximizes $\pi(\theta|x)$, while the generalized MLE maximizes $\pi(\theta)\,\ell(\theta)$.
  - Quite funny we call that Maximum Aposteriori (MAP) estimator!!!

The MAP estimator is since it is often simpler to calculate given that

$$\arg\max_{\theta} \pi(\theta|x) = \arg\max_{\theta} f(x|\theta)\,\pi(\theta)$$

- Given that the normalization factor is a constant

Remarks

Something Notable

- The standard MLE maximizes $\pi(\theta|x)$, while the generalized MLE maximizes $\pi(\theta)\,\ell(\theta)$.
  - Quite funny we call that Maximum Aposteriori (MAP) estimator!!!

The MAP estimator is since it is often simpler to calculate given that

$$\arg\max_{\theta} \pi(\theta|x) = \arg\max_{\theta} f(x|\theta)\,\pi(\theta)$$

- Given that the normalization factor is a constant

Cinvestav

Outline

Cinvestav

Properties

First

- **MAP** estimation "pulls" the estimate toward the prior.

Second

- The more focused our prior belief, the larger the pull toward the prior.

Example

- If $\alpha = \beta =$equal to large value
  - It will make the MAP estimate to move closer to the prior.

Properties

First

- **MAP** estimation “pulls” the estimate toward the prior.

Second

- The more focused our prior belief, the larger the pull toward the prior.

Example

- If $\alpha = \beta =$equal to large value
  - It will make the MAP estimate to move closer to the prior.

Properties

First

- **MAP** estimation "pulls" the estimate toward the prior.

Second

- The more focused our prior belief, the larger the pull toward the prior.

Example

- If $\alpha = \beta =$equal to large value
  - It will make the MAP estimate to move closer to the prior.

Properties

Third

- In the expression we derived for $\widehat{q}_{MAP}$, the parameters α and β play a "smoothing" role vis-a-vis the measurement n_{PRI}.

Fourth

- Since we referred to q as the parameter to be estimated, we can refer to α and β as the hyper-parameters in the estimation calculations.

Properties

Third

- In the expression we derived for $\widehat{q}_{MAP}$, the parameters α and β play a "smoothing" role vis-a-vis the measurement n_{PRI}.

Fourth

- Since we referred to q as the parameter to be estimated, we can refer to α and β as the hyper-parameters in the estimation calculations.

Beyond simple derivation

In the previous technique

- We took an logarithm of the likelihood $\times$ the prior to obtain a function that can be derived in order to obtain each of the parameters to be estimated.

What if we cannot derive?

- For example when we have something like $|\theta_i|$.

We can try the following

- Expectation Maximization + MAP to be able to estimate the sought parameters.

Beyond simple derivation

In the previous technique

- We took an logarithm of the likelihood $\times$ the prior to obtain a function that can be derived in order to obtain each of the parameters to be estimated.

What if we cannot derive?

- For example when we have something like $|\theta_i|$.

We can try the following

- Expectation Maximization + MAP to be able to estimate the sought parameters.

Beyond simple derivation

In the previous technique

- We took an logarithm of the likelihood $\times$ the prior to obtain a function that can be derived in order to obtain each of the parameters to be estimated.

What if we cannot derive?

- For example when we have something like $|\theta_i|$.

We can try the following

- Expectation Maximization $+$ MAP to be able to estimate the sought parameters.

Outline

Cinvestav

Imagine an action space and $a \in \mathcal{A}$

For example

- In estimation problems, $\mathcal{A}$ is the set of real numbers and a is a number, say $a = 2$ is adopted as an estimator of $\theta \in \Theta$.

Another One

- In testing problems, the action space is $\mathcal{A} = \{accept, reject\}$

Cinvestav

Imagine an action space and $a \in \mathcal{A}$

For example

- In estimation problems, $\mathcal{A}$ is the set of real numbers and a is a number, say $a = 2$ is adopted as an estimator of $\theta \in \Theta$.

Another One

- In testing problems, the action space is $\mathcal{A} = \{accept, reject\}$

Cinvestav

Everytime you make a decision you have a Loss

Actually

- Statisticians are pessimistic creatures that replaced nicely coined term utility to a more somber term loss!!!

How do we denote such losses?

- A classic one $L(\theta, a)$
  - representing the payoff by a decision maker (statistician) if he takes any action $a \in \mathcal{A}$ in certina state of nature θ

Cinvestav

Everytime you make a decision you have a Loss

Actually

- Statisticians are pessimistic creatures that replaced nicely coined term utility to a more somber term loss!!!

