NTNU

Norwegian University of
Science and Technology

Department of Mathematical Sciences

Examination paper for **TMA4268 Statistical learning**

Academic contact during examination: Mette Langaas

Phone: 988 47 649

Examination date: 24 May 2018

Examination time (from–to): 09:00–13:00

Permitted examination support material: C: One yellow A5 sheet with your own handwritten notes (stamped by the Department of Mathematical Sciences), specified calculator.

Other information:

- All answers must be justified, and relevant calculations provided.
- For each problem the maximum possible score is noted.
- Please use the Q1–Q28 in front of your answers.

Language: English

Number of pages: 10

Number of pages enclosed: 0

Checked by:

Informasjon om trykking av eksamensoppgave	
Originalen er:	
1-sidig ☐	**2-sidig** ☒
sort/hvit ☒	**farger** ☐
skal ha flervalgskjema ☐	

Date Signature

Problem 1 ***K*-nearest neighbour regression** [10 points]

We have a univariate continuous random variable Y and one covariate x. Further, we have observed a training set consisting of independent observation pairs $\{x_i, y_i\}$ for $i = 1, \ldots, 100$. A scatter plot of the training data is given in Figure 1.

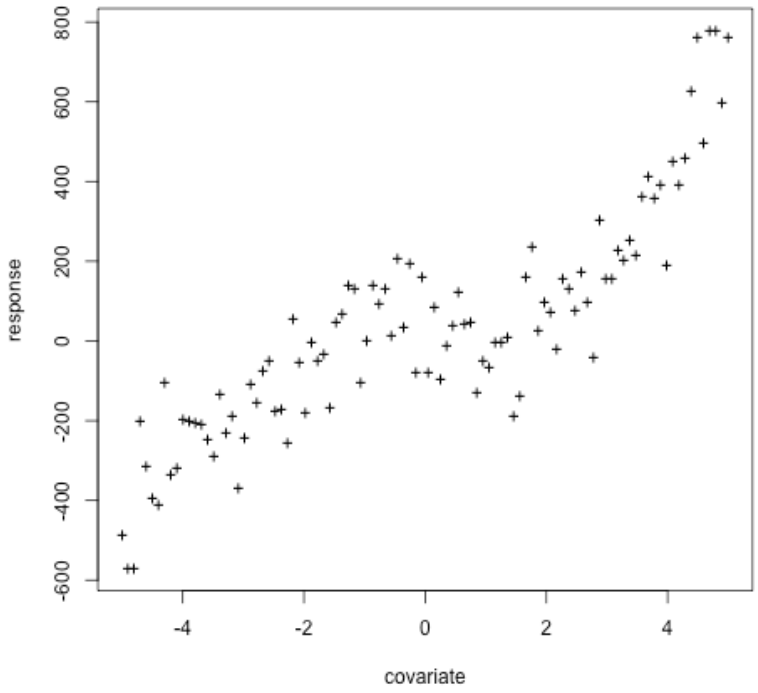

Figure 1: Training data. Observations shown as '+'.

Assume the following regression model

$$Y_i = f(x_i) + \varepsilon_i$$

where f is the true regression curve, and ε_i is an unobserved random variable with mean zero and constant variance σ^2 (not dependent on the covariate).

Our aim is now to produce an estimate of the true regression curve using K-nearest neighbour regression.

Q1: Write down the formula for the K-nearest neighbour regression curve estimate at a covariate value x_0, and explain your notation.

In Figure 2 we have used the K-nearest neighbour regression method, with K equal to 3, 15, 50 and 100, to estimate a regression curve based the training data. The four panels A–D correspond to K equal to 3, 15, 50 and 100, but not necessarily in that order.

Q2: Match panels A–D to values of K (3,15,50,100). Justify your choice.

To use this method the parameter K needs to be chosen. One possible method is 5-fold cross-validation.

Q3: Explain how 5-fold cross-validation is performed, and specify which error measure you would use. Your answer should include a formula to specify how the error is calculated during the cross-validation. A drawing is appreciated.

An alternative method to choose K is the leave-one-out cross-validation.

Q4: Would you prefer leave-one-out cross-validation to 5-fold cross-validation in this situation? Justify your answer.

