UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	STK2100 — Machine Learning and Statistical Methods for Prediction and Classification
Day of examination:	May 31 - 2023
Examination hours:	15.00 – 19.00.

This problem set consists of 8 pages.

Appendices:	List of formulas for STK1100/STK1110 and STK2100
Permitted aids:	Approved calculator

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

There are three problems:

- Problem 1 deals with regression
- Problem 2 deals with classification
- Problem 3 is a more mathematical exercise related to linear regression

The three problems can be solved independently, but you need to read through the description of the College data in Problem 1 in order to solve Problem 2.
All subquestions are counted equally!
When commenting on results, include arguments for your answers.

Problem 1

We will in this exercise look at a dataset giving statistics for a large number of US Colleges from the 1995 issue of US News and World Report. There are two main reponse variables of interest:

Apps Number of applications received

Accept Number of applications accepted

There are in total 16 covariates, one categorical and 15 numerical. All of these are related to features of the college, none are related to the individual applicants. A list of these are given at the end of the whole exam exercise, but the actual meaning of these are not important for answering the different questions.
We will in this problem consider models/method for predicting the number of applications received. There are in total 776 observations, ordered alphabetically according to the names of the colleges.

(Continued on page 2.)

(a) $n_{tr} = 388$ are randomly selected for training while the remaining 388 observations are saved for validation.

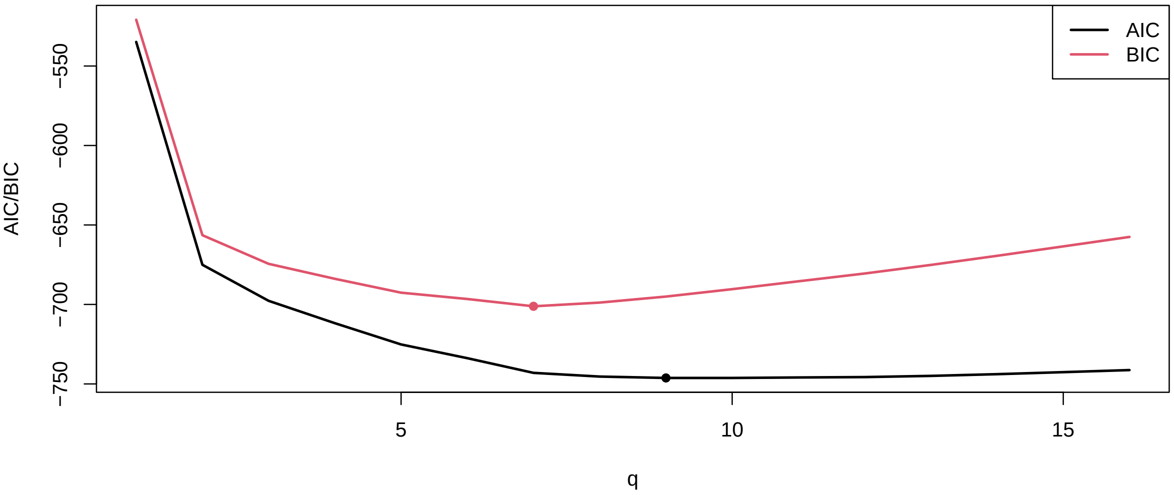
Discuss strengths and weaknesses in dividing the data set into such subsets.

Why is it important to do this division randomly?

(b) Assume now linear regression models for prediction of **Apps**. The plot below shows AIC and BIC values (based on the training data) for different number of covariates q where for each value of q the best subset is shown.

Explain why it is reasonable that after a decrease, the curves seems to increase.

Aslo explain why it is reasonable that the q giving the minimum value of AIC is larger than the value of q giving the minimum value if BIC.



The table below shows a regression table based on a linear model using the 9 variables selected by the AIC criterion (left) and 7 variables selected by the BIC criterion (right). The root mean square error (RMSE) on the test data was 1356.735 and 1361.025 for the AIC and BIC models, respectively.

