Bayesian inference for time series

Wen Yu

Departamento de Control Automático
CINVESTAV-IPN

Probability Theory

The probability of the variable A takes a is

$$0 \leq P\left(A = a\right) \leq 1$$

Alternatives ->"add"

$$P\left(A = a_1 \text{ or } A = a_2\right) = P\left(A = a_1\right) + P\left(A = a_2\right)$$

Normalisation

$$\sum_{\text{all possible } a} P\left(A = a\right) = 1$$

Joint probability

$$P\left(A = a, B = b\right) \text{ the probability that both } A = a \text{ and } B = b \text{ occur}$$

Probability Theory

Conditional probability

$P(A = a \mid B = b)$ the probability that $A = a$ occurs given the knowledge B

Product rule

$$\begin{gathered} P(A = a, B = b) \\ = P(A = a)\, P(B = b \mid A = a) \\ = P(B = b)\, P(A = a \mid B = b) \end{gathered}$$

Independence, iff A and B are independent:

$$\begin{gathered} P(A = a \mid B = b) = P(A = a) \\ P(B = b \mid A = a) = P(B = b) \\ P(A = a, B = b) = P(A = a)\, P(B = b) \end{gathered}$$

Continuous variables, "Sum" $\rightarrow$ "Integral", the probability that X lies between x and $(x + dx)$ is $p(x)dx$, $p(x)$ is a probability density function

$$P\left(x_1 \leq X \leq x_2\right) = \int_{x_1}^{x_2} p(x)dx$$
$$\int_{-\infty}^{+\infty} p(x)dx = 1$$
$$\int_{-\infty}^{+\infty} p(x, y)dy = p(x)$$

Probability Theory

Expectations: averages over a time series X

$$E\left[A\right]=\sum_a P\left(A=a\right)A$$
$$E\left[x\right]=\int_{-\infty}^{+\infty}xp(x)dx$$

If X and Y are independent

$$E\left[x\times y\right]=E\left[x\right]\times E\left[y\right]$$

Bayes' theorem, Bayes' law, Bayes' rule

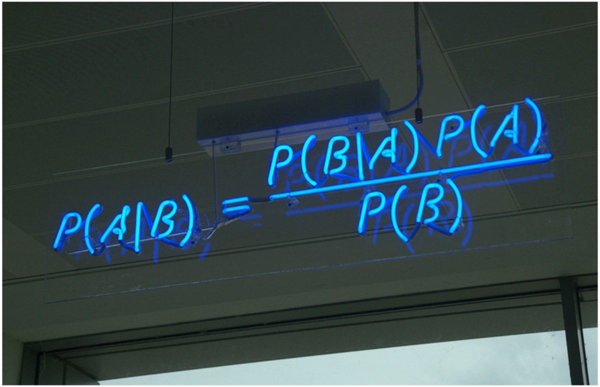


Figure: Blue neon at the offices of Autonomy in Cambridge

Bayes' theorem

10 red balls and 5 blue balls in one bag

- "Forward" probability: pick up 1 ball, the probability of red ball is $\frac{10}{15}$, the probability of blue ball is $\frac{3}{15}$.
- "Reverse" probability: there are 15 red and 15 blue balls in one bag, pick up 2 balls, the probability of the color are similar
- "Reverse" probability: there are two bags: 10 red + 5 blue, 5 red + 10 blue. Randomly select one bag, pick up 12 balls and they are 8 red + 4 blue, the probability of bag A ?

Bayes' theorem

Relates current probability to prior probability (degree of belief to account for evidence)

$$P(A \mid B) = \frac{P(B \mid A) P(A)}{P(B)}$$

where

A and B are events,

$P(A)$ and $P(B)$ are the probabilities of A and B without regard to one other

$P(A \mid B)$ is the conditional probability (probability of A given that B is true)

$P(B \mid A)$ is the probability of B given that A is true.

