Tutorial 6

Machine Learning and Big Data for Economics and Finance

List of activities

I. Complete **Section 5.3 Lab: Cross-validation and the Boostrap**, subsection 5.3.4.

II. Complete the list of exercises in this tutorial.

Exercise 1. Bootstrap for the logistic regression model
Consider the logistic regression model

$$\Pr\{Y=1|X=x\} = \frac{1}{1+\mathrm{e}^{-\beta_0-\beta_1 x-\beta_2 x^2}}$$

Write your own code to maximize the likelihood function with respect to $(\beta_0, \beta_1, \beta_2)$ and to compute the standard errors of the parameter estimators by the bootstrap method. Test your code on the dataset in `LR2.csv`.

Exercise 2. Linear discriminant analysis
Derive formula 4.13 in the textbook.

Solution.
Method 1: (much shorter than method 2)
Follow the argument in the slides while using $p_k(x) \propto \pi_k \mathrm{e}^{-\frac{1}{2\sigma^2}(x-\mu_k)^2}$.

Method 2:
Given that the posterior probability is

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2\sigma^2}(x-\mu_k)^2}}{\sum_{k'} \pi_{k'} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2\sigma^2}(x-\mu_{k'})^2}}$$

we deduce that $p_k > p_j$ for some $j \neq k$ if and only if

$$\frac{\pi_k \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2\sigma^2}(x-\mu_k)^2}}{\sum_{k'} \pi_{k'} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2\sigma^2}(x-\mu_{k'})^2}} > \frac{\pi_j \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2\sigma^2}(x-\mu_j)^2}}{\sum_{k'} \pi_{k'} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2\sigma^2}(x-\mu_{k'})^2}}$$

Simplifying the denominators
$p_k > p_j$ if and only if

$$\pi_k \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2\sigma^2}(x-\mu_k)^2} \quad > \quad \pi_j \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2\sigma^2}(x-\mu_j)^2}$$

Again $p_k > p_j$ only if $\log(p_k) > \log(p_j)$ which implies that $p_k > p_j$ if and only if

$$\log(\pi_k) + \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}(x-\mu_k)^2 > \log(\pi_j) + \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}(x-\mu_j)^2$$

Simplifying $\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)$ away yields $p_k > p_j$ if and only if

$$\log(\pi_k) - \frac{1}{2\sigma^2}(x-\mu_k)^2 > \log(\pi_j) - \frac{1}{2\sigma^2}(x-\mu_j)^2$$

which is the same as

$$\log(\pi_k) - \frac{1}{2\sigma^2}x^2 + x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} > \log(\pi_j) - \frac{1}{2\sigma^2}x^2 + x\frac{\mu_j}{\sigma^2} - \frac{\mu_j^2}{2\sigma^2}$$

Since $-\frac{1}{2\sigma^2}x^2$ does not depend on either k or j, then simplyfing it out yields $p_k > p_j$ if and only if

$$\log(\pi_k) + x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} > \log(\pi_j) + x\frac{\mu_j}{\sigma^2} - \frac{\mu_j^2}{2\sigma^2}$$

which is the same as sying that $p_k(x) > p_j(x)$ if and only if $\delta_k(x) > \delta_j(x)$ where $\delta_k(x)$ is defined by equation 4.13.