How do we denote such losses?

- A classic one $L(\theta, a)$
  - representing the payoff by a decision maker (statistician) if he takes any action $a \in \mathcal{A}$ in certina state of nature θ

Cinvestav

Outline

Cinvestav

Examples

Squared Error Loss

$$L(\theta, a) = (\theta - a)^2$$

Absolute Loss

$$L(\theta, a) = |\theta - a|$$

0-1 Loss example

$$L(\theta, a) = I[|\theta - a| > m]$$

Cinvestav

Examples

Squared Error Loss

$$L(\theta, a) = (\theta - a)^2$$

Absolute Loss

$$L(\theta, a) = |\theta - a|$$

0-1 Loss example

$$L(\theta, a) = I\left[|\theta - a| > m\right]$$

Examples

Squared Error Loss

$$L(\theta, a) = (\theta - a)^2$$

Absolute Loss

$$L(\theta, a) = |\theta - a|$$

0-1 Loss example

$$L(\theta, a) = I\left[|\theta - a| > m\right]$$

Clearly the easiest mathematically SEL

Additionally, it is linked with

$$E_{X|\theta}\left[\theta-\delta\left(X\right)\right]^{2}=Var\left(\delta\left(X\right)\right)+\left[bias\left(\delta\left(X\right)\right)\right]^{2}$$

- Where $bias\left(\delta\left(X\right)\right)=E_{X|\theta}\left[\delta\left(X\right)\right]-\theta$

In another example

The median, m, of random variable X is defined as

$$
\begin{aligned}
P\left(X \geq m\right) &\geq \frac{1}{2},\\
P\left(X \leq m\right) &\leq \frac{1}{2}
\end{aligned}
$$

In another example

The median, m, of random variable X is defined as

$$P\left(X \geq m\right) \geq \frac{1}{2},$$
$$P\left(X \leq m\right) \leq \frac{1}{2}$$

Assuming the absolute loss

$$\begin{aligned}\varphi\left(a\right) =& E_{\theta|X}\left[\left|\theta - a\right|\right]\\ =& \int_{\theta\geq a}\left(\theta - a\right)\pi\left(\theta|X\right)d\theta + \int_{\theta\leq a}\left(a-\theta\right)\pi\left(\theta|X\right)d\theta\\ =& \int_{a}^{\infty}\left(\theta - a\right)\pi\left(\theta|X\right)d\theta + \int_{\infty}^{a}\left(a-\theta\right)\pi\left(\theta|X\right)d\theta\end{aligned}$$

In another example

The median, m, of random variable X is defined as

$$P\left(X \geq m\right) \geq \frac{1}{2},$$
$$P\left(X \leq m\right) \leq \frac{1}{2}$$

Assuming the absolute loss

$$\begin{aligned}\varphi\left(a\right) =& E_{\theta|X}\left[\left|\theta - a\right|\right] \\ =& \int_{\theta \geq a} \left(\theta - a\right)\pi\left(\theta|X\right)d\theta + \int_{\theta \leq a}\left(a - \theta\right)\pi\left(\theta|X\right)d\theta \\ =& \int_{a}^{\infty} \left(\theta - a\right)\pi\left(\theta|X\right)d\theta + \int_{\infty}^{a}\left(a - \theta\right)\pi\left(\theta|X\right)d\theta\end{aligned}$$

In another example

The median, m, of random variable X is defined as

$$P\left(X \geq m\right) \geq \frac{1}{2},$$
$$P\left(X \leq m\right) \leq \frac{1}{2}$$

Assuming the absolute loss

$$\begin{aligned}\varphi\left(a\right) =& E_{\theta|X}\left[\left|\theta - a\right|\right]\\ =& \int_{\theta \geq a} \left(\theta - a\right) \pi\left(\theta|X\right) d\theta + \int_{\theta \leq a} \left(a - \theta\right) \pi\left(\theta|X\right) d\theta\\ =& \int_{a}^{\infty} \left(\theta - a\right) \pi\left(\theta|X\right) d\theta + \int_{\infty}^{a} \left(a - \theta\right) \pi\left(\theta|X\right) d\theta\end{aligned}$$

Then

Using the following equivalence

$$\frac{\partial}{\partial x}\left[\int_{f(x)}^{g(x)} \phi\left(x,t\right)dt\right] = \int_{f(x)}^{g(x)} \frac{\partial}{\partial x}\phi\left(x,t\right)dt + \phi\left(x, g\left(x\right)\right)\frac{\partial g\left(x\right)}{\partial x} - \dots$$

$$\phi\left(x, f\left(x\right)\right)\frac{\partial f\left(x\right)}{\partial x}$$

