Practical Session 04

Linear Classification

1 Naive Bayes

Problem 1: In LDA we assume that all classes share the same covariance matrix $\boldsymbol{\Sigma}$. The Naive Bayes classifier assumes that all d features of a sample $\boldsymbol{x} = (x_1, x_2, \dots, x_d)$ are conditionally independent given the class, i.e.

$$p(x_1, x_2, \dots, x_d | y) = \prod_{i=1}^{d} p(x_i | y)$$

In the case of continuous data where the likelihood is a normal distribution, this corresponds to **diagonal** covariance matrices that however are different for each class (not shared). The generative process is thus:

$$p(\boldsymbol{x} | y = c) = \mathcal{N}(\boldsymbol{x} | \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

Show that using different $\boldsymbol{\Sigma}_c$'s for each class leads to quadratic decision boundaries.

2 Multi-Class Classification

Problem 2: Consider a generative classification model for C classes defined by prior class probabilities $p(y = c) = \pi_c$ and general class-conditional densities $p(\boldsymbol{x} | y = c, \boldsymbol{\theta}_c)$ where $\boldsymbol{x}$ is the input feature vector and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_c\}_{c=1}^{C}$ are further model parameters. Suppose we are given a training set $\mathcal{D} = \{(\boldsymbol{x}^{(n)}, y^{(n)})\}_{n=1}^{N}$ where $y^{(n)}$ is a binary target vector of length C that uses the 1-of-C(one-hot) encoding scheme, so that it has components $y_c^{(n)} = \delta_{ck}$ if pattern n is from class $y = k$. Assuming that the data points are iid, show that the maximum-likelihood solution for the prior probabilities is given by

$$\pi_c = \frac{N_c}{N}$$

where N_c is the number of data points assigned to class $y = c$.

Problem 3: Using the same classification model as in the previous question, now suppose that the class-conditional densities are given by Gaussian distributions with a shared covariance matrix, so that

$$p(x | y = c, \boldsymbol{\theta}_c) = p(x | \boldsymbol{\theta}_c) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_c, \boldsymbol{\Sigma}).$$

Show that the maximum likelihood solution for the mean of the Gaussian distribution for class C_c is given by

$$\boldsymbol{\mu}_c = \frac{1}{N_c} \sum_{\{n | \boldsymbol{x}^{(n)} \in C_c\}} \boldsymbol{x}^{(n)}$$

which represents the mean of those feature vectors assigned to class C_c.

Similarly, show that the maximum likelihood solution for the shared covariance matrix is given by

$$\mathbf{\Sigma} = \sum_{c=1}^{C} \frac{N_c}{N} \mathbf{S}_c$$

where

$$\mathbf{S}_c = \frac{1}{N_c} \sum_{\{n | \boldsymbol{x}^{(n)} \in C_c\}} (\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)^T.$$

Thus $\mathbf{\Sigma}$ is given by a weighted average of the covariances of the data associated with each class, in which the weighting coefficients N_c/N are the prior probabilities of the classes.

Problem 4: Error measures for classification

- ROC curve and AUC
- PR curve and AUC a.k.a. average precision