Homework2 Report

Qi Liu

April 1, 2016

1 MLE And MAP

1.1 Maximum Likelihood Estimation

The MLE method will choose a parameter which maximizes the probability of observation. For this problem, the observation data $X = x_1, x_2, \dots, x_n$ are drawn form $\mathcal{N}(\mu; \sigma^2)$ with known variance σ^2 and unknown mean μ. Thus the likelihood function

$$P(X|\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x_i - \mu)^2}{2\sigma^2}) = (\frac{1}{\sqrt{2\pi\sigma^2}})^n \exp(-\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}).$$

To maximize the likelihood function is the same as to maximize its log-likelihood function, which is

$$\mathcal{L}(X|\mu) = \log(P(X|\mu)) = n\log(\frac{1}{\sqrt{2\pi\sigma^2}}) - \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}.$$

Calculate the partial derivative of $\mathcal{L}(X|\mu)$ respect to μ,

$$\frac{\partial}{\partial\mu}\mathcal{L}(X|\mu) = \frac{\sum_{i=1}^{n}(x_i - \mu)}{\sigma^2}.$$

The best parameter will make the partial derivative to zero, which means

$$\begin{aligned}
\frac{\partial}{\partial\mu}\mathcal{L}(X|\mu) &= 0 \\
\frac{\sum_{i=1}^{n}(x_i - \mu)}{\sigma^2} &= 0 \\
\sum_{i=1}^{n}(x_i - \mu) &= 0 \\
\sum_{i=1}^{n} x_i &= n\mu \\
\mu &= \frac{\sum_{i=1}^{n} x_i}{n},
\end{aligned}$$

which means μ is just the mean value of the samples.

1.2 Maximum A Posteriori Estimation

The MLE method maximizes the function $P(X|\mu)$ but the MAP method maximize the function $P(X|\mu)P(\mu)$ and for this problem

$$P(\mu) = \frac{1}{\sqrt{2\pi\beta^2}} \exp(-\frac{(\mu-\nu)^2}{2\beta^2}).$$

To maximize $P(X|\mu)P(\mu)$, we can maximize its log value $\log(P(X|\mu)P(\mu)) = \mathcal{L}(X|\mu) + \mathcal{L}(\mu)$. Here

$$\mathcal{L}(\mu) = \log(P(\mu)) = \log(\frac{1}{\sqrt{2\pi\beta^2}}) - \frac{(\mu-\nu)^2}{2\beta^2}$$

and the partial derivative of $\mathcal{L}(\mu)$ respect to μ is

$$\frac{\partial}{\partial\mu}\mathcal{L}(\mu) = -\frac{\mu-\nu}{\beta^2}.$$

The answer satisfies

$$\begin{aligned}
\frac{\partial}{\partial\mu}(\mathcal{L}(X|\mu) + \mathcal{L}(\mu)) &= 0 \\
\frac{\partial}{\partial\mu}\mathcal{L}(X|\mu) + \frac{\partial}{\partial\mu}\mathcal{L}(\mu) &= 0 \\
\frac{\sum_{i=1}^{n}(x_i-\mu)}{\sigma^2} - \frac{\mu-\nu}{\beta^2} &= 0 \\
\frac{\sum_{i=1}^{n}(x_i-\mu)}{\sigma^2} &= \frac{\mu-\nu}{\beta^2} \\
\beta^2\sum_{i=1}^{n}(x_i-\mu) &= \sigma^2(\mu-\nu) \\
\sigma^2\mu + n\beta^2\mu &= \beta^2\sum_{i=1}^{n}x_i + \sigma^2\nu \\
\mu &= \frac{\beta^2\sum_{i=1}^{n}x_i + \sigma^2\nu}{\sigma^2+n\beta^2}.
\end{aligned}$$

1.3 Two Methods Comparison

We can see

$$\mu_{\mathrm{MLE}} = \frac{\sum_{i=1}^{n}x_i}{n}$$

and

$$\mu_{\mathrm{MAP}} = \frac{\beta^2\sum_{i=1}^{n}x_i + \sigma^2\nu}{\sigma^2+n\beta^2}.$$

Due to σ^2 and $\sigma^2\nu$ are all constants,

$$\mu_{\text{MAP}} \to \frac{\beta^2 \sum_{i=1}^n x_i}{n\beta^2} = \frac{\sum_{i=1}^n x_i}{n} = \mu_{\text{MLE}}$$

when $n \to \infty$, which means MLE and MAP get the same results when n goes to infinity.

2 Naive Bayes

2.1 No Smoothing

Let

$$f(a_1, a_2, C) = P(A_1 = a_1|C)P(A_2 = a_2|C)P(C)$$

be the confidence function. We can get

$$f(2, 2, X) = P(A_1 = 2|X)P(A_2 = 2|X)P(X) = \frac{1}{4} \times \frac{1}{4} \times \frac{2}{3} = \frac{1}{24}$$

and

$$f(2, 2, Y) = P(A_1 = 2|Y)P(A_2 = 2|Y)P(Y) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{3} = \frac{1}{12}.$$

Thus the predict class is Y.

2.2 Laplace Smoothing

With no smoothing,

$$P(A_i = a_i|C) = \frac{\#(A_i = a_i, C)}{\#C}$$

and with Laplace smoothing,

$$P(A_i = a_i|C) = \frac{\#(A_i = a_i, C) + \alpha}{\#C + \alpha k_i},$$

here α is the smoothing parameter and k_i is the number of possible values of attribute A_i. Recalculate the confidence functions,

$$f(2, 2, X) = P(A_1 = 2|X)P(A_2 = 2|X)P(X) = \frac{2}{7} \times \frac{2}{7} \times \frac{2}{3} = \frac{8}{147}$$

and

$$f(2, 2, Y) = P(A_1 = 2|Y)P(A_2 = 2|Y)P(Y) = \frac{2}{5} \times \frac{2}{5} \times \frac{1}{3} = \frac{4}{75}.$$

The predict class has changed to X.

2.3 Spam Email Filter

Due to all the attributes of the data are continuous, we use a normal distribution to estimate the likelihood. We use a class independent variance in our implementation. The final accuracy is 87.1%. And after we removed the attribute **word_freq_make**, the accuracy increases to 87.5%.