Then

$$\frac{\partial \varphi\left(a\right)}{\partial a} = -\int_{a}^{\infty} \pi\left(\theta|X\right)d\theta + 0 - 0 + \int_{\infty}^{a} \pi\left(\theta|X\right)d\theta + 0 - 0$$

Then

Using the following equivalence

$$\frac{\partial}{\partial x}\left[\int_{f(x)}^{g(x)} \phi\left(x,t\right)dt\right] = \int_{f(x)}^{g(x)} \frac{\partial}{\partial x}\phi\left(x,t\right)dt + \phi\left(x,g\left(x\right)\right)\frac{\partial g\left(x\right)}{\partial x} - \ldots$$
$$\phi\left(x,f\left(x\right)\right)\frac{\partial f\left(x\right)}{\partial x}$$

Then

$$\frac{\partial \varphi\left(a\right)}{\partial a} = -\int_{a}^{\infty} \pi\left(\theta|X\right)d\theta + 0 - 0 + \int_{\infty}^{a} \pi\left(\theta|X\right)d\theta + 0 - 0$$

Cinvestav

Therefore

We have then

$$\frac{\partial \varphi (a)}{\partial a} = -P_{\theta|X} (\theta \geq a) + P_{\theta|X} (\theta \leq a) = 0$$

The value of a for which $P_{\theta|X} (\theta \geq a) = P_{\theta|X} (\theta \leq a)$ is the median

- Since $\frac{\partial^2 \varphi(a)}{\partial a^2} = 2\pi (a|X) > 0$ by the Fundamental theorem of calculus

Cinvestav

Therefore

We have then

$$\frac{\partial \varphi(a)}{\partial a} = -P_{\theta|X}(\theta \geq a) + P_{\theta|X}(\theta \leq a) = 0$$

The value of a for which $P_{\theta|X}(\theta \geq a) = P_{\theta|X}(\theta \leq a)$ is the median

- Since $\frac{\partial^2 \varphi(a)}{\partial a^2} = 2\pi(a|X) > 0$ by the Fundamental theorem of calculus

Finally

The Median Minimize

$$\varphi(a)$$

Outline

Cinvestav

Bayesian Expected Loss

Definition

- Bayesian expected loss is the expectation of the loss function with respect to posterior measure,

$$\rho(a, \pi) = E_{\theta|X}\left[L(a, \theta)\right] = \int_{\Theta} L(\theta, a)\, \pi(\theta|x)\, d\theta$$

Here, we have an important principle

- Referring to the less possible loss!!!

Bayesian Expected Loss

Definition

- Bayesian expected loss is the expectation of the loss function with respect to posterior measure,

$$\rho\left(a,\pi\right)=E_{\theta|X}\left[L\left(a,\theta\right)\right]=\int_{\Theta}L\left(\theta,a\right)\pi\left(\theta|x\right)d\theta$$

Here, we have an important principle

- Referring to the less possible loss!!!

Cinvestav

The Expected Loss Principle

Definition

- In comparing two actions $a_1 = \delta_1(X)$ and $a_2 = \delta_2(X)$, after data X had been observed, preferred action is the one for which the posterior expected loss is smaller.

Therefore

- An action a^* that minimizes the posterior expected loss is called Bayes action.

The Expected Loss Principle

Definition

- In comparing two actions $a_1 = \delta_1(X)$ and $a_2 = \delta_2(X)$, after data X had been observed, preferred action is the one for which the posterior expected loss is smaller.

Therefore

- An action a^* that minimizes the posterior expected loss is called Bayes action.

Cinvestav

Outline

Cinvestav

Example

If the loss is squared error

- The Bayes action a^* is found by minimizing

$$\varphi(a) = E_{\theta|X}(\theta - a)^2 = a^2 - 2E_{\theta|X}[\theta]\,a + E_{\theta|X}\theta^2$$

Then, we want $\varphi'(a) = 0$

- Solving for it, we have $a = E_{\theta|X}[\theta]$

Additionally

- $\varphi''(a) < 0$ then $a^* = E_{\theta|X}[\theta]$ is a Bayesian Action.

