TMA4268 V2022 Exam without solutions

TMA4268 Statistical Learning V2022

Stefanie Muff, Department of Mathematical Sciences, NTNU

June 3, 2022

Warming up

Problem 1 (Fill-in-the-blank text, 5P)

Read the whole text and fill in the blanks such that the whole text makes sense (you might only understand which answer is correct after you continued reading):

In our course we were mainly discussing supervised statistical learning methods, broadly divided into regression and (prediction, inference, supervised,classification, unsupervised) problems. We were thereby discriminating between models that we use for prediction versus inference. In the latter case, we prefer to use (non-parametric, parametric, flexible, classification, least squares) models, for example (hierarchical clustering, linear regression, logistic regression, support vector machines, neural networks, KNN classification, K-means clustering, local regression) for a classification problem. However, when the aim is pure prediction, we are free to choose any model that is giving good predictions. (Non-parametric, regression, classification, clustering, supervised) methods are very flexible and thus more complex and (better interpretable, more appealing, better suited for inference, less interpretable) than linear models.

An important topic that we were discussing throughout the course was the bias-variance trade-off. The main idea is that the (training error, prediction bias, irrediculble error, test error, reducible error) is smallest for a model that balances bias and variance. With "variance' " we here mean (the residual variance, the variance of the covariates, the variance of the response, the reducible error, the variance of the estimated function).

In the context of the bias-variance trade-off we were also discussing variable selection and methods based on shrinkage. Lasso and ridge regression were two shrinkage methods we learned about. Both Lasso and ridge regression are (more flexible, more interpretable, less flexible, less biased, easier to use) compared to least squares. Lasso tends to lead to (more accurate models, models with lower prediction error, simpler models, better classification models, more flexible models) than ridge regression, thus Lasso is suitable for variable selection. In contrast to (linear regression, ridge regression, the boostrap, AIC minimization, boosted regression trees), for example, Lasso does not suffer from model selection bias and should thus be preferred over other methods when the aim is to select a subset of variables.

Problem 2 (Dropdown menu, 2P)

Choose from the dropdown menu below the correct optimization method that corresponds to the problem formulation:

a)

Maximise M by choosing $\beta_0, \beta_1, \ldots, \beta_p$ subject to $\sum_{j=1}^{p} \beta_j^2 = 1$, $y_i(\beta_0 + \beta_1 x_{i1} + \ldots \beta_p x_{ip} \geq M(1 - \epsilon_i) \quad \forall i = 1, \ldots, n$, $\epsilon_i > 0, \sum_i \epsilon_i \leq C$ for some $C > 0$. (Support Vector Classifier, logistic regression, regression tree, least squares regression, Lasso)

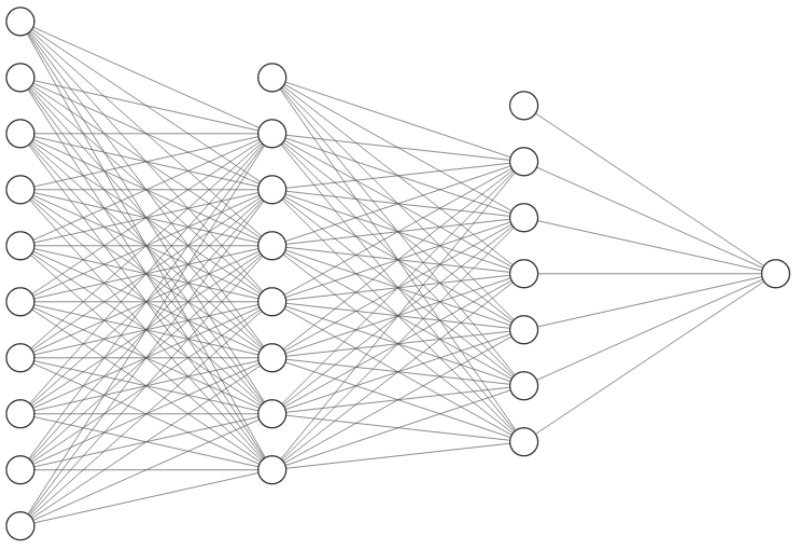
b)

