m/foreslåtte rettelse 30/5-13

**Norges teknisk–
naturvitenskapelige universitet
Institutt for matematiske fag**

EXAM IN
MA8701 GENERAL STATISTICAL METHODS

Wednesday May 15, 2013
09:00 – 13:00

No aids permitted.

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 STATISTICAL LEARNING

Answer very briefly the following questions:

a) What is meant by *supervised learning*? What are the main objectives of supervised learning?

Formalize the problem of supervised learning using Y and X in

- a *regression* setting
- a *classification* setting

b) Explain by simple examples what is meant by *model selection.*

c) It is often suggested to divide the data into three parts: *Training set, validation set and test set.* What is the reason for this, and it what cases is such a division recommended? How big - in percent of the full data set - is it recommended to make the three parts?

d) If we do not find it appropriate to divide the data like above, which other methods can alternatively be used for the above purposes? (Be short).

Problem 2 ESTIMATION AND ASYMPTOTICS

Let $Z_1, \ldots, Z_n$ be i.i.d. random variables, each with density function $f(y;\theta)$, where θ is an r-dimensional vector of unknown parameters and f is a known function.

Let $\hat{\theta}$ be the solution for θ of the equation

$$\sum_{i=1}^{n} \eta(Z_i, \theta) = 0$$

a) What is such an equation called? What is meant by *Fisher consistency* in this connection?

b) Which function η leads to the the maximum likelihood estimator?

Let

$$\mu_j(\theta) = E_\theta(Z_i^j) \text{ for } j = 1, 2, \ldots, r.$$

The moment estimators for the components of θ are found by equating the $\mu_j(\theta)$ to their empirical versions based on the data $Z_1, \ldots, Z_n$.

c) Write down the function η which leads to the moment estimator $\hat{\theta}$ of θ.

Suppose now that the Z_i are gamma-distributed with density

$$f(y;\lambda,k) = \frac{1}{\Gamma(k)} \lambda (\lambda y)^{k-1} e^{-\lambda y} \text{ for } y > 0,$$

where $\lambda > 0$, $k > 0$ are unknown parameters. You may use without proof that

$$E(Z_i) = \frac{k}{\lambda},\ E(Z_i^2) = \frac{k(k+1)}{\lambda^2}$$

[illegible] $E(Z_i^3)$, $E(Z_i^4)$

d) Write down the function η which leads to the moment estimator for (λ, k).

Let θ_0 be the true value of θ. Recall the general result (you are *not* asked to prove this),

$$n^{1/2}(\hat{\theta} - \theta_0) \to N_r(0, B_{\theta_0}^{-1} A_{\theta_0} B_{\theta_0}^{-1})$$

where

$$\begin{aligned} A_{\theta_0} &= E_{\theta_0}[\eta(Z_1, \theta_0)\eta^T(Z_1, \theta_0)] \\ B_{\theta_0} &= E_{\theta_0}[\eta'(Z_1, \theta_0)] \end{aligned}$$

and $\eta'(z,t)$ is the matrix with (i,j)th entry equal to $\partial \eta_i(z,t)/\partial t_j$.

Show how one can

e) Use this to derive the asymptotic distribution of the moment estimator of (λ, k) in the above gamma case. (a complete solution is not asked for).

Problem 3 SPLINES

a) What is the definition of a cubic spline with K knots at $\xi_1, \ldots, \xi_K$?

Argue that the following is a basis of the cubic spline above:

$$\begin{aligned} h_j(X) &= X^{j-1},\ j = 1, 2, 3, 4 \\ h_{4+\ell}(X) &= (X - \xi_\ell)_+^3,\ \ell = 1, \ldots, K \end{aligned}$$

b) Define what is meant by a natural cubic spline with K knots at $\xi_1, \ldots, \xi_K$.

What is the number of basis functions for this natural spline? How do you derive this number?

c) Write down the minimization problem that leads to a smoothing spline.

What are the characteristics of the smoothing spline?

Problem 4 CLASSIFICATION

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) Under which loss function is it optimal for an observed $X = x$ to classify to the class ~~k~~ g which maximizes

$$Pr(G = \cancel{k}\,g | X = x)?$$

Double use of k

(You need not prove this).

The task is to to derive a classification rule from training data $(x_1, g_1), (x_2, g_2), \ldots, (x_N, g_N)$ drawn from the distribution of (X, G).

b) Describe how *k-nearest-neighbor* methods can be used to classify new inputs x.

Why is the method you suggest reasonable in view of the target of maximizing $Pr(G = \cancel{k}\,g | X = x)$?

Why does this approach break down in high dimensions?

c) Describe different approaches using linear methods for the classification. Describe possible limitations, differences and similarities between the methods.

d) Consider logistic regression with a single quantitative input X and two classes represented by $G = 1$ and $G = 0$, respectively. Let the model be

$$\log \frac{Pr(G = 1|X = x)}{P(G = 0|X = x)} = f(x)$$

so that

$$Pr(G = 1|X = x) = \frac{e^{f(x)}}{1 + e^{f(x)}}$$

Write down the log likelihood function for training data $(x_i, y_i), i = 1, \ldots, N$, and add a suitable penalization term which will lead to a smoothing spline estimate for $f(x)$.