Assignment 1: Naive Bayes (I)

Xiyou Zhou, 13307130189
Computer Science and Technology

March 26, 2017

1 Problem 1

1.1 Question

What's the relationship between Maximum-A-Posteriori (MAP) estimation and Maximum likelihood (ML) estimation in Naive Bayes? In what sense will the MAP estimation be equal to the ML estimation in Naive Bayes?

1.2 Answer

The Maximum-A-Posteriori estimation maximizes the posteriori distribution using likelihood while the Maximum likelihood estimation maximizes the likelihood function, i.e. fitting the data without considering the data bias. Thus the MAP estimation requires prior distribution to derive the actual posteriori distribution.

When **the prior function is a constant**, the MAP estimation would be equal to the ML estimation in Naive Bayes.

2 Problem 2

2.1 Question

Determine the mean, mode and variance for the beta distribution. Please write down your solution step-by-step.

2.2 Answer

The beta function is defined as follows:

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx, \alpha > 0, \beta > 0$$

And the probability density function of beta distribution is:

$$p(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$$

2.2.1 Mean

$$\mu = E(x) = \int_0^1 \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} x dx = \frac{\int_0^1 x^{\alpha}(1-x)^{\beta-1}dx}{B(\alpha, \beta)} = \frac{B(\alpha+1, \beta)}{B(\alpha, \beta)}$$
$$= \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} = \frac{\alpha}{\alpha+\beta}$$

Here, $\Gamma(k)$ is Gamma function where $p(k) = \int_0^\infty x^{k-1}e^{-x}dx, k \in (0, \infty)$, thus:

$$\Gamma(k+1) = k\Gamma(k)$$

since

$$\Gamma(k+1) = \int_0^\infty x^k e^{-x}\,dx = \left(-x^k e^{-x}\right)_0^\infty + \int_0^\infty kx^{k-1}e^{-x}\,dx = 0 + k\,\Gamma(k)$$

and

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

2.2.2 Mode

Obviously, when $\alpha = \beta = 1$, the mode is 1. Otherwise, when $\alpha \leq 1$ or when $\beta \leq 1$ there is no mode. When $\alpha > 1$ and $\beta > 1$, we differentiate the pdf and get:

$$p'(x) = \frac{(\alpha-1)x^{\alpha-2}(1-x)^{\beta-1} - x^{\alpha-1}(\beta-1)(1-x)^{\beta-2}}{B(\alpha,\beta)}$$

let $p'(x) = 0$ and we get:

$$\begin{aligned}(\alpha-1)(1-x) - (\beta-1)x &= 0\\ \alpha - 1 + x - \alpha x - \beta x + x &= 0\\ (2-\alpha-\beta)x &= 1-\alpha\\ x &= \frac{\alpha-1}{\alpha+\beta-2}\end{aligned}$$

Therefore, the mode is at the peak of probability density function where $x = \frac{\alpha-1}{\alpha+\beta-2}$.

2.2.3 Variance

$E(x) = \frac{\alpha}{\alpha+\beta}$ and here we want $E(x^2)$.

$$\begin{aligned}E(x^2) = \int_0^1 \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}x^2 dx &= \frac{\int_0^1 x^{\alpha+1}(1-x)^{\beta-1}dx}{B(\alpha,\beta)} = \frac{B(\alpha+2,\beta)}{B(\alpha,\beta)}\\ &= \frac{\Gamma(\alpha+2)\Gamma(\beta)}{\Gamma(\alpha+\beta+2)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} = \frac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)}\end{aligned}$$

Thus,

$$\begin{aligned}Var(x) = E(x^2) - E(x)^2 &= \frac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)} - (\frac{\alpha}{\alpha+\beta})^2\\ &= \frac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)} - \frac{\alpha^2}{(\alpha+\beta)^2} = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}\end{aligned}$$

3 Problem 3

3.1 Question

Show that the Maximum likelihood estimation for the Bernoulli / binomial model is

$$\hat{\theta}_{MLE} = \frac{N_1}{N}$$

by optimizing the log of the likelihood $p(D|\theta) = \theta^{N_1}(1-\theta)^{N_0}$. Please write down your solution step-by-step.