Example

If the loss is squared error

- The Bayes action a^* is found by minimizing

$$\varphi(a) = E_{\theta|X}(\theta - a)^2 = a^2 - 2E_{\theta|X}[\theta]\,a + E_{\theta|X}\theta^2$$

Then, we want $\varphi'(a) = 0$

- Solving for it, we have $a = E_{\theta|X}[\theta]$

Additionally

- $\varphi''(a) < 0$ then $a^* = E_{\theta|X}[\theta]$ is a Bayesian Action.

Example

If the loss is squared error

- The Bayes action a^* is found by minimizing

$$\varphi(a) = E_{\theta|X}(\theta - a)^2 = a^2 - 2E_{\theta|X}[\theta]\,a + E_{\theta|X}\theta^2$$

Then, we want $\varphi'(a) = 0$

- Solving for it, we have $a = E_{\theta|X}[\theta]$

Additionally

- $\varphi''(a) < 0$ then $a^* = E_{\theta|X}[\theta]$ is a Bayesian Action.

Outline

Cinvestav

Given $X \in \{P_\theta, \theta \in \Theta\}$

A family which is indexed by a parameter (random variable) θ

- Here, we change our Bayesian hat to the frequentist one

This allows to make inferences about θ

- A solution is a decision procedure (decision rule) $\delta(x)$, that identifies particular inference for each value of x that can be observed.

Given $X \in \{P_\theta, \theta \in \Theta\}$

A family which is indexed by a parameter (random variable) θ

- Here, we change our Bayesian hat to the frequentist one

This allows to make inferences about θ

- A solution is a decision procedure (decision rule) $\delta(x)$, that identifies particular inference for each value of x that can be observed.

Cinvestav

A be the class of all possible realizations of $\delta(x)$, i.e. actions

The Loss function $L(\theta, a)$ maps $\Theta \times \mathcal{A} \longrightarrow \mathbb{R}$

- Defining a cost to the statistician when he takes the action a and the true value of the parameter is θ.

Then we can define a decision function called Risk

$$R(\theta, \delta) = E_{X|\theta}\left[L(\theta|\delta(X))\right] = \int_{\mathcal{X}} L(\theta|\delta(X)) f(x|\theta)\, dx$$

- A frequentist risk on the performance of δ.

$\mathcal{A}$ be the class of all possible realizations of $\delta(x)$, i.e. actions

The Loss function $L(\theta, a)$ maps $\Theta \times \mathcal{A} \longrightarrow \mathbb{R}$

- Defining a cost to the statistician when he takes the action a and the true value of the parameter is θ.

Then we can define a decision function called Risk

$$R(\theta, \delta) = E_{X|\theta}[L(\theta|\delta(X))] = \int_{\mathcal{X}} L(\theta|\delta(X)) f(x|\theta)\, dx$$

- A frequentist risk on the performance of δ.

Therefore

Since the risk function is defined as an average loss with respect to a sample space

- it is called the frequentist risk.

Let $\mathcal{D}$ be the collection of all measurable decision rules

- There are several ways for assigning the preference among the rules in $\mathcal{D}$.

Cinvestav

Therefore

Since the risk function is defined as an average loss with respect to a sample space

- it is called the frequentist risk.

Let $\mathcal{D}$ be the collection of all measurable decision rules

- There are several ways for assigning the preference among the rules in $\mathcal{D}$.

Furthermore

Some of them are

- The Minimax Principle
- Γ-minimax Principle
- Minimax Principle
- etc

Furthermore

Some of them are

- The Minimax Principle
- Γ-minimax Principle
- Minimax Principle
- etc

The one we are interested is

- The Bayes principle

Furthermore

Some of them are

- The Minimax Principle
- Γ-minimax Principle
- Minimax Principle
- etc

The one we are interested is

- The Bayes principle

Furthermore

Some of them are

- The Minimax Principle
- Γ-minimax Principle
- Minimax Principle
- etc

The one we are interested is

- The Bayes principle

Furthermore

Some of them are

- The Minimax Principle
- Γ-minimax Principle
- Minimax Principle
- etc

The one we are interested is

- The Bayes principle

Under the Bayes principle

Bayes risk

$$r(\pi, \delta) = \int R(\theta, \delta)\, \pi(d\theta) = E_\theta R(\theta, \delta)$$

Where there is a δ_π, called Bayes rule, minimizing the risk

$$\delta_\pi = \arg \inf_{\delta \in \mathcal{D}} r(\pi, \delta)$$

Bayes risk of the prior distribution π (Bayes envelope function) is

$$r(\pi) = r(\pi, \delta_\pi)$$

Under the Bayes principle

Bayes risk

$$r(\pi, \delta) = \int R(\theta, \delta)\, \pi(d\theta) = E_\theta R(\theta, \delta)$$