Problem 2 An important decomposition in regression [10 points]

We have a univariate continuous random variable Y and a covariate x. Further, we have observed a training set of independent observation pairs $\{x_i, y_i\}$ for $i = 1, \ldots, n$. Assume a regression model

$$Y_i = f(x_i) + \varepsilon_i$$

where f is the true regression function, and ε_i is an unobserved random variable with mean zero and constant variance σ^2 (not dependent on the covariate). Using the training set we can find an estimate of the regression function f, and we denote this by $\hat{f}$. We want to use $\hat{f}$ to make a prediction for a new observation (not dependent on the observations in the training set) at a covariate value x_0. The predicted response value is then $\hat{f}(x_0)$. We are interested in the error associated with this prediction.

Q5: Write down the definition of the expected test mean squared error (MSE) at x_0.
Q6: Derive the decomposition of the expected test MSE into three terms.
Q7: Explain with words how we can interpret the three terms.

Assume that we have a method to estimate the regression function, where this method has a tuning parameter that controls the complexity of the model and that a large value of the tuning parameter gives high model complexity.

Q8: Make a sketch of how the expected test MSE (at x_0) could look as a function of the tuning parameter. Do the same for the three terms.

This decomposition has played a central role in our course.

Q9: In your opinion, what is the most important implication of this decomposition? Answer with *only one* sentence.

