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 14 2018.
Examination hours:	14.30 – 18.30.

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

We will in this exercise look at a dataset on the survival after the Titanic catastrophy.
The variables available are

Survival 0=No, 1=Yes, factor

Age Age of in months, a numerical variable

Pclass Ticket class, 1=1st, 2=2nd, 3=3rd, factor

Sex Sex (male/female), factor

Sibsp Number of siblings/spouses onboard, numerical.

Parch Number of parents/children onboard, numerical.

Fare Ticketprice, numerical.

Cabin Cabin number, factor which originally had 148 different values, but which is reduced to 9; N (no cabin), A, B, C, D, E, F, G, T.

Embarked Harbour for embarking, C=Cherbourg, Q=Queenstown, S=Southampton, factor.

We will consider a subset of the total set consisting of 712 individuals.
We will start with a simple logistic regression model. Fitting of such a model gave the following table:

(Continued on page 2.)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.8723	0.6692	5.79	0.0000
Pclass2	-0.6793	0.5053	-1.34	0.1788
Pclass3	-1.8027	0.5182	-3.48	0.0005
Sexmale	-2.6900	0.2279	-11.80	0.0000
Age	-0.0439	0.0085	-5.15	0.0000
SibSp	-0.3553	0.1306	-2.72	0.0065
Parch	-0.0691	0.1251	-0.55	0.5805
Fare	0.0029	0.0030	0.97	0.3298
CabinA	1.1274	0.7877	1.43	0.1524
CabinB	0.5580	0.6381	0.87	0.3819
CabinC	-0.0680	0.5821	-0.12	0.9070
CabinD	0.9392	0.6146	1.53	0.1265
CabinE	1.5267	0.6049	2.52	0.0116
CabinF	1.2172	0.7936	1.53	0.1251
CabinG	-0.8919	1.0124	-0.88	0.3783
EmbarkedQ	-0.7989	0.6051	-1.32	0.1867
EmbarkedS	-0.4351	0.2838	-1.53	0.1252

When we use this model to predict the same data (by predicting to the most probable class), we obtain an error rate of 19.10%. The log-likelihood value for this modellen is -308.8.

(a) Explain why the regression model lists fewer rows than the number of levels for factor variables.

Given that we here have a "Treatment" constraint (we put the coefficient related to the first level to zero), what kind of interpretation do then the regression coefficients have for the other levels?

(b) Calculate the AIC-value for this model. Discuss why it may be reasonable to simplify the model somewhat.

(c) Below is a regression table based on an alternative model:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	4.3254	0.4507	9.60	0.0000
Pclass2	-1.4063	0.2848	-4.94	0.0000
Pclass3	-2.6450	0.2859	-9.25	0.0000
Sexmale	-2.6190	0.2150	-12.18	0.0000
Age	-0.0449	0.0082	-5.46	0.0000
SibSp	-0.3786	0.1214	-3.12	0.0018

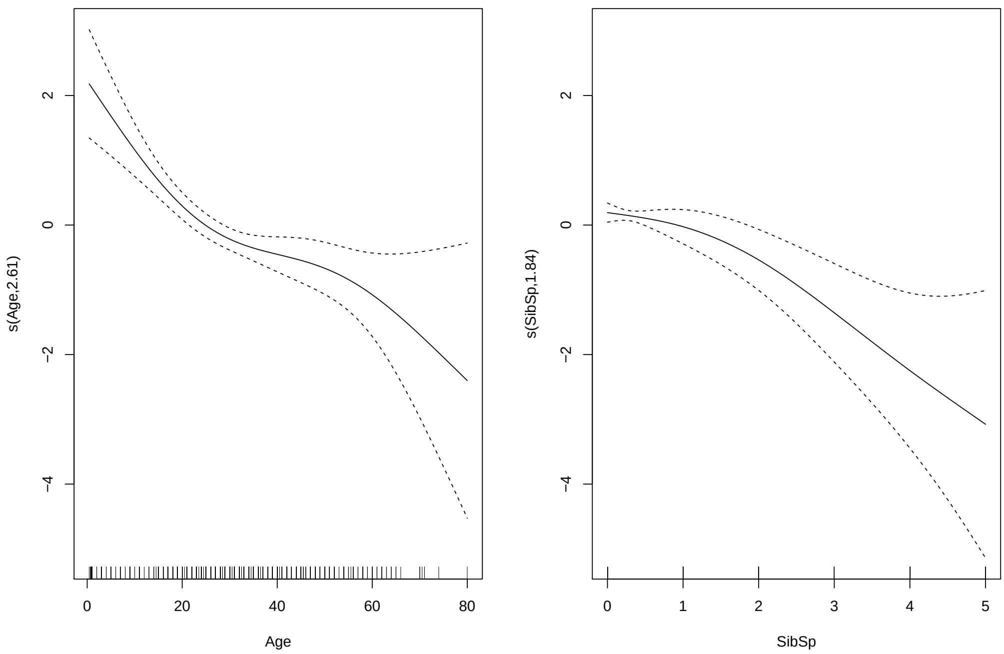
When one uses this model to predict on the same data (by predicting to the most probable class), we obtain an error rate of 19.38%. The log-likelihood value for this modellen is -318.0.