Where there is a δ_π , called Bayes rule, minimizing the risk

$$\delta_\pi = \arg \inf_{\delta \in \mathcal{D}} r(\pi, \delta)$$

Bayes risk of the prior distribution π (Bayes envelope function) is

$$r(\pi) = r(\pi, \delta_\pi)$$

Under the Bayes principle

Bayes risk

$$r(\pi, \delta) = \int R(\theta, \delta)\, \pi(d\theta) = E_\theta R(\theta, \delta)$$

Where there is a δ_π , called Bayes rule, minimizing the risk

$$\delta_\pi = \arg \inf_{\delta \in \mathcal{D}} r(\pi, \delta)$$

Bayes risk of the prior distribution π (Bayes envelope function) is

$$r(\pi) = r(\pi, \delta_\pi)$$

Bayes Envelope Function Definition

Definition

- The Bayes Envelope is the maximal reward rate a player could achieve had he known in advance the relative frequencies of the other players.

In particular, we define the following function as

$$r(\pi,\delta) = E_{\theta}\left[E_{X|\theta}\left[L(\theta,\delta(X))\right]\right]$$

Therefore the Bayes action as Bayes Rules looks like

$$\delta^{*}(x) = \arg\min_{\delta\in D} r(\pi,\delta)$$

Bayes Envelope Function Definition

Definition

- The Bayes Envelope is the maximal reward rate a player could achieve had he known in advance the relative frequencies of the other players.

In particular, we define the following function as

$$r(\pi, \delta) = E_\theta \left[E_{X|\theta} \left[L(\theta, \delta(X)) \right] \right]$$

Therefore the Bayes action as Bayes Rules looks like

$$\delta^*(x) = \arg\min_{\delta \in D} r(\pi, \delta)$$

Bayes Envelope Function Definition

Definition

- The Bayes Envelope is the maximal reward rate a player could achieve had he known in advance the relative frequencies of the other players.

In particular, we define the following function as

$$r(\pi, \delta) = E_{\theta}\left[E_{X|\theta}\left[L(\theta, \delta(X))\right]\right]$$

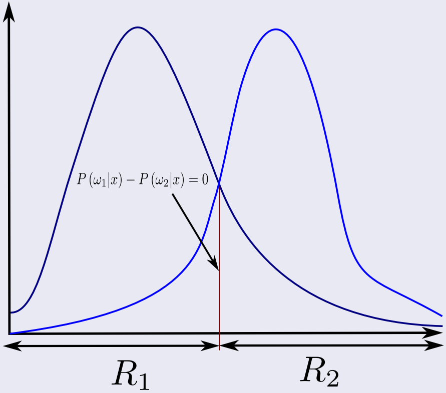
Therefore the Bayes action as Bayes Rules looks like

$$\delta^{*}(x) = \arg\min_{\delta \in \mathcal{D}} r(\pi, \delta)$$

Actually

A classic Bayes Rule

- The Naive Bayes Rules for classification using Gaussian's for classification



Outline

Cinvestav

The Fubini's Theorem (Informal Version)

Theorem

- Suppose X and Y are σ-finite measure spaces, and suppose that $X \times Y$ is given the product measure:

$$(\mu \times \nu)(E) = \inf \left\{ \sum_{j=1}^{\infty} \mu(A_j)\, \nu(B_j) \,|\, E \subset \cup_{j=1}^{\infty} A_j \times B_j \right\}$$

With any non-negative $\mu \times \nu$-measurable function f, then

$$\int_{X \times Y} f(x,y)\, d(\mu \times \nu)(x,y) = \int_Y \left(\int_X f(x,y)\, d\mu(x) \right) d\nu(y)$$

The Fubini's Theorem (Informal Version)

Theorem

- Suppose X and Y are σ-finite measure spaces, and suppose that $X \times Y$ is given the product measure:

$$(\mu \times \nu)(E) = \inf \left\{ \sum_{j=1}^{\infty} \mu(A_j)\,\nu(B_j) \,|E \subset \cup_{j=1}^{\infty} A_j \times B_j \right\}$$

With any non-negative $\mu \times \nu$-measurable function f, then

$$\int_{X\times Y} f(x,y)\, d(\mu \times \nu)(x,y) = \int_{Y} \left(\int_{X} f(x,y)\, d\mu(x) \right) d\nu(y)$$