3.2 Answer

Here the likelihood function is given as $P(D|\theta) = \theta^{N_1}(1-\theta)^{N_0}$.
Thus, use log on this function and we get:

$$\log P(D|\theta) = \log\left(\theta^{N_1}(1-\theta)^{N_0}\right) = N_1 \log\theta + N_0 \log(1 - theta)$$

And then, we use the differentiation of the log likelihood:

$$\frac{d \log P(D|\theta)}{d\theta} = \frac{N_1}{\theta} - \frac{N_0}{1-\theta}$$

Let the differentiation equal to zero and we get:

$$\hat{\theta}_{MLE} = \frac{N_1}{N}$$

4 Problem 4

4.1 Question (PRML exercise 1.22)

Given a loss matrix with elements L_{kj} , the expected risk is minimized if, for each x, we choose the class that minimizes $\sum_k L_{kj} p(\mathcal{C}_k|x)$. Verify that, when the loss matrix is given by $L_{kj} = 1 - I_{kj}$, where I_{kj} are the elements of the identity matrix, this reduces to the criterion of choosing the class having the largest posterior probability. What is the interpretation of this form of loss matrix?

4.2 Answer

Since the sum of all posterior probability is 1 and we are using $L_{kj} = 1 - I_{kj}$ as the corresponding loss function, the loss matrix would give 1 as a punishment for misclassification and 0 for correct classification to maximize $p(\mathcal{C}_j|x)$. Thus, misclassification rate is minimized using this loss matrix.

5 Problem 5

5.1 Question (PRML exercise 2.3)

Prove the probability function of binomial distribution

$$\text{Bin}(m|N,\mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$$

adds up to 1.

5.2 Answer

Here, since $(1+x)^{N+1} = \sum_{m=0}^{N+1} \binom{N+1}{m} x^m$ stands trivially for $N = 0$, we can prove accordingly.

$$\begin{aligned}
(1+x)^{N+1} &= (1+x)(1+x)^N = (1+x)^N + x(1+x)^N \\
&= \sum_{m=0}^{N} \binom{N}{m} x^m + x \sum_{m=0}^{N} \binom{N}{m} x^m \\
&= \sum_{m=0}^{N} \binom{N}{m} x^m + \sum_{m=1}^{N+1} \binom{N}{m-1} x^m \\
&= \binom{N}{0} x^0 + \sum_{m=1}^{N} \left\{ \binom{N}{m} + \binom{N}{m-1} \right\} x^m + \binom{N}{N} x^{m+1} \\
&= \binom{N+1}{0} x^0 + \sum_{m=1}^{N} \binom{N+1}{m} x^m + \binom{N+1}{N+1} x^{m+1} \\
&= \sum_{m=0}^{N+1} \binom{N+1}{m} x^m
\end{aligned}$$

Therefore,

$$\begin{aligned}
\sum_{m=0}^{N} \mathrm{Bin}(m|N,\mu) &= \sum_{m=0}^{N} \binom{N}{m} \mu^m (1-\mu)^{N-m} \\
&= (1-\mu)^N \sum_{m=0}^{N} \binom{N}{m} (\frac{\mu}{1-\mu})^m \\
&= (1-\mu)^N (1 + \frac{\mu}{1-\mu})^N = 1
\end{aligned}$$

Thus, the proof is complete.

References

[1] Wikipedia: Maximum a posteriori estimation,
`https://en.wikipedia.org/wiki/Maximum_a_posteriori_estimation`

[2] Quora: What is the difference between Maximum Likelihood (ML) and Maximum a Posteri (MAP) estimation?
`https://www.quora.com/What-is-the-difference-between-Maximum-Likelihood-ML-and-Maximum-a-Posteri-MAP-estimation`

[3] StackExchange: Mean And Variance Of Beta Distributions,
`http://math.stackexchange.com/questions/497577/mean-and-variance-of-beta-distributions`

[4] Uah: The Gamma Distribution,
`http://www.math.uah.edu/stat/special/Gamma.html`

[5] Wikipedia: Binomial distribution,
`https://en.wikipedia.org/wiki/Binomial_distribution`