Explain why the log-likelihood value will be *smaller* in this case.

Argue why this model still is preferable.

(Continued on page 3.)

Another alternative is a generalised additive model (GAM). The plots below show the non-linear functions that were included in the model, based on the same explanatory variables as in exercise (c). The log-likelihood value for this model is -312.2 while the estimated degrees of freedom is 8.4.

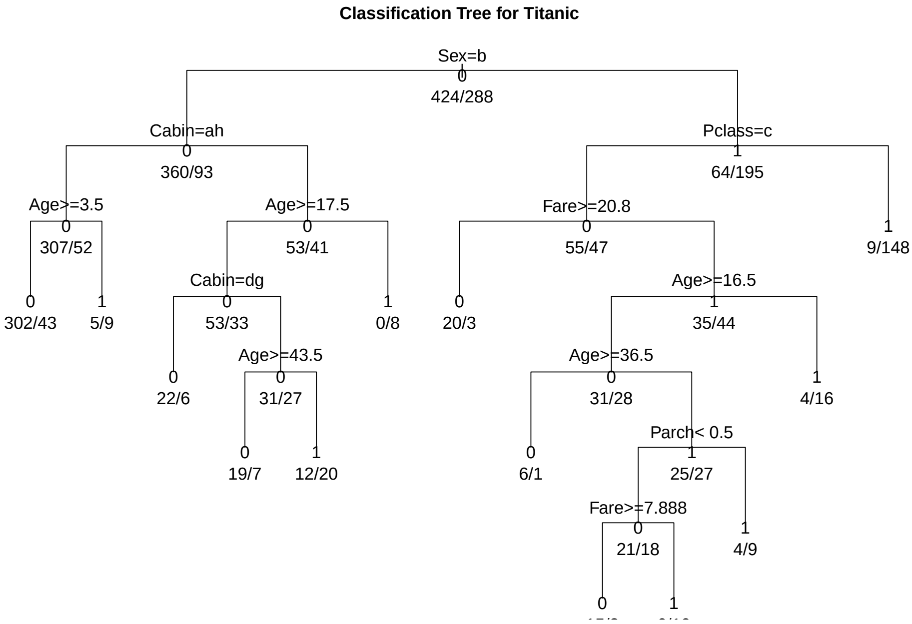


(d) Explain how the degrees of freedom is calculated in this case. Use this to compare this model with earlier models.

Comment on whether the plots shows significant non-linearities.

(e) Another alternative model can be obtained by classification trees. Below is a plot of a classification tree based on 11 end nodes.

(Continued on page 4.)



Discuss why classification trees give the posssibility of including *interactions* between explanatory variables.

Explain why a likelihood function for classification trees with a response within two classes can be written as

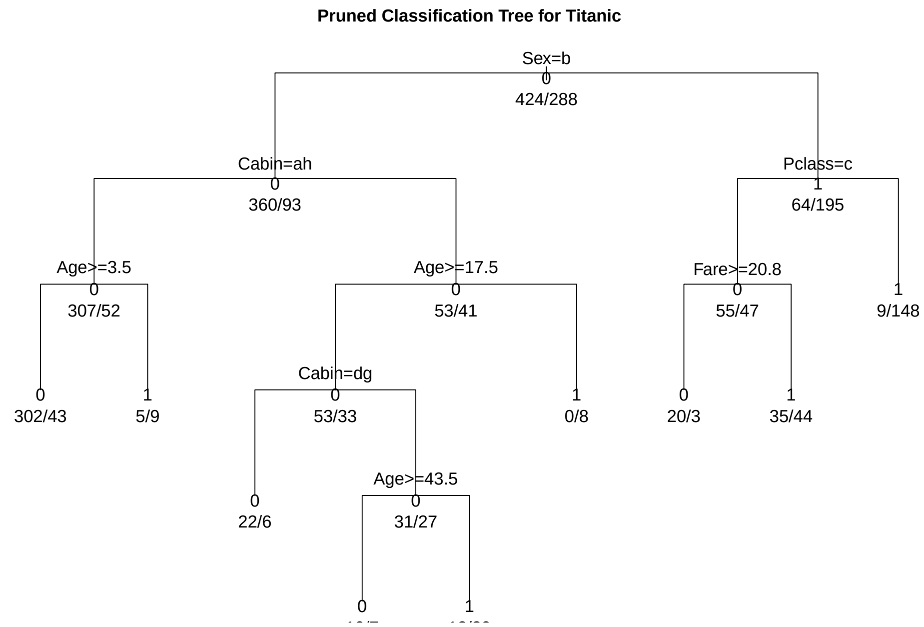
$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}$$

where $p_i = c_m$ for $\boldsymbol{x}_i \in R_m$.