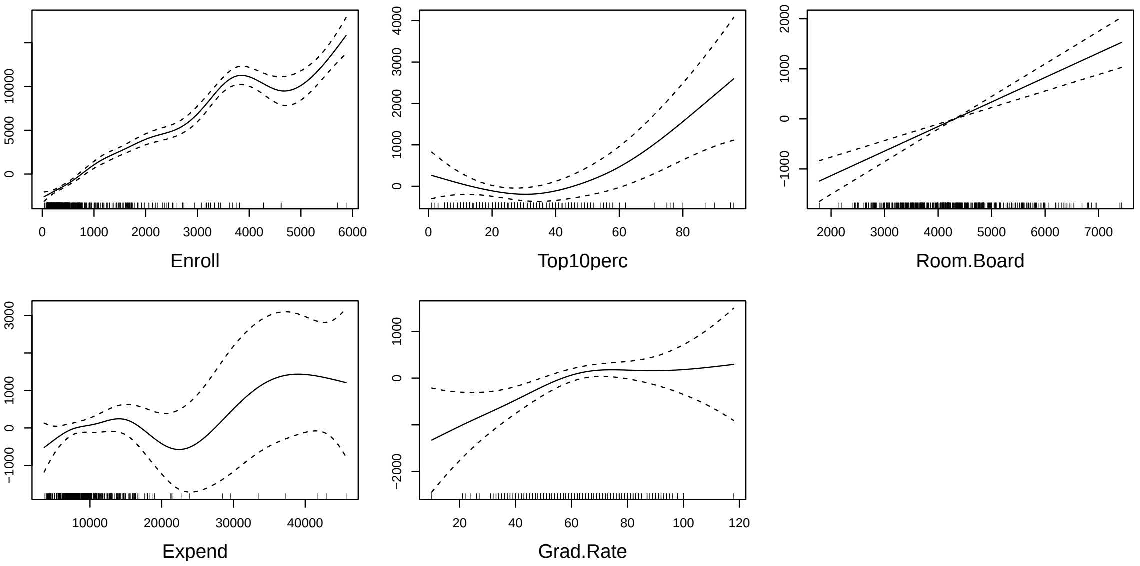
	AIC model				BIC model			
	Estimate	Std. Error	t value	P-value	Estimate	Std. Error	t value	P-value
(Intercept)	-2889.283	468.334	-6.169	0.000	5.673	0.143	39.736	0.000
PrivateYes	-602.801	245.631	-2.454	0.015	-0.619	0.086	-7.194	0.000
Enroll	2.374	0.284	8.354	0.000	0.001	0.000	17.520	0.000
Top10perc	4.496	6.521	0.689	0.491	0.006	0.002	2.455	0.015
F.Undergrad	0.178	0.056	3.207	0.001				
Outstate	0.036	0.033	1.082	0.280				
Room.Board	0.472	0.087	5.450	0.000	0.000	0.000	6.044	0.000
PhD	-5.552	5.604	-0.991	0.322				
Expend	0.060	0.020	2.960	0.003	0.000	0.000	2.335	0.020
Grad.Rate	15.641	5.367	2.914	0.004	0.010	0.002	4.845	0.000
perc.alumni					-0.006	0.003	-2.085	0.038

(Continued on page 3.)

(c) Discuss possible reasons for why some variables selected by the BIC criterion is not included in the AIC model and vice versa.

Further, discuss why the P-values related to some of the variables that are common in both models have (in some cases very large) different values.

Consider now generalized additive models (GAMs). For simplicity, we will only consider the variables selected by the BIC criterion above. Further, in all cases, smoothing splines are used. The plot below shows the nonlinear functions obtained by a fit to the training data. The RMSE on the test data was 1263.51 in this case.



