UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in: STK4030 — Modern Data Analysis

Day of examination: Thursday December 13'th 2012

Examination hours: 14.30 – 18.30

This problem set consists of 3 pages.

Appendices: None

Permitted aids: Approved calculator

Please make sure that your copy of the problem set is complete before you attempt to answer anything.

Solution proposal

Problem 1

In addition to the definition of the methods as described in the textbook in sections 3.3-3.6 a discussion of how the methods are used in model selection is appropriate.

Problem 2

a) $trace[\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T] = trace[(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}] = trace(\mathbf{I}_{p+1}) = p + 1$.

b)

$$\hat{y}_i = x_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

By independence $cov(\hat{y}_i, y_i) = x_i^T(\mathbf{X}^T\mathbf{X})^{-1}x_i\sigma_\varepsilon^2 = trace(x_i x_i^T(\mathbf{X}^T\mathbf{X})^{-1})\sigma_\varepsilon^2$, so $\sum_{i=1}^N cov(\hat{y}_i, y_i) = \sum_{i=1}^N trace(x_i x_i^T(\mathbf{X}^T\mathbf{X})^{-1})\sigma_\varepsilon^2 = trace((\mathbf{X}^T\mathbf{X}))(\mathbf{X}^T\mathbf{X})^{-1})\sigma_\varepsilon^2 = (p+1)\sigma_\varepsilon^2$.

c) A linear fitting method is one for which the fitted values can he written $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$ for a $N \times N$ matrix which only depends on the input vectors x_i but not on the responses y_i. The *effective degrees-of-freedom* is defined as $df(\mathbf{S}) = trace(\mathbf{S})$.

d) By independence $cov(\hat{f}(x_i), y_i) = cov(\frac{1}{k}\sum_{x_j \in N_k(x)} y_j, y_i) = \frac{1}{k}$. Thus $\sum_{i=1}^N cov(\frac{1}{k}\sum_{x_j \in N_k(x)} y_j, y_i) = \frac{N}{k}$.

Each row corresponds to an observation with input x_i. Let $x_{i_1}, \ldots, x_{i_k}$ be the inputs which are in $N_k(x_i)$. Let the elements in $\mathbf{S}$ be

$$\mathbf{S}_{ii_j} = \begin{cases} 1/k & j = 1, \ldots, k \\ 0 & else \end{cases}$$

(Continued on page 2.)

Then $\mathbf{S}$ does not depend on the y_i's and $\frac{1}{k}\sum_{x_j \in N_k(x)} y_j = \mathbf{S}\mathbf{y}$.

Problem 3

a) Model:

$$Pr(G=1|X=x) = \frac{\exp(\beta_0+\beta^T x)}{1+\exp(\beta_0+\beta^T x)}, Pr(G=0|X=x) = 1-P(G=1|X=x)$$

Log-likelihood:

$$l(\beta) = \sum_{i=1}^{N}\{y_i \log p(x_i,\beta) + (1-y_i)\log(1-p(x_i,\beta))\}$$

b) Model:

$$\begin{aligned} Pr(G=k|X=x) &= \frac{\exp(\beta_{k0}+\beta_k^T x)}{1+\sum_{l=1}^{K-1}\exp(\beta_{l0}+\beta_l^T x)}, k=1,\ldots,K-1 \\ Pr(G=K|X=x) &= \frac{1}{1+\sum_{l=1}^{K-1}\exp(\beta_{l0}+\beta_l^T x)} \end{aligned}$$

Denote responses as $y_{ij},\ j=1,\ldots,K, i=1,\ldots,N$ where $\sum_{j=1}^{K} y_{ij} = 1,\ i=1,\ldots,N$. Then log likelihood may be expressed as

$$\sum_{i=1}^{N}\{\sum_{j=1}^{K}[\beta_{j0}+\beta_j^T x_i]y_{ij} - \log[1+\sum_{l=1}^{K-1}\exp(\beta_{l0}+\beta_l^T x_i)]\}$$

where $\beta_{K0}=0$ and $\beta_K=0$.

A localized log-likelihood is constructed by weighing the terms in the sum by a kernel, i.e.

$$\sum_{i=1}^{N} K(x_0,x_i)\{\sum_{j=1}^{K}[\beta_{j0}+\beta_j^T x_i]y_{ij} - \log[1+\sum_{l=1}^{K-1}\exp(\beta_{l0}+\beta_l^T x_i)]\}$$

c) Fitting a locally constant logistic model consists of minimizing

$$min_{\alpha(x_0)}\sum_{i=1}^{N} K(x_0,x_i)\{\sum_{j=1}^{K}\alpha_{j0}(x_0)y_{ij}\} - \sum_{i=1}^{N} K(x_0,x_i)\{\log[1+\sum_{l=1}^{K-1}\exp(\alpha_{l0}(x_0))]\}$$

where $\alpha(x_0) = (\alpha_1(x_0),\ldots,\alpha_{K-1}(x_0))^T$ and $\alpha_K(x_0)=0$. But

$$\begin{aligned} \frac{\partial}{\partial\alpha_j(x_0)} &= \sum_{i=1}^{N} K(x_0,x_i)y_{ij} \\ &- \sum_{i=1}^{N} K(x_0,x_i)\{\frac{\exp(\alpha_j(x_0))}{1+\sum_{l=1}^{K-1}\exp(\alpha_0(x_0)}\}, j=1,\ldots,K-1 \end{aligned}$$

Thus $\frac{\partial}{\partial\alpha_j(x_0)} = 0$ has solution

$$\sum_{i=1}^{N} K(x_0,x_i)y_{ij} = \frac{\exp(\hat\alpha_j(x_0))}{1+\sum_{l=1}^{K-1}\exp(\hat\alpha_0(x_0)}\sum_{i=1}^{N} K(x_0,x_i), j=1,\ldots,K-1$$

(Continued on page 3.)

or

$$\hat{f}_j(x_0) = \frac{\exp(\hat{\alpha}_j(x_0))}{1 + \sum_{l=1}^{K-1} \exp(\hat{\alpha}_0(x_0)} = \frac{\sum_{i=1}^N K(x_0, x_i) y_{ij}}{\sum_{i=1}^N K(x_0, x_i)}, j = 1, \dots, K-1$$

Also,

$$\hat{f}_K(x_0) = 1 - \sum_{j=1}^{K-1} \hat{f}_j(x_0) = \frac{\sum_{i=1}^N K(x_0, x_i)[1 - \sum_{j=1}^{K-1} y_{ij}]}{\sum_{i=1}^N K(x_0, x_i)} = \frac{\sum_{i=1}^N K(x_0, x_i) y_{iK}}{\sum_{i=1}^N K(x_0, x_i).}$$

Thus for all the categories the fitted values are the smoothed response indicators separately using the Nadaraya-Watson kernel smoother.

SLUTT