Bayes' theorem

- $P(A)$ is the probability of cause A before effect B is known, it is called the *prior probability* of A.
- $P(A \mid B)$ is the probability of cause A after effect B is known, and it is called the *posterior probability* of A.
- $P(B \mid A)$ is the likelihood of B when A

Bayes' theorem

Proof.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad P(B \mid A) = \frac{P(B \cap A)}{P(A)}$$
$$P(B)\,P(A \mid B) = P(A)\,P(B \mid A) = P(A \cap B)$$
$$P(A \mid B) = \frac{P(A)P(B|A)}{P(B)}$$

$\square$

Bayes' theorem

$$P(A \mid B) = \frac{P(B|A)P(A)}{P(B)}$$
$$= \frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|\neg A)P(\neg A)}$$
$$= \frac{P(B|A_i)P(A_i)}{\sum_i P(B|A_i)P(A_i)}$$

Bayes' theorem- application

Example

A patient comes into the doctor's office. A blood test for cancer is given and the test result is POS. The test is 95% accurate. Only 0.8% of the people in the U.S. have this form of cancer. What is probability of haveing cancer of this patient $P(cancer \mid POS)$?

$$
\begin{aligned}
&P(cancer) = 0.008, P(POS \mid cancer) = 0.95 \\
&P(POS) = P(POS, cancer) + P(POS, \neg cancer) \\
&= P(POS \mid cancer) P(cancer) + P(POS \mid \neg cancer) P(\neg cancer) \\
&= 0.95 \times 0.008 + 0.05 \times 0.992
\end{aligned}
$$

The probability of haveing cancer is

$$
P(cancer \mid POS) = \frac{P(POS \mid cancer) P(cancer)}{P(POS)} = 0.133
$$

This patient has a 13% chance of having cancer.

Bayes' theorem- application

Spelling check

"lates" is "late" or " latest "? we need probability, not rules

Gived a written word W, we need to find a correct word C in all correct words,

$$\max P(C \mid W)$$

By Bayes' theorem

$$\max P(C \mid W) = \max \frac{P(W \mid C)\, P(C)}{P(W)}$$

Since we can wirte any word, $P(W)$ is the same always

$$\max P(C \mid W) = \max P(W \mid C)\, P(C)$$

Bayes' theorem- application

- $P(C)$ is the probability of all English words. It is the language model. For example $P(''the'')$ is very high, $P(''teh'')$ is very low.
- $P(W \mid C)$ is the probability: the person want input C, however he types W. It is error model.
- Language Model: $P(C)$ can be obtained by extracting a big text file
- Error Model: edit distance, in 80-95% spelling error have "edit distance"=1

Bayes' theorem- countinous time

For continuous random variables X and Y , Bayes' Theorem is formulated in terms of densities:

$$P(y \mid x) = \frac{P(x, y)}{P(x)} = \frac{P(x \mid y)\, P(y)}{P(x)}$$

where
$P(x, y)$ is the joint probability distribution of X and Y
$P(x)$ is the prior probability of X, $P(x \mid y)$ is the likelihood of Y when $X = x$
$P(y \mid x)$ is the posterior probability of Y when $X = x$

$$P(y \mid x) = \frac{P(x \mid y)\, P(y)}{\int_{-\infty}^{\infty} P(x \mid y)\, P(y)\, dy}$$

$$P(A \mid B) = \frac{P(B|A)P(A)}{P(B)}$$

$$posterior = \frac{prior \times linkhood}{evidence}$$

Naive Bayesian

For input-ouput model

$$y = f(x) = P(Y \mid X) = \frac{P(X \mid Y_i)\, P(Y_i)}{\sum_i P(X \mid Y_i)\, P(Y_i)}$$

where $X = [x_1 \cdots x_n]$, Y_i has i values c_i

Assume that each feature x_i is conditionally independent of every other feature

$$P(X \mid Y = c_i) = P(x_1 \cdots x_n \mid Y = c_i) = \Pi_j P(X = x_j \mid Y = c_i)$$

So

$$P(Y = c_i \mid X) = \frac{\Pi_j P(X = x_j \mid Y = c_i)\, P(Y = c_i)}{\sum_i P(Y = c_i)\, \Pi_j P(X = x_j \mid Y = c_i)}$$

We want to the best value of Y, i.e.,

$$y = f(x) = \arg\max_{c_i} \frac{\Pi_j P(X = x_j \mid Y = c_i)\, P(Y = c_i)}{\sum_i P(Y = c_i)\, \Pi_j P(X = x_j \mid Y = c_i)}$$

the denominator for all c_i is the same

$$y = f(x) = \arg\max P(Y = c_i)\, \Pi_j P(X = x_j \mid Y = c_i)$$

Naive Bayesian

Naive Bayes Classifier: select the most likely classification V given the attribute values $x_1 \cdots x_n$

$$V = \max_{a_j \in A} P\left(a_i \mid x_1 \cdots x_n\right) = \max_{a_j \in A} P\left(a_j\right) \prod_{i=1}^{n} P\left(x_i \mid a_j\right)$$

It is equivalent to the joint probability model, because $P\left(A \mid B\right) = \frac{P(B|A)P(A)}{P(B)}$ the denominator does not depend on A, it is effectively constant.

