UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Examination in:	STK4030 — Modern data analysis - FASIT
Day of examination:	Friday 13. Desember 2013.
Examination hours:	14.30 – 18.30.
This examination set consists of 5 pages.	
Appendices:	None
Permitted aids:	Approved calculator

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

Problem 1.

(a) We have that

$$\begin{aligned}
E[(Y-\widehat{Y})^2] =& E[(Y-E[\widehat{Y}]-E[\widehat{Y}]-E[\widehat{Y}]+E[\widehat{Y}]-\widehat{Y}(\boldsymbol{X}))^2] \\
=& E[(Y-E[Y|\boldsymbol{X}])^2]+E[(E[Y|\boldsymbol{X}]-E[\widehat{Y}])^2]+E[(E[\widehat{Y}]-\widehat{Y}(\boldsymbol{X}))^2]+ \\
& 2E[(Y-E[Y|\boldsymbol{X}])(E[Y|\boldsymbol{X}]-E[\widehat{Y}])]+ \\
& 2E[(Y-E[Y|\boldsymbol{X}])(E[\widehat{Y}]-\widehat{Y}(\boldsymbol{X}))]+ \\
& 2E[(E[Y|\boldsymbol{X}]-E[\widehat{Y}])(E[\widehat{Y}]-\widehat{Y}(\boldsymbol{X}))] \\
=& \mathrm{Var}(Y)+[\mathrm{Bias}(\widehat{Y}(X))]^2+\mathrm{Var}(\widehat{Y})
\end{aligned}$$

(all the cross-terms becomes zero).

This result shows that there typically will be a tradeoff between variance and bias (the first term is independent of the choise of $\widehat{Y}$). Predictors with less constraints will have small bias but can potentially have large variance, while more constrained predictors will have less variance but can have large bias if the constraints do not fit well with how the data is generated.

(Continued on page 2.)

We further have

$$E[L[Y,\widehat{Y}(\boldsymbol{X})]] = E_{\boldsymbol{X}} E_{Y|\boldsymbol{X}}[L[Y,\widehat{Y}(\boldsymbol{X})]|\boldsymbol{X}]$$

and it is enough to minimize $E_{Y|\boldsymbol{X}}[L[Y,\widehat{Y}(\boldsymbol{X})]|\boldsymbol{X}]$ for each $\boldsymbol{X}$. We have

$$\begin{aligned} E_{Y|\boldsymbol{X}}[L[Y,\widehat{Y}(\boldsymbol{X})]|\boldsymbol{X}] =& E[(Y-\widehat{Y}(\boldsymbol{X}))^2|\boldsymbol{X}] \\ =& E[(Y-E[Y|\boldsymbol{X}]+E[Y|\boldsymbol{X}]-\widehat{Y}(\boldsymbol{X}))^2|\boldsymbol{X}] \\ =& E[(Y-E[Y|\boldsymbol{X}])^2|\boldsymbol{X}]+(E[Y|\boldsymbol{X}]-\widehat{Y}(\boldsymbol{X}))^2 \end{aligned}$$

The prediction is minimized with $\widehat{Y}(\boldsymbol{x}) = E[Y|\boldsymbol{X}=\boldsymbol{x}] = f(\boldsymbol{x})$.

(b) The first predictor is based on a linear model. It will have small variance due to hard constraints on the model. But if $f(\boldsymbol{X})$ is far from linear, the bias can be severe. The other estimator puts less constraints on the model, although it will depend on the choice of k. For small k, the bias will be small, but the variance will be high. For larger k, the bias decreases at the cost of higher variance. The degrees of freedom in the linear model is p while the degrees of freedom in the k-nearest neighbor predictor is N/k. Typically N/k needs to be smaller than p in order to obtain flexibility.

Problem 2.

(a) This is Ridge regression. It puts penalties on the β_j's shrinking these to zero compared to least squares. λ is a complexity parameter that controls the amount of shrinkage with $\lambda = 0$ corresponding to ordinary least squares. Such a method can reduce the variance but also increase the bias. λ can be chosen by e.g. cross-validation. Note that there is no penalty on β_0, making no restriction on the overall level.

(b) We can write the criterion as

$$RSS(\lambda) = (\boldsymbol{y}-\beta_0\mathbf{1}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\beta_0\mathbf{1}-\boldsymbol{X}\boldsymbol{\beta})+\lambda\boldsymbol{\beta}^T\boldsymbol{\beta}$$

Then

$$\frac{\partial}{\partial\beta_0}RSS(\lambda) = -2((\boldsymbol{y}-\beta_0\mathbf{1}-\boldsymbol{X}\boldsymbol{\beta})^T\mathbf{1}$$

showing that

$$\hat{\beta}_0 = \bar{y}-\bar{\boldsymbol{x}}^T\boldsymbol{\beta}.$$

(Continued on page 3.)

