UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	STK2100 — Machine learning and statistical methods for prediction and classification
Day of examination:	Thursday June 2017.
Examination hours:	09.00 – 13.00.
This problem set consists of 7 pages.	
Appendices:	Ingen
Permitted aids:	Approved calculator and List of formulas for STK1100/STK1110 and STK2100

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

Problem 1

In this exercise we will look at a data set for housing prices in suburbs to Boston. As a response variable we will have

MEDV Median value of housing within an area (in $1000)

There are 11 explanatory variables. It is not important to understand what these are in the following questions, but a description of these is given below. All variables, except CHAS (binary) and RAD (categorically with 9 levels) are numeric.

CRIM per capita crime rate by town

ZN proportion of residential land zoned for lots over 25,000 sq.ft.

CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)

NOX nitric oxides concentration (parts per 10 million)

RM average number of rooms per dwelling

AGE proportion of owner-occupied units built prior to 1940

DIS weighted distances to five Boston employment centres

RAD index of accessibility to radial highways

PTRATIO Pupil-teacher ratio by town

B $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town

(Continued on page 2.)

LSTAT % lower status of the population

There are a total of $n = 506$ areas (observations), but all the analysis below is based on the division into a training set and a test set.
The division in the training and test sets is done by randomly drawing 253 observations used for training and the remaining for testing. As a reference, we will use results from a linear regression model. The least square method gave the following regression table:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	38.071527	7.841521	4.855	2.20e−06
CRIM	−0.140742	0.040019	−3.517	0.000524
ZN	0.047965	0.019805	2.422	0.016204
CHAS1	5.205026	1.401956	3.713	0.000256
NOX	−16.911410	5.702240	−2.966	0.003332
RM	2.614760	0.617887	4.232	3.33e−05
AGE	0.019537	0.019267	1.014	0.311638
DIS	−1.262763	0.294073	−4.294	2.57e−05
RAD2	4.604395	2.072367	2.222	0.027254
RAD3	6.099546	1.914405	3.186	0.001638
RAD4	2.636360	1.776669	1.484	0.139187
RAD5	4.019631	1.781579	2.256	0.024981
RAD6	2.250257	2.255377	0.998	0.319441
RAD7	6.901910	2.114468	3.264	0.001262
RAD8	5.704424	2.151248	2.652	0.008557
RAD24	7.544125	1.989127	3.793	0.000190
PTRATIO	−0.983504	0.213207	−4.613	6.54e−06
B	0.007554	0.003570	2.116	0.035400
LSTAT	−0.695442	0.073822	−9.421	< 2e−16

The log likelihood value for this linear model is -742.34. The average error rate for the test data based on this model is 25.14.

(a) Why is it advisable to divide into training and test sets randomly in relation to other strategies?

Given this table, arguments why it may be reasonable to remove AGE from the model.

(b) Below is a regression table given where AGE is removed.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	37.268866	7.801931	4.777	3.14e−06
CRIM	−0.141442	0.040016	−3.535	0.000492
ZN	0.046455	0.019750	2.352	0.019493
CHAS1	5.339371	1.395765	3.825	0.000167
NOX	−15.305548	5.478227	−2.794	0.005637
RM	2.719095	0.609295	4.463	1.26e−05
DIS	−1.359251	0.278268	−4.885	1.92e−06

(Continued on page 3.)

RAD2	4.647958	2.072045	2.243	0.025818
RAD3	6.042215	1.913684	3.157	0.001800
RAD4	2.556993	1.775050	1.441	0.151052
RAD5	4.009193	1.781656	2.250	0.025358
RAD6	2.223959	2.255363	0.986	0.325110
RAD7	6.935950	2.114328	3.280	0.001194
RAD8	5.818663	2.148425	2.708	0.007259
RAD24	7.358077	1.980765	3.715	0.000254
PTRATIO	−0.955146	0.211378	−4.519	9.86e−06
B	0.007894	0.003554	2.221	0.027306
LSTAT	−0.665377	0.067610	−9.841	< 2e−16

The log likelihood value for this model is -742.90 whereas the average error rate on the test data based on this model was 24.86.

If you ignore the test data, perform a procedure for model selection between the two linear regression models. What is the conclusion of this? Does it seem reasonable compared to what came out on the test data?

