Trial-exam in STK4030, fall semester 2006.

Introduction.

This is slightly more than will be expected at a 3 hour exam, but the types of assignments will be similar. The set will be discussed on Monday, November 6.

Exercise 1.

a) Define the concepts of cubic splies and natural cubic splines in one dimension. Show 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 points 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 the natural splines:

$$N_X = 1, \quad N_2(X) = X,$$

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

where

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

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

Exercise 2.

Assume that in a $p-$dimensional space you have some collection of points of two types: crosses and rings. Your task is to find a classification boundary between these points.

The nearest neighbor method is defined as follows: Let the coordinates of the points be x_i, and put $y_i = 0$ for a ring, $y_i = 1$ for a cross. Introduce for fixed k:

$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i,$$

where $N_k(x)$ is, in some metric, the set of k points closest to x. Classify x as belonging to a cross-region if $\hat{Y}(x) > 0.5$.

a) Show that, if $k = 1$, every point in the collection above is given a correct classification.

b) Does this mean that the above classification rule always is good? Give reasons for your answer.

c) Describe two different ways of assessing the goodness of such a classification rule.

d) Let N be the number of points in the collection. Assume that the values (x_i, y_i) are independent observations of stochastic variables (X, Y). Give qualitative reasons for the following: When k and N tend to infinity in such a way that $k/N \to 0$, then $\hat{Y}(x)$ approaches $\mathrm{E}(Y|X = x)$.

e) Assume that the conditional distribution of Y given X is Bernoulli. Find what the parameter of this Bernoulli distribution must be in order that the limiting classification rule from d) shall give the same rule as logistic regression.

f) Argue for the following: When the dimension p is large, the approximation described in d) may be poor.