The table below shows degrees of freedom and AIC/BIC values for the GAM fit corresponding to the plot above (first line) and then five alternative models where the non-linear parts are turned off on the variable listed (so Enroll lin means that Enroll is included as a linear term while all the others have a non-linear term).

	df	AIC	BIC
All non-lin	22.76	6688.54	6778.70
Enroll lin	16.11	6708.49	6772.28
Top10perc lin	21.00	6701.77	6784.94
Room.Board lin	22.74	6688.57	6778.66
Expend lin	19.18	6690.49	6766.45
Grad.rate lin	21.23	6692.74	6776.83

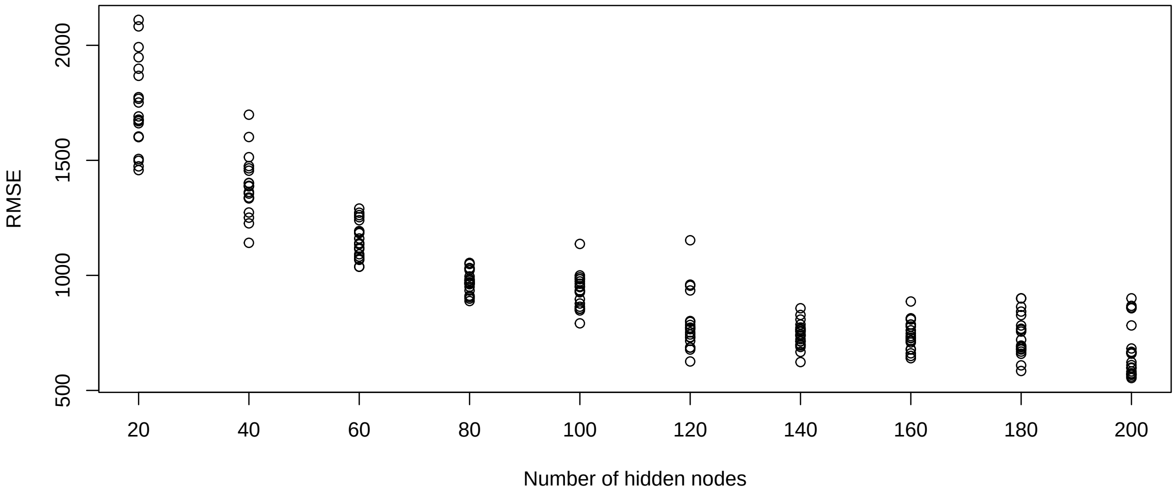
(d) Explain how the degrees of freedom are calculated for GAM models.

Do you find the AIC/BIC values reasonable based on the plots of the estimated non-linear effects in the plot above?

We will finally consider neural networks. One hidden layer, but varying number of hidden nodes was tried out. The decay parameter was equal to 1 in this case. For each setting, 20 repetitions were performed. The plot below shows the results where each point correspond

(Continued on page 4.)

to the RMSE on the test data. The best RMSE value (553.82) was obtained for 200 hidden nodes.



(e) Discuss these results in relation to properties of neural networks.

How would you choose the number of hidden nodes to be used for your final model? And what would you say about the performance of your chosen model?

Problem 2

We will in this problem continue to consider the College data set but now focus on the **Accept** reponse variable. For this problem, denote by y_i the number of applications accepted for college i and n_i the corresponding number of applicants received.

(a) Discuss why it may be reasonable to assume that y_i is binomial distributed with n_i trial and a success probability p_i.

Also discuss possible reasons for why the binomial distribution assumption may be violated.