$\text{argmax}_\beta \sum_{i=1}^{n}(y_i \log(p_i) + (1 - y_i)\log(1 - p_i))$, where $p_i = \frac{1}{1+\exp(-(\beta_0+\beta_{i1}+\ldots+\beta_p x_{ip}))}$. (Support Vector Classifier, logistic regression, classification tree, least squares regression, Lasso, feed-forward neural network)

Problem 3 (6P)

a) (4P)

Look at the following graphical description of a neural network:



(i) (2P) Write down the equation that describes how the input is related to output in this network, using ReLU activation functions in the two hidden layers, and sigmoid activation function in the output layer, assuming that there is one bias node in each of the layers (except the output layer).
(ii) (1P) What kind of task can this network be used for?
(iii) (1P) How many parameters are estimated in this network?

b) (2P)

We have four observations for which we know the distance matrix in Euclidean space:

$$\begin{bmatrix} 0 & 3 & 5 & 7 \\ 3 & 0 & 6 & 4 \\ 5 & 6 & 0 & 5.5 \\ 7 & 4 & 5.5 & 0 \end{bmatrix}.$$

Based on this dissimilarity matrix, sketch the dendogram that results from hierarchical clustering using complete linkage. On the plot, indicate the height where each fusion occurs, as well as the observastions that correspond to the leafs in the dendogram (enumerated as 1, 2, 3, 4).

Problem 4 – Data analysis 1 (17P)

In this data analysis problem we are using a dataset that is very similar to the bodyfat data we used in our course. The bodyfat dataset we use here was taken from kaggle https://www.kaggle.com/datasets/fedesoriano/body-fat-prediction-dataset . When clicking on the link you will also find a detailed description of the variables

Load the dataset and create a training and a test set using the following code:

```
id <- "1kGOLsnKAOUq2lWKlMjhAF8h71scOWcLO"  # google file ID
d.bodyfat <- read.csv(sprintf("https://docs.google.com/uc?id=%s&export=download",
    id))[, -c(1)]

set.seed(1234)
training_set_size <- floor(0.8 * nrow(d.bodyfat))

samples <- sample(1:nrow(d.bodyfat), training_set_size, replace = F)
d.body.train <- d.bodyfat[samples, ]
d.body.test <- d.bodyfat[-samples, ]
```

Before you get started, it is smart to make yourself familiar with the data. Use `str(d.body.train)`, `summary(d.body.train)` or other functions or graphical tools you learned about in the course.

a)

(i) (1P) Fit a linear regression model on the training data, with `BodyFat` as the response variable. Use all predictors linearly plus a quadratic term for `Abdomen` (but no interactions)

(ii) (1P) Report and interpret R^2, and relate it to the adjusted R^2. What does the difference tell you?

(iii) (1P) Do a residual analysis using the Tukey-Anscombe plot and the QQ-diagram. Explain which plot addresses which assumption(s) and whether the respective assumptions are met here.

b)

(i) (2P) To make the model efficient and easy to use in a practical application, we are interested in finding a model that has as few variables as possible and still gives good predictions. To this end, perform forward selection on the model you used in a) (including the quadratic term for `Abdomen`). Choose the model with the lowest BIC.

(ii) (1P) Fit the model selected in (i) to the training data and calculate the MSE on the test data.

c) Lasso

(i) (2P) Now use Lasso to do model selection. To this end, use the training data and choose the largest λ within 1 standard error from the lambda with the minimal error in a 5-fold cross-validation. As above, include the quadratic term for `Abdomen`. **Requirement**: Use `set.seed(4268)` before running the cross-validation.

(ii) (2P) Fit the Lasso-model with $\lambda_{1\text{se}}$ selected in (i) to the training data and calculate the MSE on the test data.

(iii) (2P) Compare the model and MSE you found in (ii) with a model where you use $\lambda_{\min}$ instead. Is the model with $\lambda_{\min}$ useful for tue purpose of task c)?

d)

Now we are looking at principal component analysis (PCA) and principal component regression (PCR).

(i) (2P) First, make a PCA for the training data, including all the regression variables (i.e., all variables except Bodyfat) plus the squared version of `Abdomen`. Display in a scree plot the variance explained by the PCs against the PC number. Looking at this plot, what could be a useful number of PCs to be included in a PCR?

(ii) (2P) Instead of directly using the number of PCs you found in (i), run a cross-validated PCR on the training data. Choose the number of PCs that lead to the smallest CV error. Then use the respective model fitted on the training data to calculate the test MSE.

