UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Examination in:	STK4030/9030 — Modern data analysis - Solutions
Day of examination:	Friday December 5. 2008.
Examination hours:	14.30 – 17.30.
This examination set consists of 4 pages.	
Appendices:	None.
Permitted aids:	Accepted calculator

Make sure that your copy of the examination set is complete before you start solving the problems.

Problem 1.

(a) We have that

$$\begin{aligned}
\text{Err}(\boldsymbol{x}_0) =& E[(Y - \hat{f}(\boldsymbol{x}_0))^2|\boldsymbol{x} = \boldsymbol{x}_0] \\
=& E[(Y - f(\boldsymbol{x}_0) + f(\boldsymbol{x}_0) - E(\hat{f}(\boldsymbol{x}_0)) + E(\hat{f}(\boldsymbol{x}_0)) - \hat{f}(\boldsymbol{x}_0))^2|\boldsymbol{x} = \boldsymbol{x}_0] \\
=& E[(Y - f(\boldsymbol{x}_0))^2|\boldsymbol{x} = \boldsymbol{x}_0] + E[(f(\boldsymbol{x}_0) - E(\hat{f}(\boldsymbol{x}_0)))^2|\boldsymbol{x} = \boldsymbol{x}_0] + \\
& E[(E(\hat{f}(\boldsymbol{x}_0) - \hat{f}(\boldsymbol{x}_0)))^2|\boldsymbol{x} = \boldsymbol{x}_0] + \\
& 2E[(Y - f(\boldsymbol{x}_0))(f(\boldsymbol{x}_0) - E(\hat{f}(\boldsymbol{x}_0)))|\boldsymbol{x} = \boldsymbol{x}_0] + \\
& 2E[(Y - f(\boldsymbol{x}_0))(E(\hat{f}(\boldsymbol{x}_0) - \hat{f}(\boldsymbol{x}_0)))|\boldsymbol{x} = \boldsymbol{x}_0] + \\
& 2E[(f(\boldsymbol{x}_0) - E(\hat{f}(\boldsymbol{x}_0)))(E(\hat{f}(\boldsymbol{x}_0) - \hat{f}(\boldsymbol{x}_0)))|\boldsymbol{x} = \boldsymbol{x}_0] \\
=& E[(Y - f(\boldsymbol{x}_0))^2|\boldsymbol{x} = \boldsymbol{x}_0] + E[(f(\boldsymbol{x}_0) - E(\hat{f}(\boldsymbol{x}_0)))^2|\boldsymbol{x} = \boldsymbol{x}_0] + \\
& E[(E(\hat{f}(\boldsymbol{x}_0) - \hat{f}(\boldsymbol{x}_0)))^2|\boldsymbol{x} = \boldsymbol{x}_0]
\end{aligned}$$

where we have used that $\hat{f}(\boldsymbol{x}_0)$ is independent of a new observation Y and that $f(\boldsymbol{x}_0) - E(\hat{f}(\boldsymbol{x}_0))$ is just a constant. Since

$$\begin{aligned}
E[(Y - f(\boldsymbol{x}_0))^2|\boldsymbol{x} = \boldsymbol{x}_0] =& \sigma_\varepsilon^2, \\
E(\hat{f}(\boldsymbol{x}_0)) - f(\boldsymbol{x}_0) =& \text{Bias}(\hat{f}(\boldsymbol{x}_0)),
\end{aligned}$$

(Continued on page 2.)

and

$$E[(E(\hat{f}(\boldsymbol{x}_0)) - \hat{f}(\boldsymbol{x}_0))^2 | \boldsymbol{x} = \boldsymbol{x}_0] = \text{Var}(\hat{f}(\boldsymbol{x}_0)),$$

the result follow.

This result shows that

- It is never possible to get a prediction error smaller than the variance of the noise.
- The two other terms are related to how well we are able to estimate $\hat{f}(\boldsymbol{x}_0)$ and show that we can divide this part into one part corresponding to bias and one part corresponding to the variability in the estimate. Usually there will be a tradeoff between bias and variance, i.e. more complex models can reduce bias but increase variance due to limited data.

(b) Define

$$l^{ridge}(\boldsymbol{\beta}) = \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

We have

$$\begin{aligned}\frac{\partial}{\partial \beta_0} l^{ridge}(\boldsymbol{\beta}) &= -2\sum_{i=1}^{N}(y_i - \beta_0 - 2\sum_{j=1}^{p} x_{ij}\beta_j) \\ &= -2\sum_{i=1}^{N} y_i + 2N\beta_0\end{aligned}$$

giving $\hat{\beta}_0 = \bar{y}$.