In practice we predict y_i by $\hat{y}_i = n_i\hat{p}_i$ where $\hat{p}_i$ is some estimate of the probability p_i. Since the y_i's are numerical numbers, this can be seen as some kind of regression problem. We will however see how this can be related to classification on individual applicant level. Consider one specific college, i say. Denote by z_{ij} the binary variable equal to 1 if applicant j on college i is accepted and 0 otherwise (so the total number of accepted $y_i = \sum_j z_{ij}$). In the dataset only y_i is available, not z_{ij}. Assume we have some classification rule which for each individual j we make a classification to $\hat{z}_{ij} \in \{0, 1\}$ and denote by $\hat{y}_i = \sum_j \hat{z}_{ij}$. Within each College we can then make a confusion matrix

$z_{ij}\backslash\hat{z}_{ij}$	0	1
0	N_i^{00}	N_i^{01}
1	N_i^{10}	N_i^{11}

(Continued on page 5.)

where $N_i^{k\ell}$ is the numbers where $z_{ij} = k$ and $\hat{z}_{ij} = \ell$. Note that $y_i = N_i^{10} + N_i^{11}$ while $\hat{y}_i = N_i^{01} + N_i^{11}$.

(b) With a ordinary zero-one loss function on the individual level, the number of errors would be $N_i^{01} + N_i^{10}$. However, if we are making predictions at college level we would look at $y_i - \hat{y}_i$ instead. Express this difference by the $N_i^{k\ell}$ terms and discuss these results.

Assuming the z_{ij}'s were available, the likelihood can (under certain assumptions) be written as

$$L = \prod_i \prod_j p_{ij}^{z_{ij}} (1 - p_{ij})^{1 - z_{ij}}$$

where i is an index for college and j is an index for individuals within college. Further p_{ij} is the probability for individual j to be accepted.

(c) Specify which assumptions this likelihood expression relies on.

The probabilities p_{ij} are assumed to be a function of the covariates available. When all covariates are only related to college, not to individuals, argue why $\hat{p}_{ij} = \hat{p}_i$ and that the likelihood then only depend on y_i and n_i and not the individual z_{ij}'s.

We will now consider using trees for estimation of p_i and prediction of y_i. A problem with many procedures for fitting trees is that they require the data to be available at individual levels (z_{ij}'s), not aggregated in groups such as the data considered here (y_i's and n_i's).

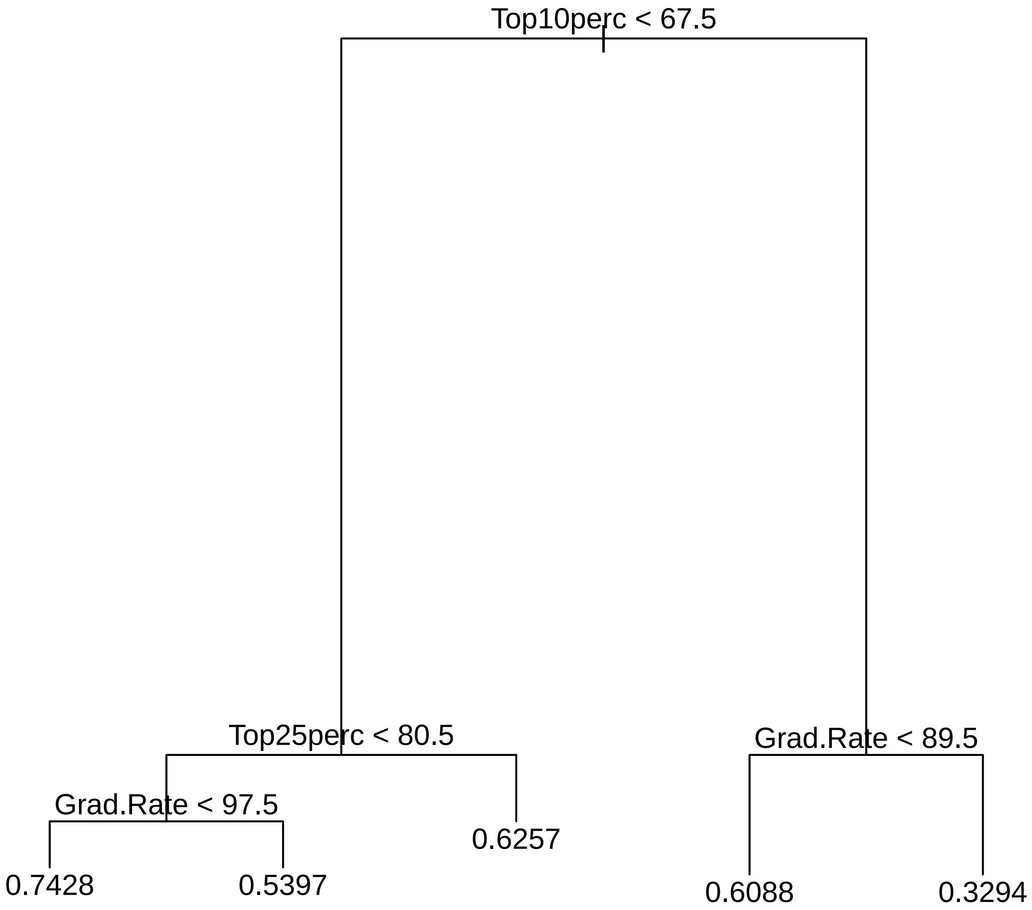
(d) Explain how you can construct a training dataset which do have information on individual level for the data at hand.