(f) For the specific tree we obtained a log-likelihood value of -279.452. Use this to compare this model with earlier models.

(g) Discuss why it may be useful to *prune* trees. Below you see a tree pruned to include 9 end nodes. The log-likelihood value is in this case -287.349. Also evaluate this model compared with the previous ones.

(Continued on page 5.)



(h) Below is given a table of estimated error rates based on *cross-validation* (divided into 8 groups). Alternative methods such as Bagging, Random Forest and neural network are also included.

Method	Error rate (%)
Logistic regression, all variables	15.59
Logistic regression, variable selection	17.84
GAM, all variables	11.24
GAM, variable selection	16.85
Classification tree, 11 noder	20.37
Classification tree, 9 noder	19.94
Bagging	20.79
Random Forrest	19.38
Neural net (150 latent nodes)	20.37
Deep net (3 latent layers with 50 nodes in each)	22.75

Discuss the benefits in using cross-validation in evaluating different methods.

Give a short description on how Bagging, Random Forrest, neural nets and deep nets work.

(i) Discuss possible explanations on why the simple methods seems to work best in this case.

Assuming you choose the method with the smallest estimated error rate, discuss how you can say something about how good this selected method works. Discuss strengths and weaknesses of your choice.

(Continued on page 6.)

Problem 2

Assume a linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \quad i = 1, ..., n$$

where $\varepsilon_i \sim N(0, \sigma^2)$ and all noise terms are independent.

(a) Show that you can rewrite the model to

$$Y_i = \tilde{\beta}_0 + \beta_1 \tilde{x}_{i1} + \beta_2 \tilde{x}_{i2} + \varepsilon_i, \quad i = 1, ..., n$$

where $\sum_i \tilde{x}_{i1} = \sum_i \tilde{x}_{i2} = 0$. What kind of interpretation do $\tilde{\beta}_0$ have in this formulation of the model?

(b) Assume you want to estimate $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ by minimisation of

$$h(\beta_0, \beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2 + \lambda_1 \beta_1^2 + \lambda_2 \beta_2^2.$$

(We will in the following call the values that minimises h the *optimal* values).

Discuss situations were it can be useful to use $\lambda_1 \neq \lambda_2$.

Show that minimisation of $h(\boldsymbol{\beta})$ can be obtained by minimisation of

$$\tilde{h}(\tilde{\beta}_0, \beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - \tilde{\beta}_0 - \beta_1 \tilde{x}_{i1} - \beta_2 \tilde{x}_{i2})^2 + \lambda_1 \beta_1^2 + \lambda_2 \beta_2^2.$$

Find the optimal value of $\tilde{\beta}_0$.

(c) Put up an equation system which the optimal values of (β_1, β_2) has to satisfy.

Under the assumption that $\sum_i (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) = 0$, derive explicit expressions for the optimal values of (β_1, β_2). What will be the optimal value of β_0 in that case?

We will now look at the `Hitters` dataset where we want to predict `Salary` based on many different explanatory variables. We will however only look at two of these here: `PutOuts` and `Hits`. A simple linear regression based on these two explanatory variables gave the following results:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	535.9259	24.6013	21.78	0.0000
PutOuts	83.7694	25.8357	3.24	0.0013
Hits	172.7897	25.8357	6.69	0.0000

(Continued on page 7.)

In order to see the effect of penalty terms, three different trials were compared:

- $\lambda_1 = \lambda_2 = 0$.
- $\lambda_1 = \lambda_2 = \lambda$ where λ is specified by minimisation of the cross-validated estimate of the sum of squared errors.
- $\lambda_1 \neq \lambda_2$ where (λ_1, λ_2) are both specified by minimisation of the cross-validated estimate of the sum of squared errors.

The cross-validated estimates for the sum of squares errors were 63367, 163166 ($\lambda = 20.0$) and 163142 ($\lambda_1 = 20.0, \lambda_2 = 12.2$) respectively.

(d) Which methods do the first two trials correspond to?

Based on the results given, why is it reasonable that the optimal common λ value in the first trial corresponds to λ_1 in the third trial?

Discuss challenges in relation to using different penalty terms for the different explanatory variables when the number of explanatory variables increases.