(iii) (1P) Compare the number of chosen PCs from (i) and (ii) and interpret the difference.

R-Hints:

- To do the PCA with all the variables including `Abdomen^2`, you can add it directly to the training and test sets:

```
d.body.train$Abd2 <- (d.body.train$Abdomen)^2
d.body.test$Abd2 <- (d.body.test$Abdomen)^2
```

- use `scale=TRUE` in the PCA and PCR.
- Use `set.seed(4268)` before running the cross-validation in task (ii).

Problem 5 – Data analysis 2 (17P)

In this example with look at a dataset taken from https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease . The data stem from a large health survey in the US and contain information about factors that are related to heart disease status of individuals. When clicking on the link you will also find a detailed description of the variables.

The entire datasets contains almost 400'000 rows. For convenience, we are looking at a portion of the data with only 20'000 instances. Load the dataset and create a training and a test set using the following code:

```
id <- "1HM1ytt-x9QkTHQu7bMvhBJSJWihzpZJ2"  # google file ID
d.heart <- read.csv(sprintf("https://docs.google.com/uc?id=%s&export=download",
    id))
d.heart$HeartDisease <- as.factor(d.heart$HeartDisease)

# 70% of the sample size for training set
training_set_size <- floor(0.7 * nrow(d.heart))

set.seed(4268)
train_ind <- sample(seq_len(nrow(d.heart)), size = training_set_size)

train <- d.heart[train_ind, ]
test <- d.heart[-train_ind, ]
```

Before you get started, it is smart to make yourself familiar with the data. Use `str(train)`, `summary(train)` or other functions or graphical tools you learned about in the course.

a) (3P)

(i) (2P) Fit a logistic regression model using the training data with `HeartDisease` as the response and `BMI`, `Smoking`, `AlcoholDrinking`, `Sex` and `AgeCategory` as covariates, including interactions between `Smoking` and `Sex` and between `AlcoholDrinking` and `Sex`. Print the `summary()` table.

(ii) (1P) How many age categories are in this dataset? How many regression parameters are estimated for `AgeCategory`?

b) (4P)

(i) (2P) Use the fitted model from a) to write down the estimated regression model of `HeartDisease` vs. `BMI` for a non-smoking male age 20 that does not drink alcohol.

(ii) (1P) Is there evidence that the effect of drinking alcohol differs between males and females? Give a reason.

(iii) (1P) Do you think the purpose of this model was inference or prediction? Give a reason.

c) (6P)

From now one we are using *all* the covariates in the dataset (but no interaction terms).

(i) (1P) Carry out a linear and a quadratic discriminant analysis (LDA, QDA) on the training data, again using `HeartDisease` as the response.
(ii) (2P) Using the fitted models from (i), calculate the test error for both the LDA and QDA.
(iii) (2P) Calculate the AUC using the test data for both models (still using the fitted model from (i)).
(iv) (1P) Why would k-nearest-neighbor (KNN) classification probably not work very well for this task?

d) (4P)

(i) (3P) The aim of this task is to find a tree-based method that gives a lower test error than both LDA and QDA above. To this end, choose a tree-based method and fit it using the training data, then predict the response using the test data. Justify the choice of any parameters you use.
(ii) (1P) Based on the model you chose in (i), which three variables are most important to predict heart disease, according to an importance measure based on node purity?

Multiple and single choice questions

Problem 6 (6P, single choice, 2P each)

a)