Defining $\tilde{y}_i = y_i - \bar{y}$ and $\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i - \bar{\boldsymbol{x}}$, the criterion is then modified to

$$RSS(\lambda) = (\tilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta})^T(\tilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}$$

and

$$\frac{\partial}{\partial \boldsymbol{\beta}^T} RSS(\lambda) = -2\widetilde{\boldsymbol{X}}^T(\tilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta}) + \lambda 2\boldsymbol{\beta}$$

giving

$$\hat{\boldsymbol{\beta}} = \left[\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}} + \lambda \boldsymbol{I}\right]^{-1}\widetilde{\boldsymbol{X}}^T\tilde{\boldsymbol{y}}$$

(c) This is called Lasso regression. In this case an L_1 norm penalty is used. This has the effect that some of the β_j's are shrinked exactly to zero and thereby works as a variable selector. This is in contrast to Ridge regression where the β_j's only are downweighted.

Given these distinct properties, the left plot corresponds to Lasso and the right to Ridge. In both cases the rightmost values correspond to least squares.

(d) Can either be chosen by considering a validation set or by cross-validation.

Problem 3.

(a) When we consider decision boundaries, these are given by

$$\Pr(G = 1|\boldsymbol{X} = \boldsymbol{x}) = \Pr(G = 2|\boldsymbol{X} = \boldsymbol{x})$$

which is equivalant to that

$$\exp(\beta_0 + \boldsymbol{\beta}^T\boldsymbol{x}) = 1$$

which again is equivalant to that

$$\beta_0 + \boldsymbol{\beta}^T\boldsymbol{x} = 0$$

which is a linear constraint and thereby we get a linear boundary.

(Continued on page 4.)

(b) The typical method is maximum likelihood, that is maximize

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{N} p_{y_i}(\boldsymbol{x}_i; \boldsymbol{\beta})$$

where $p_{y_i}(\boldsymbol{x}_i; \boldsymbol{\beta}) = \Pr(G = y_i | \boldsymbol{X} = x_i)$. Equivalently we can minimize

$$-2 \log L(\boldsymbol{\beta}) = -2 \sum_{i=1}^{N} \log p_{y_i}(\boldsymbol{x}_i; \boldsymbol{\beta})$$

Note that Ridge regression with Gaussian observations corresponds to minimizing

$$-2 \log L(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \beta_j^2$$

Penalities on the β_j's in a logistic regression setting can be done similarly. Also Lasso penalties can be done in the same way.

Problem 4.

(a) All variables except "regn1uke" seems to give significant contribution to the prediction, All the other variables show nonlinear relationships. The first four variables give increased responses when the explanatory variables increases although all except "hastighet" seems to flatten out for large values. For the first three variables is is certainly reasonable that we have a postive effect. The relation to temparature is not that obvious. For "regn4time" is seems like there is a more discrete type of behaviour. When this value is below a certain threshold, we get high values (with a linear decrease), while it decreases to a stable value for values above 1. This is also reasonable in that when there is rain, this will force the particles to the ground, and it is enough with a certain hour. It also makes sence that it is the most recent precipitation that matters, not an average of the last week.

(b) The data do not contain speed values over 60 km/hour, but by extrapolating the relationship for "hastighet" (which seems linear above 50), we can predict that an increase of 20 km/hour corresponds to an increase of 0.3 in the respons (using that it seems like the increase from 0.5 to 0.6 is about 0.15). There is however a huge uncertainty in this prediction since we need to assume that the behaviour within 50-60 continues for higher speeds!

(Continued on page 5.)

Problem 5.

(a) The model is described by

$$Y = f(\boldsymbol{x}) + \varepsilon = \beta_0 + \sum_{m=1}^{M} g_m(\boldsymbol{\omega}_m^T \boldsymbol{x}) + \varepsilon$$

where β_0 is an intercept term (sometimes dropped when we write the model), the $g_m(\cdot)$'s are smooth functions that are to be estimated from the data and the ω_m's are coefficient (or direction) vectors that defines lnear combinations of the explanatory variables. Finally ε is an error term.

Thus, the response is described by a sum of non-linear functions of linear combinations of the explanatory variables.

The tuning parameters are the number of terms M and the smoothing parameter(s) for the non-linear functions. Given these, the model is estimated by least squares.

(b) Interpretation: This model has two terms. The first term is essentially the difference between a weighted sum of x_1, x_2 and x_3 and x_4 . The response is increasing exponentally by this linear combination. The second term is essentially the difference between a weighted sum of x_1, x_2 and x_4 and x_3. The response is a quadratic function of this linear combination.

The model explains the relationship between y and the x-es reasonably well, because the originl standard deviation of y is 3.76, and this is reduced to a residual standard error of 0.93, and a cross-validated RMSE of 0.98, both much lower than 3.76.

A GAM model with only univariate (one x per time) terms would probably not model the relationship beetween y and the x-es equally well as the PPR model. This is because i) the scatter plot matrix show only rough relationships between y and each of the x-es, and no non-linear relationships, whereas the PPR model shows that y is non-linearly dependent on differences between the x-es.

END