RICE UNIVERSITY, DEPARTMENT OF COMPUTER SCIENCE

Assignment 3, COMP 540

Chen Zeng(cz39), Zhihui Xie(zx18)

February 6, 2018

1 MAP and MLE parameter estimation

Answer1 According to the properties of Bernoulli Distribution, we can get:

$$P(D|\theta) = \theta^{\sum x^{(i)}} (1-\theta)^{m-\sum x^{(i)}}$$

According to the principle of maximum likelihood, we can get:

$$\theta^*_{MLE} = argmax_\theta logP(D|\theta) = argmax_\theta \left(\sum x^{(i)} log\theta + \left(m - \sum x^{(i)} \right) log(1-\theta) \right)$$

Make the derivation of $logP(D|\theta)$ on θ and make it equal to 0, we can get the result:

$$\theta^*_{MLE} = \frac{\sum x^{(i)}}{m}$$

This is the estimate of θ.

Answer2 Because of:

$$Beta(\theta|a,b) \propto \theta^{a-1}(1-\theta)^{b-1}$$

Therefore:

$$P(\theta|D) \propto P(D|\theta)P(\theta) \propto \left(\theta^{\sum x^{(i)}} (1-\theta)^{m-\sum x^{(i)}} \right) \left(\theta^{a-1}(1-\theta)^{b-1} \right) = \theta^{\sum x^{(i)}+a-1} (1-\theta)^{m+b-\sum x^{(i)}-1}$$

According to the MAP estimation of θ, we can get:

$$\theta_{MAP} = argmax_\theta logP(\theta|D) = argmax_\theta \left(\left(\sum x^{(i)} + a - 1 \right) log\theta + \left(m + b - \sum x^{(i)} - 1 \right) log(1-\theta) \right)$$

Make the derivation of $logP(\theta|D)$ on θ and make it equal to 0, we can get the result:

$$\theta^*_{MAP} = \frac{a + \sum x^{(i)} - 1}{a + b + m - 2}$$

When it's under a uniform prior, $a = b = 1$, we can get:

$$\theta^*_{MAP} = \frac{\sum x^{(i)}}{m}$$

Combing the answer from problem1, we can get:

$$\theta^*_{MLE} = \theta^*_{MAP}$$

Therefore, the MAP and MLE estimates of θ are the same under a uniform prior.

2 Logistic regression and Gaussian Naive Bayes

Answer1 We can easily get the results from the logistic regression shown as below:

$$P\left(y = 1|x\right) = g\left(\theta^T x\right)$$

$$P\left(y = 0|x\right) = 1 - g\left(\theta^T x\right)$$

They are the expression in terms of the parameter θ and the sigmoid function.

Answer2 At first, we define the const A as below:

$$A = \frac{1}{P\left(x, y' = 1\right) + P\left(x, y' = 0\right)} = \frac{1}{P\left(x|y' = 1\right)\gamma + P\left(x|y' = 0\right)\left(1 - \gamma\right)}$$

Then, according to the Bayes rule, and the definition and properties of Gaussian distribution and Gaussian Naive Bayes model, we can get the results:

$$P\left(y = 1|x\right) = \frac{P\left(x|y = 1\right)P\left(y = 1\right)}{\sum_{y' \in \{0,1\}} P\left(x|y'\right)P\left(y'\right)} = A\gamma \prod_{j=0}^{d} \frac{1}{\sqrt{2\pi\sigma_j^2}} exp\left(-\frac{\left(x^{(j)} - \mu_j^1\right)^2}{2\sigma_j^2}\right)$$

$$P\left(y = 0|x\right) = \frac{P\left(x|y = 0\right)P\left(y = 0\right)}{\sum_{y' \in \{0,1\}} P\left(x|y'\right)P\left(y'\right)} = A\gamma \prod_{j=0}^{d} \frac{1}{\sqrt{2\pi\sigma_j^2}} exp\left(-\frac{\left(x^{(j)} - \mu_j^0\right)^2}{2\sigma_j^2}\right)$$

Here const A is defined at the beginning.

Answer3 Because class 1 and class 0 are equally likely, $\gamma = \frac{1}{2}$. Therefore, to simplify the expression for $P\left(y=1|x\right)$, we can get:

$$P\left(y=1|x\right) = \frac{A}{2}\prod_{j=0}^{d}\frac{1}{\sqrt{2\pi\sigma_j^2}}exp\left(-\frac{\left(x^{(j)}-\mu_j^1\right)^2}{2\sigma_j^2}\right)$$

For the definition of const A:

$$A = \frac{1}{P\left(x,y'=1\right)+P\left(x,y'=0\right)} = \frac{2}{P\left(x|y'=1\right)+P\left(x|y'=0\right)}$$

