Exercise 05 of Machine Learning [IN 2064]

Name: Yiman Li
Matr-Nr: 03724352
cooperate with Kejia Chen(03729686)

Problem 1

a) We can try to calculate the distribution of the posterior distribution as below:

$$p(y=1|x) = \frac{p(x|y=1)p(y=1)}{p(x|y=1)\cdot p(y=1) + p(x|y=0)\cdot p(y=0)} = \frac{\lambda_1 e^{-\lambda_1 x}}{\lambda_1 e^{-\lambda_1 x} + \lambda_0 e^{\lambda_0}} = \theta \quad (1)$$

$$p(y=0|x) = \frac{p(x|y=0)p(y=0)}{p(x|y=1)\cdot p(y=1) + p(x|y=0)\cdot p(y=0)} = \frac{\lambda_0 e^{-\lambda_0 x}}{\lambda_1 e^{-\lambda_1 x} + \lambda_0 e^{\lambda_0}} = 1-\theta \quad (2)$$

So we can write that

$$p(y|x) = \theta^{\prod(y=1)}(1-\theta)^{\prod(y=0)} \quad (3)$$

Which is a Bernoulli distribution.
b) Now that x are classified as class 1, which means

$$p(y=1|x) > p(y=0|x) \quad (4)$$

So we have

$$\lambda_1 e^{-\lambda_1 x} > \lambda_0 e^{-\lambda_0 x} \quad (5)$$

$$\frac{\lambda_1}{\lambda_0} > e^{(\lambda_1-\lambda_0)x} \quad (6)$$

$$\ln\frac{\lambda_1}{\lambda_0} > (\lambda_1-\lambda_0)x \quad (7)$$

- When $\lambda_1 > \lambda_0$, $x < \frac{\ln\lambda_1 - \ln\lambda_0}{\lambda_1-\lambda_0}$
- When $\lambda_1 < \lambda_0$, $x > \frac{\ln\lambda_1 - \ln\lambda_0}{\lambda_1-\lambda_0}$

Problem 2

According to the Book of Bishop [1], the maximum likelihood solution lack robustness to outliers. Namely, maximum likelihood estimation may often lead to overfitting for data sets that are linearly separable. This arises because the maximum likelihood solution occurs when the hyperplane corresponding to $\sigma = 0.5$, equivalent to $\boldsymbol{w}^T\boldsymbol{\phi} = 0$, separates the two classes and the magnitude of $\boldsymbol{w}$ goes to infinity.
Considering the picture shown in Figure 1: there are two classes of data, denoted bu red crosses and blue circles, together with the decision boundary found by least squares(magenta curve) and also by logistic regression model(green curve). We can see that the additional data points in the right figure produce a significant change in the location of the decision boundary, even though these points would be correctly classified by the original decision boundary in the left figure.
We can control this by penalizing large weights, just as what we do in linear regression:

$$E(\boldsymbol{w}) = -\ln p(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{X}) + \lambda||\boldsymbol{w}||_q^2 \quad (8)$$

Like before, for $q = 2$ this corresponds to MAP estimation with a Gaussian prior on $\boldsymbol{w}$.

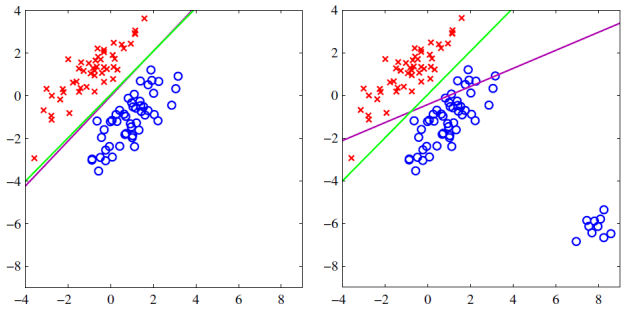

Figure 1: Contrast between least squares and logistic regression model

Problem 3

For the case of $K > 2$ classes, we have

$$\begin{aligned} p(\mathcal{C}_k|x) &= \frac{p(\boldsymbol{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{\sum_j p(\boldsymbol{x}|\mathcal{C}_j)p(\mathcal{C}_j)} \\ &= \frac{\exp(a_k)}{\sum_j \exp(a_j)} \end{aligned} \tag{9}$$

which is known as the normalized exponential, also softmax function. Here the quantities a_k are defined by

$$a_k = \ln p(\boldsymbol{x}|\mathcal{C}_k)p(\mathcal{C}_k) \tag{10}$$

When we set $K = 2$, then we have, for instance:

$$\begin{aligned} p(\mathcal{C}_1|\boldsymbol{x}) &= \frac{\exp(a_1)}{\exp(a_1) + \exp(a_2)} \\ &= \frac{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \\ &= \frac{1}{1 + \exp(-\ln \frac{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2)})} \\ &= \frac{1}{1 + \exp(-a)} \\ &= \sigma(a) \end{aligned} \tag{11}$$

Which is a logistic sigmoid function, where we have defined

$$a = \ln \frac{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \tag{12}$$

So we can see that the sofymax function is equivalent to a sigmoid function in the 2-class case.

Problem 4

We can define $\phi(x_1, x_2) = x_1 \cdot x_2$, then we have

$$\phi(x_1, x_2) = x_1 \cdot x_2 \begin{cases} < 0, \text{ if } x_1, x_2 \in \text{blue cross;} \\ > 0, \text{ if } x_1, x_2 \in \text{black circle.} \end{cases} \tag{13}$$

References

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, 2007.