Assume now that you make a prediction for z_{ij} such that $z_{ij} = 1$ if the estimated $p_{ij} > 0.5$ and zero otherwise. Explain why $\hat{y}_i$ then either will be 0 or n_i.

Based on this, argue why a better prediction is $\hat{y}_i = \sum_j \hat{p}_{ij}$.

The plot below shows a fitted tree to the training data (where one goes to the left if the criterion is satisfied).

(Continued on page 6.)



The table below shows the covariates for Abilene Christian University. In addition, the number of applicants were 1660 while the number of accepted were 1232 for this college.

Private	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board
Yes	721	23	52	2885	537	7440	3300
Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
450	2200	70	78	18.10	12	7041	60

(e) Use the tree to make a prediction on the number of accepted applicants.

Does the prediction do well in this case?

What is the main advantage of tree classifiers compared to for instance logistic regression?

(Continued on page 7.)

Problem 3

Assume a linear regression model

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, w_i\sigma^2), \qquad (*)$$

where w_i are known quantities and $i = 1, ..., N$. We can write this model in vector/matrix form:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{W}).$$

Here $\boldsymbol{W}$ is a diagonal matrix with w_i on the ith diagonal.

(a) Argue why the *weighted* least squares criterion RSS $= \sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2/w_i$ is reasonable to use in this case.

Show that this criterion also can be written as

$$\text{RSS} = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{W}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}).$$

(b) Show that the optimal estimate in this case becomes

$$\widehat{\beta} = [\boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{Y}$$

Also show that $\hat{\beta}$ is an unbiased estimate of $\boldsymbol{\beta}$.

What is the covariance matrix for $\hat{\boldsymbol{\beta}}$?

(c) Argue that you can reformulate the model (*) to a standard linear regression model (that is with equal variances on the noise terms) by dividing both sides by $\sqrt{w_i}$.

Based on this, argue why the results that was shown in (b) are reasonable.

Description of covariates for the College data

Private A factor with levels No and Yes indicating private or public university

Enroll Number of new students enrolled

Top10perc Pct. new students from top 10% of H.S. class

Top25perc Pct. new students from top 25% of H.S. class

F.Undergrad Number of fulltime undergraduates

P.Undergrad Number of parttime undergraduates

Outstate Out-of-state tuition

Room.Board Room and board costs

(Continued on page 8.)

Books Estimated book costs

Personal Estimated personal spending

PhD Pct. of faculty with Ph.D.'s

Terminal Pct. of faculty with terminal degree

S.F.Ratio Student/faculty ratio

perc.alumni Pct. alumni who donate

Expend Instructional expenditure per student

Grad.Rate Graduation rate