Maximum-likelihood estimation (MLE)

For independent and identically distributed samples, $x_1 \cdots x_n$, coming from unknown probability density function P_0. Assume P_0 belongs to a certain family of distributions $\{\, P(x|\,\theta), \theta \in \Theta \,\}$. where θ is parameters for this family. The parametric model is

$$P_0 = P(x|\,\theta_0)$$

The joint density function for all observations

$$P\left(x_1 \cdots x_n \middle|\, \theta\right) = \prod_{i=1}^{n} P\left(x_i \middle|\, \theta\right)$$

When θ can be changed and x_i is fixed, the $P\left(x_1 \cdots x_n \middle|\, \theta\right)$ is called likelihood,

$$L\left(\theta; x_1 \cdots x_n\right) = P\left(x_1 \cdots x_n \middle|\, \theta\right)$$

where $x_1 \cdots x_n$ are observations

Maximum-likelihood estimation (MLE)

In practice, we use the log-likelihood:

$$\ln L\left(\theta; x_1 \cdots x_n\right) = \sum_{i=1}^{n} \ln P\left(x_1 \cdots x_n \middle| \theta\right), \textit{or} \quad \hat{L} = \frac{1}{n} \ln L$$

$\hat{L}$ estimates the expected log-likelihood of a single observation in the model.

MLE

$$\theta^{*} = \max_{\theta} \hat{L}$$

MLE vs. Bayesian estimator

Bayes' theorem:

$$P\left(\theta;\middle|\, x_1\cdots x_n\right)=\frac{P\left(x_1\cdots x_n\middle|\,\theta\right)P\left(\theta\right)}{P\left(x_1\cdots x_n\right)}$$

Bayesian estimator is obtained by maximizing $P\left(x_1\cdots x_n\middle|\,\theta\right)P\left(\theta\right)$ with respect to θ. If we further assume that the prior $P\left(\theta\right)$ is a uniform distribution, the Bayesian estimator is obtained by maximizing the likelihood function $P\left(x_1\cdots x_n\middle|\,\theta\right)$.
Thus the Bayesian estimator coincides with MLE for a uniform prior distribution$P\left(\theta\right)$.

MLE - example

For the normal distribution $N\left(\mu, \sigma^2\right)$, the probability density function is

$$P(x|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

the likelihood of n independent identically distributed samples is

$$L\left(\theta; x_1 \cdots x_n\right) = P\left(x_1 \cdots x_n \middle| \mu, \sigma^2\right) = \prod_{i=1}^{n} P\left(x_i \middle| \mu, \sigma^2\right) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp \Bigg[$$

$$\ln L = -\frac{n}{2} \ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} \sum \left(x_i - \mu\right)^2$$

MLE

Because

$$\exp\left[-\frac{\sum(x-\mu)^2}{2\sigma^2}\right] = \exp\left[-\frac{\sum(x-\bar{x})^2 + n\sum(\bar{x}-\mu)^2}{2\sigma^2}\right]$$

$$\frac{\partial}{\partial\mu}\ln L = 0$$
$$\mu = \frac{1}{n}\sum x_i$$

$$\frac{\partial}{\partial\sigma}\ln L = 0$$
$$\sigma^2 = \frac{1}{n}\sum(x_i-\mu)^2$$

MLE

If $x_1 \cdots x_n$ have the same probability density function, such as Bernoulli distribution

$$p(x,\theta) = \begin{cases} (1+\theta)\, x^{\theta} & 0 < x < 1 \\ 0 & otherwise \end{cases}$$

Construct the likelihood function

$$L = \begin{cases} (1+\theta)^n \prod_{i=1}^{n} x_i^{\theta} & 0 < x < 1 \\ 0 & otherwise \end{cases}$$

and

$$\ln L = \begin{cases} n \ln(1+\theta) + \theta \sum \ln x_i & 0 < x_i < 1 \\ 0 & otherwise \end{cases}$$

MLE

$$\frac{\partial}{\partial\theta}\ln L = \frac{n}{1+\theta} + \sum \ln x_i = 0$$

MLE of θ is

$$\hat{\theta} = \frac{n}{\sum \ln x_i} - 1$$

Normal model
A random variable X is said to be normally distributed with mean θ and variance σ^2, if the density of X is

$$p\left(x \mid \theta, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(x-\theta\right)^2\right] \tag{1}$$

If $\left(x_1 \ldots x_n \mid \theta, \sigma^2\right) \sim i.i.d.normal\left(\theta, \sigma^2\right)$, then the sampling density is

$$\begin{aligned} p\left(x_1 \ldots x_n \mid \theta, \sigma^2\right) &= \prod p\left(x_i \mid \theta, \sigma^2\right) \\ &= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\tfrac{1}{2\sigma^2} \sum \left(x_i - \theta\right)\right] \end{aligned} \tag{2}$$

The problem of the model training is to estimate the two parameters, θ and σ^2, from the data $x_1 \ldots x_n$.

