Norges teknisk–
naturvitenskapelige universitet
Institutt for matematiske fag

TRIAL EXAM IN
MA8701 GENERAL STATISTICAL METHODS

Tuesday March 22, 2011

Problem 1

Let (X, Y) be a random pair where Y is 1-dimensional. Consider the task of predicting Y by a function $f(X)$.

a) Define what is meant by a loss function, and in particular what is meant by squared error loss.

Show that the optimal strategy under squared error loss is to use

$$f(x) = E(Y|X = x)$$

b) If, in the above point, we insist on using a linear function of the form $f(x) = x^T\beta$, what is the optimal β?

c) Suppose now that Y is categorical, $Y \in \{1, 2, \ldots, K\}$, and that the loss is of so-called $0-1$ type.

What is now the loss function?

Show that the optimal strategy is to use

$$f(x) = \text{argmax}_y Pr(Y = y|X = x)$$

d) In practice the joint distribution of (X, Y) is not known, but instead we may have training data $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$ drawn from the distribution of (X, Y).

Suppose that we use the training data to obtain prediction functions $\hat{f}(X)$.

Discuss, assuming a general loss function $L(Y, f(X))$, how we can measure the prediction ability of such functions.

Why is the training error rate not a good measure?

How can we improve on it?

e) Describe how K-fold cross-validation and bootstrapping can be used to decide between models of different complexity.

In what cases are these approaches recommended?

Try to be as concise as possible.

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:

$$\beta_2 = 0, \quad \sum_{k=1}^{K} \theta_k = 0,$$
$$\beta_3 = 0, \quad \sum_{k=1}^{K} \xi_k \theta_k = 0.$$

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?