Cinvestav

Implications with the Expected Value

We have by the Fubini's Theorem

$$
\begin{aligned}
r(\pi,\delta) &= E_\theta\left[E_{X|\theta}\left[L(\theta,\delta(X))\right]\right]\\
&= E_X\left[E_{\theta|X}\left[L(\theta,\delta(X))\right]\right]
\end{aligned}
$$

Where the posterior expected loss

$$
\rho(\pi,\delta) = E_{\theta|X}\left[L(\theta,\delta(X))\right]
$$

Implications with the Expected Value

We have by the Fubini's Theorem

$$r(\pi, \delta) = E_{\theta}\left[E_{X|\theta}\left[L(\theta, \delta(X))\right]\right]$$
$$= E_{X}\left[E_{\theta|X}\left[L(\theta, \delta(X))\right]\right]$$

Where the posterior expected loss

$$\rho(\pi, \delta) = E_{\theta|X}\left[L(\theta, \delta(X))\right]$$

Cinvestav

Therefore

$r(\pi, \delta)$ is minimized for any fixed x

- When $\rho(\pi, \delta)$ is minimized, for any fixed x, $\delta_B(x) = a^*(x)$ where * represent the optimal action.

Basically

- This result links the conditional Bayesian and decision theoretic frequentist inference:
  - The frequentist Bayes rule conditional on X is the Bayes action.

Therefore

$r(\pi, \delta)$ is minimized for any fixed x

- When $\rho(\pi, \delta)$ is minimized, for any fixed x, $\delta_B(x) = a^*(x)$ where * represent the optimal action.

Basically

- This result links the conditional Bayesian and decision theoretic frequentist inference:
  - The frequentist Bayes rule conditional on X is the Bayes action.

What happens when we have the Squared Loss?

The Bayes rule is the posterior expectation

$$\delta_B(x) = \frac{\int_\Theta \theta f(x|\theta)\,\pi(\theta)\,d\theta}{\int_\Theta f(x|\theta)\,\pi(\theta)\,d\theta}$$

Not only that, in the case of

$$L(\theta, a) = w(\theta)(\theta - a)^2$$

What happens when we have the Squared Loss?

The Bayes rule is the posterior expectation

$$\delta_B(x) = \frac{\int_\Theta \theta f(x|\theta)\,\pi(\theta)\,d\theta}{\int_\Theta f(x|\theta)\,\pi(\theta)\,d\theta}$$

Not only that, in the case of

$$L(\theta, a) = w(\theta)(\theta - a)^2$$

We have

The following Bayes Rule

$$\delta_B(x) = \frac{\int_\Theta w(\theta)\,\theta f(x|\theta)\,\pi(\theta)\,d\theta}{\int_\Theta w(\theta)\,f(x|\theta)\,\pi(\theta)\,d\theta}$$

Furthermore

According to a Bayes principle

- A rule $\delta_1(X)$ is preferred to $\delta_2(X)$ if $r(\pi, \delta_1) < r(\pi, \delta_2)$

The frequentists use Bayes principle

- to compare frequentist risks of the rules $R(\theta, \delta_1)$ and $R(\theta, \delta_2)$.

Furthermore

According to a Bayes principle

- A rule $\delta_1(X)$ is preferred to $\delta_2(X)$ if $r(\pi, \delta_1) < r(\pi, \delta_2)$

The frequentists use Bayes principle

- to compare frequentist risks of the rules $R(\theta, \delta_1)$ and $R(\theta, \delta_2)$.

Analysis of frequentist risk

It leads to various concepts as

1. minimaxity,
2. admissibility,
3. unbiasedness,
4. equivariance,
5. etc.

D. V. Lindley and L. D. Phillips, "Inference for a bernoulli process (a bayesian view)," *The American Statistician*, vol. 30, no. 3, pp. 112–119, 1976.

C. Aschwanden, "Not even scientists can easily explain p-values." https://fivethirtyeight.com/features/not-even-scientists-can-easily-explain-p-values.

Online; accessed 24 Noviembre 2015.

K. P. F.R.S., "X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 50, no. 302, pp. 157–175, 1900.

P. Bickel and K. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics*.
No. v. 1 in Mathematical Statistics: Basic Ideas and Selected Topics, Prentice Hall, 2001.

R. A. Fisher, "On the mathematical foundations of theoretical statistics," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 222, no. 594-604, pp. 309–368, 1922.

G. Casella and R. L. Berger, *Statistical inference*, vol. 2.
Duxbury Pacific Grove, CA, 2002.

S. M. Kay, *Fundamentals of statistical signal processing*.
Prentice Hall PTR, 1993.