Norges teknisk–
naturvitenskapelige universitet
Institutt for matematiske fag

EXAM IN
MA8701 GENERAL STATISTICAL METHODS

Thursday May 19, 2011
09:00 – 13:00

No aids permitted.

The exam consists of 3 problems which are given equal weight in the grading.

You may in the solution of the exercises need the density of the multinormal distribution with dimension p, expectation vector μ and covariance matrix Σ:

$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}.$$

Problem 1

Let (X, G) be a random pair where the input X is a random vector and G is a categorical variable, $G \in \{1, 2, \ldots, K\}$, 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; $k = 1, 2, \ldots, K$.

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 $k \in \{1, 2, \ldots, K\}$.

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

Let (X, Y) have joint distribution on $R \times R$ and define

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

The task is to estimate the function $f(x)$ from training data $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$ drawn from the distribution of (X, Y).

a) The *k-nearest-neighbor* estimate is defined as

$$\hat{f}(x) = \text{Ave}(y_i | x_i \in N_k(x)) \tag{1}$$

Explain the various ingredients of this definition.

What is the reasoning behind the method? What is the role of k?

b) Give an informal explanation of the following: When k and N tend to infinity in such a way that $k/N \to 0$, then $\hat{f}(x)$ in (1) approaches $E(Y|X = x)$.

c) Assume *in this question only* that the joint distribution of (X, Y) can be described by the relation

$$Y = f(X) + \epsilon,$$

where X and ϵ are independent and $E(\epsilon) = 0$, $Var(\epsilon) = \sigma^2$.

Consider estimation of $f(x_0)$ for some fixed x_0, using $\hat{f}(x_0)$ with $\hat{f}$ given in (1).

Suppose that the x_i of the training data are fixed in advance, and let $x_{(\ell)}$ for $\ell = 1, 2, \ldots, N$ be the x_i ordered according to the distance from x_0, with $x_{(1)}$ being the nearest point etc.

Find an expression for the mean squared error

$$E[\hat{f}(x_0) - f(x_0)]^2$$

and discuss the so-called *Bias-Variance Tradeoff* for determination of k in light of this.

d) One disadvantage of the nearest-neighbor method is that it usually leads to a discontinuous $\hat{f}(x)$.

An alternative is the so-called Nadaraya-Watson estimator

$$\hat{f}(x_0) = \frac{\sum_{i=1}^{N} K_\lambda(x_0, x_i) y_i}{\sum_{i=1}^{N} K_\lambda(x_0, x_i)}.$$

How is the function $K_\lambda(x_0, x_i)$ defined? What is the meaning of λ?

What is the reasoning behind the method?

e) What are possible problems with the Nadaraya-Watson estimator? What is the possible advantage of using a local linear regression method instead?

Describe the main ideas of local polynomial regression.

Why can the Nadaraya-Watson estimator be viewed as a special case of local polynomial regression? (Answer the question by writing down a least squares expression for which the Nadaraya-Watson estimator is the minimizer).

Problem 3

a) Describe briefly the linear regression model and the least squares solution for the coefficient vector. (Assume that the response is 1-dimensional).

Give reasons why ordinary least squares estimates are often not satisfactory in applications of linear models.

b) Give an overview of methods that can be used to overcome the problems with ordinary least squares in linear models, in particular consider: subset selection; shrinkage methods; methods using derived input directions.

Rough descriptions of main ideas, definitions and properties are satisfactory.