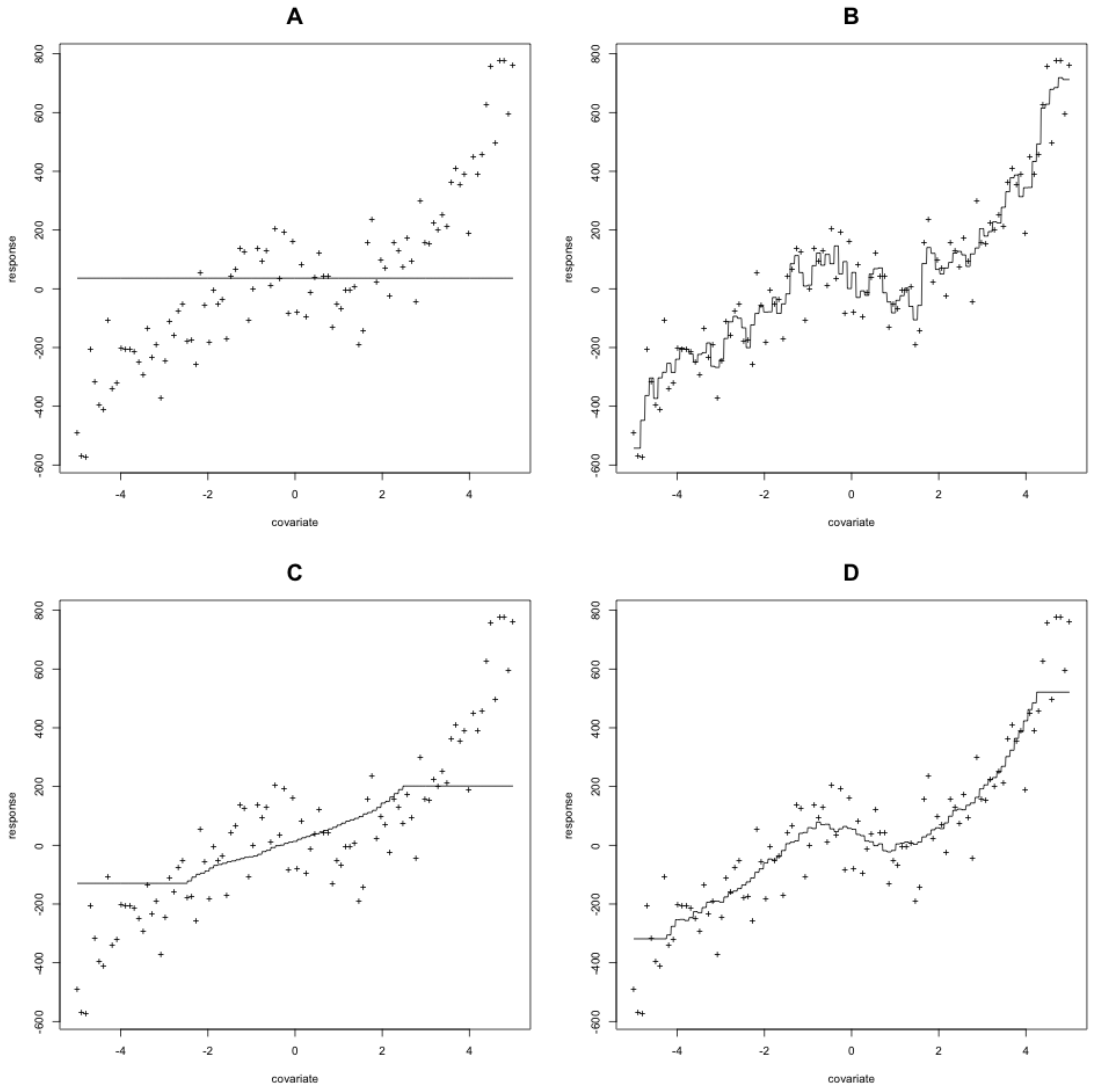

Figure 2: Observations shown as '+' and solid curves give the solution to a K-nearest neighbour regression with values of K equal to 3,15, 50 and 100 (in some order).

Problem 3 Linear discriminant analysis [10 points]

In this problem you may use that the probability density function for a p-dimensional multivariate normal random variable $\mathbf{X}$ with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$ is given as

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\det(\boldsymbol{\Sigma})^{1/2}} \exp\Big(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big).$$

We will not discuss how to estimate the mean and variance-covariance matrix here, so you may assume that they are known.

Q10: Write down the mathematical model assumptions for a linear discriminant analysis with two classes (coded as 0 and 1) and p explanatory variables and explain what the different ingredients are.
Q11: Explain how you derive the mathematical formula for the posterior probability for class 1.
Q12: Derive the mathematical formula for the class boundary between the two classes, given that the classification rule is to classify to the class with the highest posterior probability.
Q13: Is this boundary linear or non-linear in the space of the explanatory variables?

Problem 4 Classifying diabetes cases

We will look at data on diabetes (`diabetes` is '0' if not present and '1' if present) from a population of women of Pima Indian heritage in the US, available in the R `MASS` package. The following covariates were collected for each woman:

- `npreg`: number of pregnancies
- `glu`: plasma glucose concentration in an oral glucose tolerance test
- `bp`: diastolic blood pressure (mmHg)
- `skin`: triceps skin fold thickness (mm)
- `bmi`: body mass index (weight in kg/(height in m)^2)
- `ped`: diabetes pedigree function.
- `age`: age in years

We will use a training set (called `train`) with 200 observations (132 non-diabetes and 68 diabetes cases) and a test set (called `test`) with 332 observations (223 non-diabetes and 109 diabetes cases). Our aim is to make a classification rule for diabetes (or not) based on the available data.

In Figure 3 you find R-code and results from fitting a logistic regression to the `train` data set.

a) [10 points]

Q14: Write down the statistical model for the logistic regression.
Q15: Explain what is the estimated effect of the `ped` covariate on getting diabetes.
Q16: Would you predict that a person with the following characteristics has diabetes?

Person: `npreg`=2, `glu`=145, `bp`=85, `skin`=35, `bmi`=37, `ped`=0.7, `age`=40.

b) [10 points]

Q17: Draw a feedforward neural network with the same architecture as the logistic regression and specify which activation function(s) is/are used. Label nodes and connections with the same notation as for the mathematical model for the logistic regression.

The `test` data set was used to evaluate the performance of the logistic regression. A receiver–operator curve (ROC) was constructed and is the solid curve in Figure 4 (the dotted and dashed curves will be studie in Q24).

Q18: Explain briefly how a ROC curve is constructed. Your explanation should include the words cut-off, confusion matrix, sensitivity, specificity.

Look at the confusion matrix (probability cut-off 0.5) reported for the test set in the bottom part of the print-out in Figure 3.