Exponential model
The inverse exponential distribution

$$p(\beta) = \frac{1}{\beta} e^{(-X/\beta)} \tag{3}$$

where $X = [x_1 \ldots x_n]$, $\beta > 0$. This distribution can be calculated with the cumulative of the previous distribution with $\lambda = 1/\beta$,

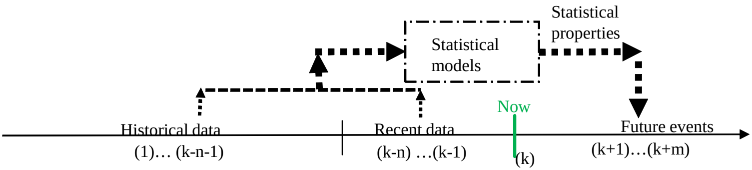
$$p(\lambda) = \lambda e^{-\lambda X} \tag{4}$$

The mean and variance of the exponential distribution can be represented as

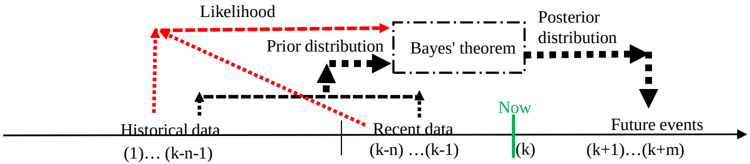
$$\mu = \beta, \quad \sigma^2 = \beta^2 \tag{5}$$

The problem of the model training is to estimate the parameter β from the data $x_1 \ldots x_n$.

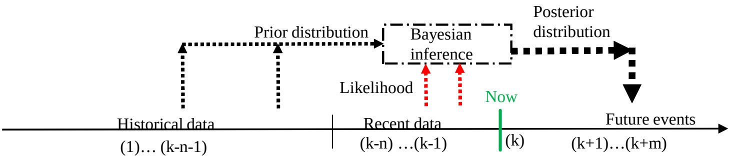
Bayesian inference for time series



Bayesian inference for time series



Bayesian inference for time series



Bayes' theorem in model form

Bayes' theorem in model form is written as:

$$p\,(parameter \mid data) = \frac{p(data|paratmeter)p(parameter)}{p(data)} \\ = \frac{p(x|\theta)p(\theta)}{p(x)} \tag{6}$$

- $p\,(parameter)$ is the prior distribution. It represents our beliefs about the true value of the parameters,
- $p\,(data \mid paratmeter)$ is the likelihood distribution
- $p\,(parameter \mid data)$ is the posterior distribution. This is the distribution representing the parameter values after we have calculated everything taking the observed data into account.

The simplified Bayes' theorem is

$$\begin{aligned} & p\left(\textit{parameter} \mid \textit{data}\right) \\ & \propto p\left(\textit{parameter}\right) \times p\left(\textit{data} \mid \textit{paratmeter}\right) \end{aligned} \tag{7}$$

- the posterior distribution can be obtained from the prior distribution and the likelihood function.
- the prior distribution and the likelihood function are approximated by Monte Carlo method.

MCMC or Gibbs sampling

The Monte Carlo procedure allows to approximate the distribution $p(x)$ in $[a, b]$ by sampling random variables.
Because

$$\begin{gathered}\int_a^b f(x)\,dx = \int_a^b \frac{f(x)}{p(x)} p(x)\,dx \\ = \int_a^b F(x)\,p(x)\,dx = E\left[F(x)\right]\end{gathered} \tag{8}$$

where $F(x) = \frac{f(x)}{p(x)}$.
We can sample $x_1 \cdots x_n$ to obtain the distribution $p(x)$ as

$$p(x_i) \approx \frac{x_i}{\sum_{i=1}^n x_i} \tag{9}$$

So

$$E\left[F(x)\right] \approx \frac{1}{n}\sum_{i=1}^n \frac{f(x_i)}{p(x_i)}$$