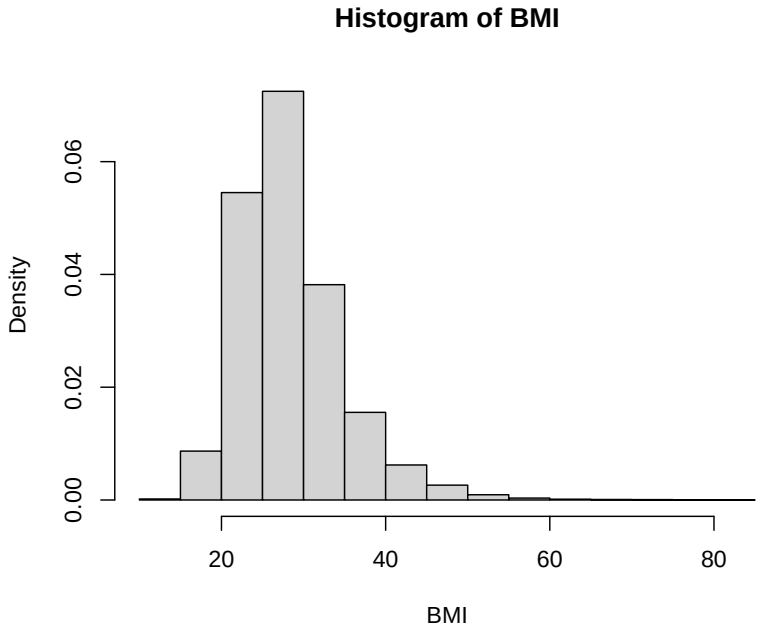
We are referring to the logistic regression model you fitted in Problem 5a). What is the probability that a smoking and alcohol drinking male age 77 with a BMI of 25 suffers from heart disease?

(i) 0.127
(ii) 0.003
(iii) 0.299
(iv) 0.077
(v) 0.23
(vi) 0.007

b)

Below you find the histogram for the distribution of BMI in our heart disease dataset from Problem 5.

```
d.bmi <- d.heart$BMI
hist(d.bmi, main = "Histogram of BMI", xlab = "BMI", freq = F)
```



You can see that the distribution is a bit skewed, which implies that the estimated mean $\hat{\mu}$ and median $\hat{\mu}_{\text{med}}$ are unequal. In fact, the *difference* between the mean and the median is 1.005, calculated as

```
mean(d.bmi) - median(d.bmi)
```

Run a bootstrap and find the (approximately) correct 95% confidence interval for the difference betwen the mean and the median. Round to three digits after the comma. Use at least 1000 bootstrap samples. Your answer might differ by a little bit due to sampling error – choose the closest anwer:

(i) 0.945, 1.066
(ii) -2.836, 4.847
(iii) 0.974, 1.036
(iv) 1.004, 1.007
(v) -0.955, 2.965

c)

$\mathbf{x} = [x_1, x_2, x_3]^T$ is a 3-dimensional random vector with covariance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} 5 & 2 & 4 \\ 2 & 7 & -3 \\ 4 & -3 & 8 \end{bmatrix}$$

The correlation between element x_2 and x_3 of the vector $\mathbf{x}$ is:

(i) -3
(ii) -0.40
(iii) 3
(iv) -0.23
(v) -0.23
(vi) 0.40
(vii) It is not possible to calculate the correlation, because the matrix is not positive definite.

Problem 7 (6P, multiple choice, 2P each)

a)

Which of the following statements about clustering methods are true, which false?

(i) The results of hierarchical clustering may differ depending on which datapoints are grouped first.
(ii) The results in K-means clustering may differ for different initial cluster assignments.
(iii) Hierarchical clustering works well when the underlying data have a hierarchical structure.
(iv) A drawback of hierarchical clustering is that we have to decide the number of clusters in advance.

b)

Which of the following statements are true, which false?

(i) A natural cubic spline is linear beyond the boundary knots.
(ii) A regression spline of order 3 with 5 knots has 9 basis functions.
(iii) A regression spline with polynomials of degree M-1 has continous derivatives up to order M-2, but not at the knots.
(iv) Regression splines generally produce more stable estimates than polynomial regression, and are therefore preferable.

c)

When talking about support vector classifiers in $p = 2$ dimensions, we were looking at hyperplanes of the form $\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$. We are now looking at the following non-linear decision boundary

$$(1 + X_1)^2 + (X_2 + 2)^3 - X_2^3 = 2 \ .$$

We assume that class 1 fulfils $(1+X_1)^2+(X_2+2)^3-X_2^3 > 2$ and class 2 fulfils $(1+X_1)^2+(X_2+2)^3-X_2^3 < 2$.

Which of the following statements are true?

(i) This decision boundary is linear in terms of X_1, X_2, X_1^2 and X_2^2.
(ii) The decision boundary has the shape of a circle.
(iii) The point $(x_1, x_2) = (1, -1)$ belongs to class 1.
(iv) The point $(x_1, x_2) = (1, 1)$ belongs to class 2.