Q19: Which point on the ROC curve does this cut-off correspond to?

```
# logistic regression
> fitlogist=glm(diabetes~npreg+glu+bp+skin+bmi+ped+age,data=train,
family=binomial(link="logit"))
> summary(fitlogist)
Call:
glm(formula = diabetes ~ npreg + glu + bp + skin + bmi + ped +
    age, family = binomial(link = "logit"), data = train)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.773062   1.770386  -5.520 3.38e-08 ***
npreg        0.103183   0.064694   1.595  0.11073
glu          0.032117   0.006787   4.732 2.22e-06 ***
bp          -0.004768   0.018541  -0.257  0.79707
skin        -0.001917   0.022500  -0.085  0.93211
bmi          0.083624   0.042827   1.953  0.05087 .
ped          1.820410   0.665514   2.735  0.00623 **
age          0.041184   0.022091   1.864  0.06228 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 256.41  on 199  degrees of freedom
Residual deviance: 178.39  on 192  degrees of freedom
AIC: 194.39
Number of Fisher Scoring iterations: 5

> predlogist=predict(fitlogist,newdata=test,type="response")
> testclasslogist=ifelse(predlogist > 0.5, 1, 0)
> table(test$diabetes, testclasslogist)
   testclasslogist
      0   1
  0 200  23
  1  43  66
>  # note: 223 true non-diabetes and 109 true diabetes cases
> library(pROC)
> roclogist=roc(test$diabetes,predlogist)
> auc(roclogist)
Area under the curve: 0.8659
> plot(roclogist, lty="solid")
```

Figure 3: R code and results from fitting logistic regression.

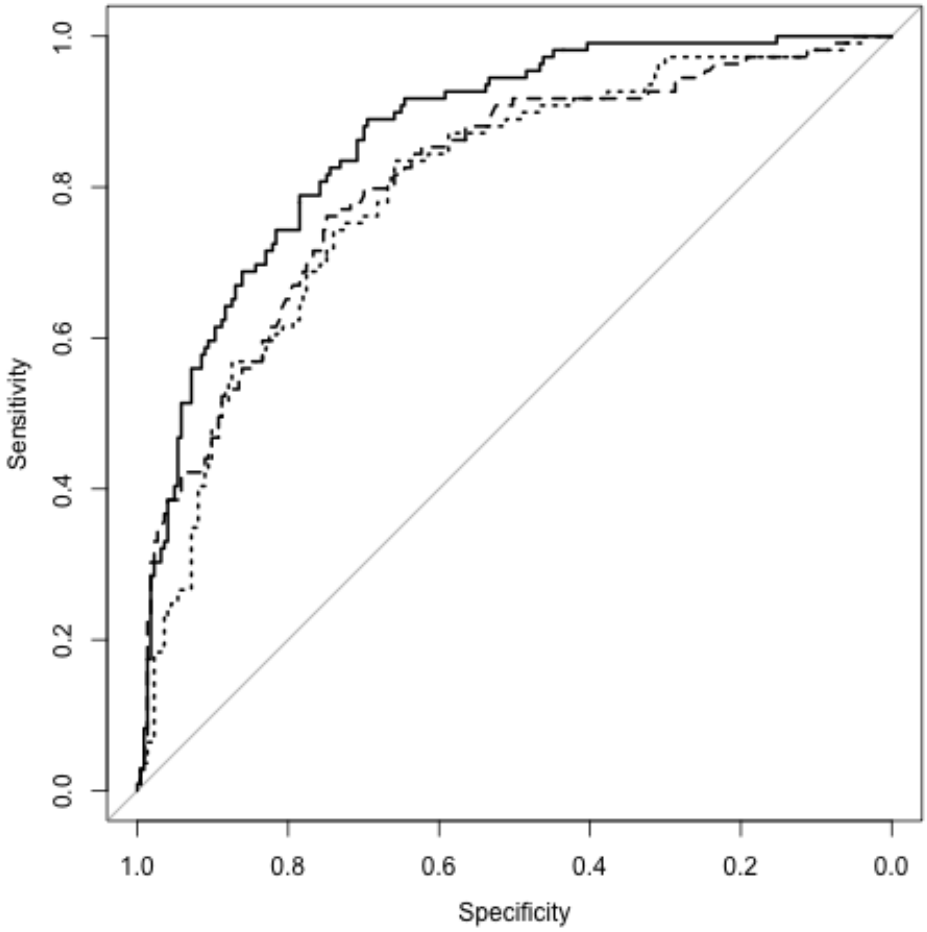

Figure 4: ROC-curves for logistic (solid), bagged trees (dashed) and quadratic discriminant analysis (dotted) for the diabetes test data set.