Let's divide $P\left(y=1|x\right)$ with $P\left(y=0|x\right)$, using the formation in problem2, we can get:

$$\begin{aligned}\frac{P\left(y=1|x\right)}{P\left(y=0|x\right)} &= \frac{A\gamma\prod_{j=0}^{d}\frac{1}{\sqrt{2\pi\sigma_j^2}}exp\left(-\frac{\left(x^{(j)}-\mu_j^1\right)^2}{2\sigma_j^2}\right)}{A\gamma\prod_{j=0}^{d}\frac{1}{\sqrt{2\pi\sigma_j^2}}exp\left(-\frac{\left(x^{(j)}-\mu_j^0\right)^2}{2\sigma_j^2}\right)} \\ &= exp\left(\sum_{j=0}^{d}\left(-\frac{\left(x^{(j)}-\mu_j^1\right)^2}{2\sigma_j^2}+\frac{\left(x^{(j)}-\mu_j^0\right)^2}{2\sigma_j^2}\right)\right) \\ &= exp\left(\sum_{j=0}^{d}\frac{\left(\mu_j^0\right)^2-\left(\mu_j^1\right)^2}{2\sigma_j^2}+\sum_{j=0}^{d}\frac{x^{(j)}\left(\mu_j^1-\mu_j^0\right)}{\sigma_j^2}\right)\end{aligned}$$

According to the answer of problem1, we have:

$$\frac{P\left(y=1|x\right)}{P\left(y=0|x\right)} = \frac{g\left(\theta^T x\right)}{1-g\left(\theta^T x\right)} = exp\left(\theta^T x\right)$$

These two are equal and comparing them, we can get the appropriate parameterization:

$$\sum_{j=0}^{d}\frac{\left(\mu_j^0\right)^2-\left(\mu_j^1\right)^2}{2\sigma_j^2} = 0$$

$$\frac{\mu_j^1-\mu_j^0}{\sigma_j^2} = \theta_j\left(j=0,1,...,d\right)$$

Therefore, it shows that with appropriate parameterization, $P\left(y=1|x\right)$ for Gaussian Naive Bayes with uniform priors is equivalent to $P\left(y=1|x\right)$ for logistic regression.

3 Reject option in classifiers

Answer1 To calculate the minimum risk, we discuss it in the following two cases.
When $i = 1, 2, ..., C$, we have:

$$\begin{aligned} R(\alpha_i|x) &= \sum_{j=1}^{C} L\left(\alpha_i|y=j\right) P\left(y_j|x\right) \\ &= \lambda_s \sum_{j=1, j\neq i}^{C} P\left(y=j|x\right) \\ &= \lambda_s \left(1 - P\left(y=i|x\right)\right) \end{aligned}$$

When $i = C+1$, we have:

$$R(\alpha_{C+1}|x) = \lambda_r$$

Therefore, if we want to decide $y = j$ to obtain the minimum risk, the following two requirements should be fulfilled:

$$R\left(\alpha_j|x\right) \leq R(\alpha_k|x), \left(j, k = 1, 2, ..., C \& j \neq k\right)$$

$$R\left(\alpha_j|x\right) \leq R(\alpha_{C+1}|x), \left(j = 1, 2, ..., C\right)$$

Combining with the results we get before, we need to solve the following two inequalities:

$$\lambda_s \left(1 - P\left(y=j|x\right)\right) \leq \lambda_s \left(1 - P\left(y=k|x\right)\right)$$

$$\lambda_s \left(1 - P\left(y=j|x\right)\right) \leq \lambda_r$$

Finally we get the results:

$$P\left(y=j|x\right) \geq P\left(y=k|x\right)$$

$$P\left(y=j|x\right) \geq 1 - \frac{\lambda_r}{\lambda_s}$$

Therefore, if it fulfilled the above requirements, we decide $y = j$ to obtain the minimum risk.

Answer2 When $\lambda_r = 0$, which makes $\lambda_r/\lambda_s = 0$, we do always reject. As it increases but not equal to 1, the relative cost of rejection increases, we are more likely not to reject. But at this moment, we still reject. When $\lambda_r/\lambda_s = 1$, we can either reject or not to reject and the results will be the same.

4 Kernelizing k-nearest neighbors

Answer To re-express this classification rule in terms of dot products, we have:

$$D = \sqrt{\sum ((X - Y) \cdot (X - Y))}$$

To make use of the kernel trick to formulate the k-nearest neighbor algorithm, we have:

$$D = K(X, Y)$$

5 Constructing kernels

Answer1 According to the Mercer's theorem, if k_1 is a valid kernel, then the Gram matrix K_1 whose elements are $k_1\left(x, x^{'}\right), 1 \leq i, j \leq m$ is positive definite. According to the definition of positive definite, the scalar $z^T K_1 z$ should always be positive for every non zero column vector z. Because we have $k\left(x, x^{'}\right), 1 \leq i, j \leq m = ck_1\left(x, x^{'}\right), 1 \leq i, j \leq m$, we can know the Gram matrix for k is $K = cK_1$. So we have $z^T Kz = cz^T K_1 z$. It should also be positive and K should be positive definite if $c > 0$. Therefore, according to the Mercer's theorem, under the case of $c > 0$, k should also be valid kernels.

Answer2 Here $k\left(x, x^{'}\right)$ may not be valid kernel and it depends on $f(x)$. For example, if $f(x) = 0$, then K will be a zero matrix and it's not positive definite. Therefore, according to the Mercer's theorem, under this case, k is not a valid kernel. Also, if $f(x) = 1$ and then $k = k_1$. Under this case, k is a valid kernel.

Answer3 This is similar to answer1. Here $K = K_1 + K_2$ and $z^T Kz = z^T K_1 z + z^T K_2 z > 0$ for every non zero column vector z. According to the Mercer's theorem, K is positive definite and therefore k should also be valid kernels.

6 One-vs-all logistic regression

Answer

7 Softmax regression

Answer