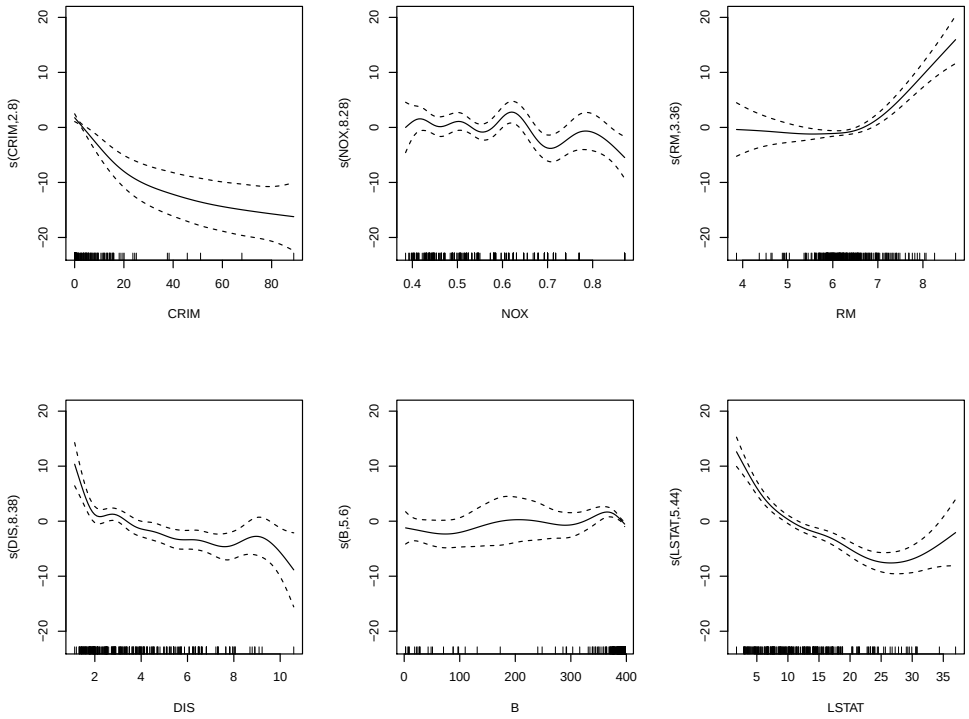
Based on the printout, could you imagine a further simplification of the model?

(c) An alternative model is GAM. The plots below show the nonlinear features included in this model. The log likelihood value for this model is -615.79 while the estimated number of degrees of freedom is 46.12.

Explain how the number of degrees of freedom is calculated in this case. Use this to compare this model with the previous models. Comment on the result.

The average error rate for test data based on this model was 15.52. Is this in accordance with the model comparisons you have made?

(Continued on page 4.)



(d) Another alternative model can be obtained using regression trees. Below is a plot of a regression tree based on 9 terminal nodes (or leaves).

Discuss why regression trees provide an opportunity to include *interactions* between explanatory variables.

Explain why a reasonable likelihood function in this case is

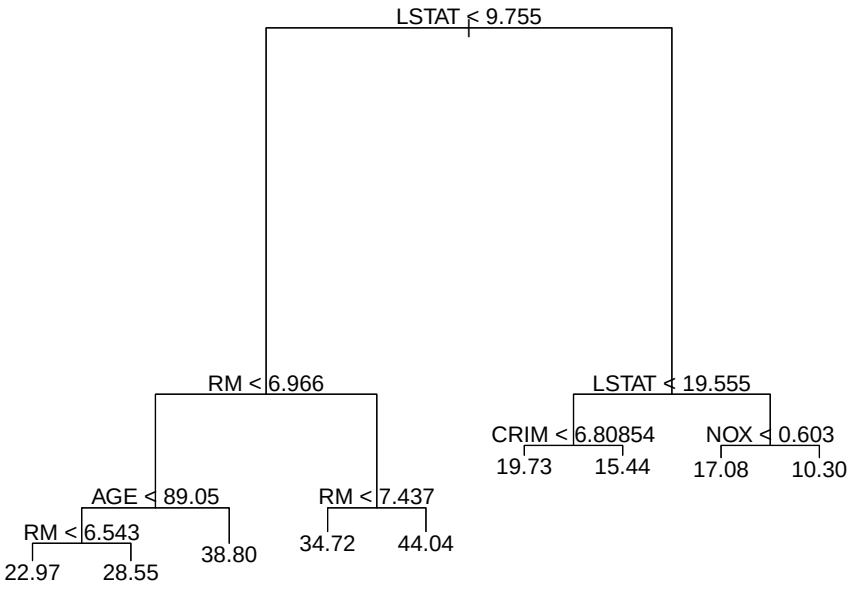
$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i - \mu_i)^2}$$

where $\mu_i = c_m$ if $\boldsymbol{x}_i \in R_m$. On what assumptions is such a likelihood based on?

How many parameters must be specified to fit a tree with 9 enddenodes?

Compare this model against previous models when the log-likelihood value (with estimated values for $\boldsymbol{\theta}$ inserted) in this case is -697.49.

(Continued on page 5.)



(e) Alternative methods like Bagging, Random Forest and Boosting gave the following results (where Error is estimated square error on the test set):

Method	Error
Regresjons tree	17.16
Bagging	11.36
Random Forrest	11.28
Boosting	11.58

Describe **short** these three methods and comment on the results. Discuss in particular the improvements in relation to the GAM model.