We have used bagging in combination with classification trees for the `train` data set, and tested on the `test` data set. R-code and results are found in Figure 5.

c) [10 points]

Q20: Explain how we build a bagged set of trees, and why we would want to fit more than one tree.

To estimate error rates (for bagged trees) it is possible to use something called an out-of-bag (OOB) sample.

Q21: Assume we have a data set of size n, calculate the probability that a given observation is in a given bootstrap sample.
Q22: What is an OOB sample?
Q23: Use the results in Figure 5 to compare the misclassification rates on the OOB sample to the misclassification rates for the `test` data set and comment on your findings.

We have now built classifiers for diabetes based on logistic regression and bagged classification trees. In addition the method quadratic discriminant analysis was also fitted (R-code and results in Figure 5). The results of all three methods are presented as ROC-curves in Figure 4. Here the solid line is the logistic regression, the dashed is the bagged trees and the dotted the quadratic discriminant analysis.

Q24: Which of the three methods would you recommend to use for predicting diabetes status?

```
> # bagged trees with call to randomForest
> library(randomForest)
> library(pROC)
> set.seed(4268)
> rf=randomForest(factor(diabetes)~npreg+glu+bp+skin+bmi+ped+age,
+                 data=train,
+                 mtry=7,ntree=1000,importance=TRUE)
> rf$confusion #error rates based on OOB data
    0  1 class.error
0 108 24   0.1818182
1  31 37   0.4558824
> yrf=predict(rf,newdata=test)
# note: test data 332 observations with
# 223 non-diabetes and 109 diabetes cases
> table(test$diabetes, yrf)
   yrf
      0   1
  0 184  39
  1  42  67
> predrf = predict(rf,test, type = "prob")
> rocrf=roc(test$diabetes, predrf[,2])
> auc(rocrf)
Area under the curve: 0.8094

> # fitting and ROC for QDA
> fitqda=qda(diabetes~npreg+glu+bp+skin+bmi+ped+age,data=train)
> predqda = predict(fitqda,newdata=test)
> testclassqda=ifelse(predqda$posterior[,2] > 0.5, 1, 0)
> table(test$diabetes, testclassqda)
   testclassqda
      0   1
  0 194  29
  1  47  62
> rocqda=roc(test$diabetes,predqda$posterior[,2])
> auc(rocqda)
Area under the curve: 0.7962

> # plotting all tree methods with ROC
> plot(roclogist) #logistic regression solid line
> plot(rocrf,add=TRUE,lty="dashed") #bagged trees dashed
> plot(rocqda,add=TRUE,lty="dotted") #QDA dotted line
```

Figure 5: R code and results from fitting bagged trees and quadratic discriminant analysis.

Problem 5 Clustering [10 points]

A data set has 5 observations (labelled a–e) of two variables (x_1, x_2) and is given in the left table of Figure 6. In addition, the Euclidean distance matrix between the observations is partially given in the right table.

Observations

label	x_1	x_2
a	1	1
b	2	1
c	4	5
d	7	7
e	5	7

Distance matrix

	a	b	c	d	e
a	0.0				
b	?	0.0			
c	5.0	4.5	0.0		
d	8.5	?	3.6	0.0	
e	7.2	6.7	2.2	?	0.0

Figure 6: Observations and Euclidean distance matrix for clustering.

Q25: There are three missing entries, labelled with a question mark, in the Euclidean distance matrix (right table of Figure 6). Calculate the missing entries.
Q26: Perform hierarchical clustering with complete linkage and draw the resulting dendrogram.
Q27: Using the dendrogram assume that we want two clusters, which observations are in each cluster?

A competing clustering method is k-means.

Q28: Discuss briefly *two differences* between hierarchical clustering and k-means clustering.