An Introduction to Bayes' Rule

中山大学 数据科学与计算机学院
计算机科学与技术理论基础
18110928 吴侃 & 18110929 罗柯

Contents

An Introduction to Bayes' Rule

Bayes'　Rule: The Basic Form

Given two events M and D, Bayes' Rule states:

$$\Pr(M|D) = \frac{\Pr(D|M)\Pr(M)}{\Pr(D)}$$

Derivation:

$$\Pr(M \cap D) = \Pr(M|D)\Pr(D)$$
$$\Pr(D \cap M) = \Pr(D|M)\Pr(M)$$
$$\Pr(M \cap D) = \Pr(D \cap \mathrm{M})$$

Bayes' Rule: An Example for Checking

Example: Checking Bayes' Rule

Consider two events M and D with the following joint probability table:

	$M=1$	$M=0$
$D=1$	0.25	0.5
$D=0$	0.2	0.05

We can observe that indeed $\mathbf{Pr}(M \mid D) = \mathbf{Pr}(M \cap D)/\mathbf{Pr}(D) = \frac{0.25}{0.75} = \frac{1}{3}$, which is equal to

$$\frac{\mathbf{Pr}(D \mid M)\mathbf{Pr}(M)}{\mathbf{Pr}(D)} = \frac{\frac{.25}{.2+.25}(.2+.25)}{.25+.5} = \frac{.25}{.75} = \frac{1}{3}.$$

Using conditional probability:

$$\Pr(M|D) = \Pr(M \cap D)/\Pr(D)$$

Using Bayes' Rule:

$$\Pr(M|D) = \Pr(D|M)\Pr(M)/\Pr(D)$$

Bayes' Rule: A Further Example

Problem:

- You have bought a new car with its windshield broken, and you would like to guess from which factory it comes.
- You know there are three factories assembling this kind of car, namely A, B, and C, as well as the proportion of cars in the market from them, namely $\Pr(A)$, $\Pr(B)$, and $\Pr(C)$.
- You know the rates of cracked windshields for each factory, namely $\Pr(W|A)$, $\Pr(W|B)$, and $\Pr(W|C)$.

Bayes' Rule: A Further Example (Cont.)

Solution:

- $\Pr(A|W) = \frac{\Pr(W|A)\Pr(A)}{\Pr(W)}$
- $\Pr(B|W) = \frac{\Pr(W|B)\Pr(B)}{\Pr(W)}$
- $\Pr(C|W) = \frac{\Pr(W|C)\Pr(C)}{\Pr(W)}$
- $\Pr(W)$ is unknown but the same to all of A, B, and C, therefore we can compare them and find the largest one.

Model Estimation Given Data

- **Model**

  A pattern which generates data, but observed with noise .

- **Data**

  A set of points generated with noise by a pattern.

Model Estimation Given Data: Examples

- **Single Point Model**

  Model: a single point M in R^d

  Data: a set of points in R^d near the point M

- **Linear Regression**

  Model: a line M in R^d

  Data: a set of points in R^d near the line M

- **Cluster**

  Model: a small set of points M in R^d

  Data: a large set of points in R^d, where each point is near one of the points in M

Model Estimation Given Data: Examples

- **PCA**

  Model: a k-dimensional subspace M in R^d $(k \ll d)$

  Data: a set of points in R^d where each point is near M

- **Linear Classification**

  Model: a half-space M in R^d

  Data: a set of labeled points (with label $+$ or $-$)

- **etc.**

MLE vs. MAP

Given data, what is the corresponding model?

$$\Pr(M|D)\,?$$

Frequentist:

- Maximum Likelihood Estimation (MLE)
- Find the model that is most likely to generate D

Bayesian:

- Maximum A Posterior estimation (MAP)
- Find the model with maximum a posterior

Difference – the prior:

- $\Pr(M|D) \propto \Pr(D|M)\,\Pr(M)$

MLE vs. MAP: An Example

Data:

$\{1, 3, 12, 5, 9\} \in R$

Model:

a point $M \in R$

Noise:

an independent Gaussian noise with $\mu = 0, \sigma = 2$

The PDF of data:

$$g(x) = \frac{1}{\sqrt{8\pi}} \exp(-\frac{1}{8}(M - x)^2)$$

MLE vs. MAP: An Example (Cont.)

$$g(x) = \frac{1}{\sqrt{8\pi}} \exp(-\frac{1}{8}(M - x)^2)$$

$$\Pr(D|M) = \prod_{x \in D} g(x)$$

$$M^* = \arg\max_M \Pr(D|M) = \arg\max_M \ln(\Pr(D|M))$$

$$\ln(\Pr(D|M)) = \sum_{x \in D} \left(-\frac{1}{8}(M - x)^2\right) + c$$

The likelihood reaches maximum when M is the mean of the data.

Bayesian Inference

When it comes to continuous random variables, the probability of any specific point is $\mathbf{0}$.

Instead, the probability density is used:

$$p(M|D) \propto f(D|M)\pi(M)$$

Bayesian Inference (Cont.)

- Though we can't calculate the probabilities, we can compare them:

$$\frac{p(M_1|D)}{p(M_2|D)} = \frac{f(D|M_1)\pi(M_1)}{f(D)} \frac{f(D)}{f(D|M_2)\pi(M_2)}$$

- We can select a range of parameter values instead of a single value.
- Marginalization: we can take a weighted average of all models.

Weight for Prior

How important is the prior?

Weight for Prior: Example

Data:

$$D = \{x_1, \dots, x_n\}, sampled\ from\ N(\mu_M, 2)$$

Prior:

$$\pi(M) = N(66, 6)$$

MAP and MLE:

$$p(M|D) = C_1 f(D|M)\pi(M)$$

$$\ln\big(p(M|D)\big) = \sum_{x \in D} \ln\big(f(x|M)\big) + \ln\big(\pi(M)\big) + C_2$$

$$\ln\big(p(M|D)\big) \propto -\sum_{x \in D} 9(\mu_M - x)^2 - (\mu_M - 66)^2 + C_3$$

Weight for Prior

How important is the prior?

- With any prior, if we get enough data, it no longer becomes important. But with a small amount of data, it can have a large influence on our model.
- MLE goes closer to MAP with more data.
- Exploit prior knowledge when only a small number of data appear.

Questions & Discussion