Problem 2

We will now look at a classification setting. Let $Y \in \{1, ..., G\}$ be the variable of interest and suppose we observe $\boldsymbol{x} \in \mathcal{R}^p$. We have as usual data $\{(y_i, \boldsymbol{x}_i), i = 1, ..., n\}$ where $y_i \in \{1, ..., G\}$ while $\boldsymbol{x}_i \in \mathcal{R}_i$. We want to predict Y based on $\boldsymbol{x}$.

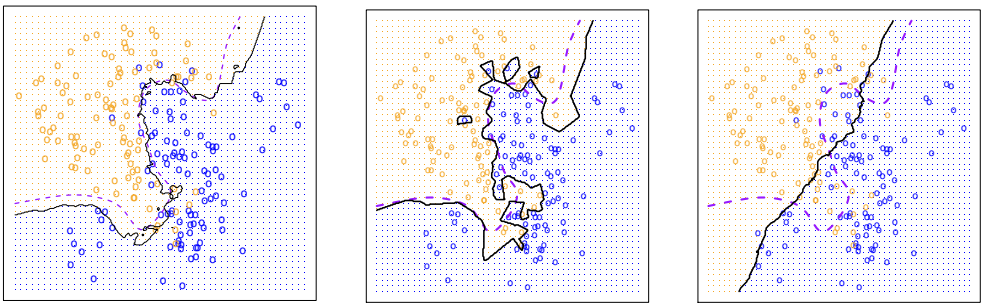
(a) Suppose we want to use the K nearest neighbor method for classification.

Explain how this method works. What are the strengths and weaknesses of this method?

(Continued on page 6.)

(b) Below are 3 plot of the K nearest neighbor method for $G = 2$ and $K = 1, 10$ or 100 (but not necessarily in this order). Here the solid line gives the classification boundary between the two classes while the dotted line gives the optimal boundary (data here are simulated so that we know the underlying true model). The colors of the points and areas indicate class values for observations and classifications, respectively.

Specify which plots that belong to the different K values. What value of K would you prefer? Give reasons for your answer.



a b c

Assume now that we introduce a loss function

$$L(y, \hat{y}) = \begin{cases} 1 & \text{hvis } \hat{y} \neq y; \\ 0 & \text{ellers.} \end{cases}$$

which says something about how serious we measure errors that are made.

(c) Show that the optimal predictor in this situation is

$$\widehat{Y}(\boldsymbol{x}) = \arg\max_g \Pr(Y = g|\boldsymbol{x}).$$

Explain why it therefore is important to estimate $f_g(\boldsymbol{x}) = E[I(Y = g)|\boldsymbol{x}]$ for $g = 1, ..., G$ where $I(A) = 1$ if the event A is true and 0 otherwise.

(d) Explain how noe can use *regression methods* for estimating $f_g(\boldsymbol{x})$ and thus use regression methods to construct classification methods.

(e) Discuss different methods for estimating expected losses in this classification setting. Take special attention to strengths and weaknesses with different methods.

(f) Assume now that $\hat{f}_g(\boldsymbol{x})$ is an estimate of $f_g(\boldsymbol{x})$. Show that

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

Provide an interpretation of the various terms on the right side.

(Continued on page 7.)

(g) Now let $\hat{f}_{g,1}(\boldsymbol{x})$ be an estimate of $f_b(\boldsymbol{x})$ based on a fairly restrictive method/model while $\hat{f}_{g,2}(\boldsymbol{x})$ is based on a more flexible approach. Discuss the different terms in the equation above in this setting.

Problem 3

In Ridge regression we want to minimize with respect to $\boldsymbol{\beta}$

$$h(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

We will assume the explanatory variables are centered so that $\sum_{i=1}^{n} x_{ij} = 0$ for all j.

(a) Explain why is is also reasonable to scale the x_{ij}'s such that $\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1$ for all j.

(b) Show that

$$\begin{aligned} \hat{\beta}_0^{ridge} =& \bar{y} \\ \hat{\boldsymbol{\beta}}^{ridge} =& \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y} \end{aligned}$$

for a suitable specification of $\boldsymbol{X}$. Here $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$.

(c) Assume now all the x's are uncorrelated so that $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}$. Assume also that the true model is $Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_j$ where $\varepsilon_1, ..., \varepsilon_p$ are independent with expectation 0 and variance σ^2.

Derive in this case the expectation vector and the covariance matrix for $\hat{\boldsymbol{\beta}}^{ridge}$.

Discuss these results in relation to the trade-offs that we usually do in regression settings.

Hint: Show first that $E[\boldsymbol{Y}] = \beta_0\mathbf{1} + \boldsymbol{X}\boldsymbol{\beta}$ where $\mathbf{1}$ is a vector of 1's and that $\boldsymbol{X}^T\mathbf{1} = \mathbf{0}$.