TRIAL EXAM IN
MA8701 GENERAL STATISTICAL METHODS

Thursday February 23, 2017

Problem 1

Let (X, G) be a random pair where the input X is a random vector and G is a categorical variable, $G \in \{1, 2, \dots, K\}$, denoting the class from which the observation X comes. The task is to predict the class G from an observation of X alone.

a) Show that if the loss is 0 for a correct classification and 1 for a wrong classification, then the optimal strategy for an observed $X = x$ is to classify to the class k which maximizes

$$Pr(G = k|X = x).$$

In questions b), c) d) below it is assumed that conditional on the class being $G = k$, X is distributed with a density function $f_k(x)$. Suppose further that apriori probabilities of the classes are given by π_k; $k = 1, 2, \dots, K$.

b) Show that under these conditions the optimal classification k for an observed $X = x$ is found by maximizing

$$f_k(x)\pi_k$$

with respect to $k \in \{1, 2, \dots, K\}$.

c) With the above as the point of departure, derive the classification criterion of *Linear Discriminant Analysis* (LDA). What is meant by the *linear discriminant functions*?

d) What is meant by *Quadratic Discriminant Analysis* (QDA)? How would you define appropriate *quadratic discriminant functions*?

e) Describe how *logistic regression* is used in the classification problem defined in the start of the exercise.

f) Discuss the relation between logistic regression and linear discriminant analysis. Which similarities and which differences can you point to? Which are the advantages or disadvantages of one method compared to the other?

Problem 2

a) Define the concepts of cubic splines and natural cubic splines in one dimension.

Argue that all cubic splines can be written in the form

$$f(X) = \sum_{j=0}^{3} \beta_j X^j + \sum_{k=1}^{K} \theta_k (X - \xi_k)_+^3$$

Define the expression $()_+$ and give an interpretation of the ξ_k.

b) Prove that the boundary conditions for natural cubic splines imply the following constraints on the coefficients:

$$\begin{aligned} &\beta_2 = 0, \quad \textstyle\sum_{k=1}^{K} \theta_k = 0, \\ &\beta_3 = 0, \quad \textstyle\sum_{k=1}^{K} \xi_k \theta_k = 0. \end{aligned}$$

c) Show that these constraints are satisfied if the following are taken as possible basis functions for natural splines:

$$N_1(X) = 1, \; N_2(X) = X,$$

$$N_{k+2}(X) = d_k(X) - d_{K-1}(X), \; k = 1, 2, \ldots, K-2,$$

where

$$d_k(X) = \frac{(X - \xi_k)_+^3 - (X - \xi_K)_+^3}{\xi_K - \xi_k}$$

Why does this prove that $N_1, N_2, \ldots, N_K$ indeed is a basis for natural splines?