Further, for $j > 0$,

$$\begin{aligned}\frac{\partial}{\partial \beta_l} l^{ridge}(\boldsymbol{\beta}) &= -2\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)x_{il} + 2\lambda\beta_l \\ &= -2\sum_{i=1}^{N} y_i x_{il} + 2\beta_0 \sum_{i=1}^{N} x_{il} + 2\sum_{j=1}^{p}\beta_j \sum_{i=1}^{N} x_{ij}x_{il} + 2\lambda\beta_l \\ &= -2\sum_{i=1}^{N} y_i x_{il} + 2\beta_l \sum_{i=1}^{N} x_{il}^2 + 2\lambda\beta_l\end{aligned}$$

giving $\hat{\beta}_l = \sum_{i=1}^{N} y_i x_{il} / [\lambda + \sum_{i=1}^{N} x_{il}^2]$.

We see that the $\hat{\beta}_j$'s are shrinked towards zero compared to the OLS estimates.

This term penalise large β_j values. By shrinking the estimates, we obtain less variability, but this can result in higher bias.

(Continued on page 3.)

Problem 2.

(a) Overfitting is a serious problem when performing estimation using complex models. Using the same data both to fit the model and to evaluate its performance will typically give a too optimistic estimate on the prediction/classification performance.

The usual way to avoid this is to put aside a part of the dataset that is not to be used for estimation but only for testing the performance, i.e. the test set. The set used for doing the estimation is called the training set.

For many methods, there are tuning (penalty/shrinkage/complexity) parameters that need to be specified or models to be selected. Using the training set to specify these turning parameters will have the same problems with overfitting. On the other hand, using the test set for this, will lead to no data that can evaluate the final selected model. A way to obtain this is to define a separate validation set for selecting tuning parameters.

For small datasets, such a split can be problematic in that neither of the datasets will be large enough for doing their task (estimation or validation).

(b) Cross-validation is a method which utilises the data much better. Typically it is used for avoiding a split between training and validation. In this case the training set is split into K, say, parts. $K-1$ of these parts are used for estimation while the last part is used for validation. By looping through all the K possible parts that can be used for validation, we obtain a validation based on the whole training set, while each estimation is based on a fraction $(K-1)/K$ of the training set.

Problem 3.

(a) We have Likelihood

$$L = \prod_{i=1}^{n} p_{m(\boldsymbol{x}),y_i}$$

(Continued on page 4.)

giving log-likelihood

$$\begin{aligned} l &= \sum_{i=1}^{n} \log p_{m(\boldsymbol{x}),y_i} \\ &= \sum_{i=1}^{n} \sum_{m=1}^{M} \sum_{k=1}^{K} \log p_{m,k} I(\boldsymbol{x}_i \in R_m, y_i = k) \\ &= \sum_{m=1}^{M} \sum_{k=1}^{K} \log p_{m,k} \sum_{i=1}^{n} I(\boldsymbol{x}_i \in R_m, y_i = k) \\ &= \sum_{m=1}^{M} \sum_{k=1}^{K} \log p_{m,k} N_{m,k} \end{aligned}$$

which shows that maximizing the log-likelihood corresponds to minimizing Q.

(b) $\hat{p}_{m,k}$ is fraction of observations within region R_m that belongs to class k and is both an unbiased and the maximum likelihood estimate (given the regions) for $p_{m,k}$.

Inserting $\hat{p}_{m,k}$, we get

$$\begin{aligned} Q &= -\sum_{m=1}^{M} \sum_{k=1}^{K} N_m \hat{p}_{m,k} \log \hat{p}_{m,k} \\ &= \sum_{m=1}^{M} N_m Q_m(T) \end{aligned}$$

where

$$Q_m(T) = -\sum_{k=1}^{K} \hat{p}_{m,k} \log \hat{p}_{m,k}$$

This corresponds to the cross-entropy or deviance measure used for classification trees.

(c) Searching for regions R_m is a difficult task due to the many possibilities in how this can be done. The great flexibility also increase the possibility of overfitting. By using a tree structure where regions are defined by sequential splits, the possibilities of regions are significantly reduced. By further restricting the splits to only depend on one variable at a time, this is further reduced.

Another benefit in this tree-structure splitting is that we get a readable model.

END