TMA4268 V2021 Exam

TMA4268 Statistical Learning V2021

Stefanie Muff, Department of Mathematical Sciences, NTNU

May 26, 2021

Warming up

Problem 1 (Fill-in-the-blank text, 7P)

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):

We have discussed a lot of methods and models, and in (*supervised*, *unsupervised*, *parametric*, *non-parametric*) methods there were two main purposes: (*prediction*, *inference*, *bias reduction*, *variance reduction*, *supervised learning*, *supervised learning*) and (*prediction*, *bias reduction*, *inference*, *variance reduction*, *unsupervised learning*, *supervised learning*). In both cases we want to learn from data and build a model that relates a set of variables to an outcome, but in the first case we do not care about the actual model parameters, because we do not want to interpret them. Some of the methods we learned about were (*non-parameteric*, *parametric*, *supervised*, *unsupervised*) and others were (*non-parametric*, *parameteric*, *supervised*, *unsupervised*), whereas the former tend to be more more rigid and thus less flexible – and thus possibly more biased – than the latter ones.

In any case, the aim of a supervised model fitting procedure in statistical learning is to minimize the (alternatives: *irreducible error*, *bias*, *variance*, *overfitting*, *underfitting*, *reducible error*) of the estimated relation between some predictor variables and a response. The (alternatives: *irreducible error*, *reducible error*, *bias*, *variance*, *overfitting*, *underfitting*), on the other hand, corresponds to the expected test error when we find the best possible function that relates the predictors to the response.

Conceptual / theoretical questions

Problem 2 (Multiple choice, 6P)

Note: There was some unclarity in 2a), wheter Ridge regression also counts as a method to do model selection (see lecture notes for module 12). Therefore, Ridge regression was counting as correct, no matter whether it was selected or not. From the total 6 points of this question, -1 point was then deducted either when one correct answer was missing (except Ridge regression, here both were ok), or when a wrong answer was clicked (except Ridge regression again).

a) (2P)

Which of the following methods is/are suitable to do model selection / variable selection?

- Support vector classifiers
- Lasso
- Ridge regression
- Partial least squares regression

b) (2P)

Which of the following are assumptions that are made when fitting a normal multiple linear regression model $y_i = \beta_0 + \beta_x x_i + \beta_z z_i + \epsilon_i$ with $1 \leq i \leq n$ and n number of data points?

- The response variables y_i are independent and identically distributed.
- The covariates x and z are not collinear.
- The error terms are normally distributed.
- The error terms have a constant variance and are independent of each other.

c) (2P)

Which of the following are tuning parameters in at least one of the methods we discussed in the course?

- The number of trees in boosted regression trees.
- The width of the margin M in support vector classifiers.
- The number of principal component included in principal component regression.
- K in k-nearest-neighbor (KNN) classification.

Problem 3 (Free text questions 7P)

a) (3P)

One of the central topics in the course was the bias-variance trade-off.

(i) (1P) Say in 2-3 sentences what the bias variance trade-off means.
(ii) (2P) Give two examples of methods in the course where there were (tuning) parameters that could be tweaked such that there was a bias variance trade-off.

b) (2P)

Assume you are a statistician that is involved in an epidemiological study. The researchers are interested in finding variables that are associated with obesity, such as physiological measures, personal behaviour, food habits or genetic components. The aim of the study is to help obese people lose weight.

(i) (1P) What is the purpose of finding a good model for this question: prediction or inference? Please justify your anwer.
(ii) (1P) When would analysing the same dataset serve the opposite purpose (i.e., if you chose prediction in (i), which question would make it an inference type of problem, and vice versa)?

c) (2P)

We have learned about K-means clustering and hierarchical clustering. Give one advantage and one disadvantage of K-means clustering compared to hierarchical clustering.

Problem 4 (5P)

In the module about support vector machines we have mentioned the *hinge loss* $L(\boldsymbol{x}_i, y_i, \boldsymbol{\beta}) = \max(0, 1 - y_i f(\boldsymbol{x}_i))$ where in the simplest case, $f(\boldsymbol{x}_i)$ is a linear function of the covariates and some parameter vector $\boldsymbol{\beta}$. In the case where $f(\boldsymbol{x}_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$ we saw that the hinge loss and the logistic regression loss were very similar, see Figure 9.12 of the course book:

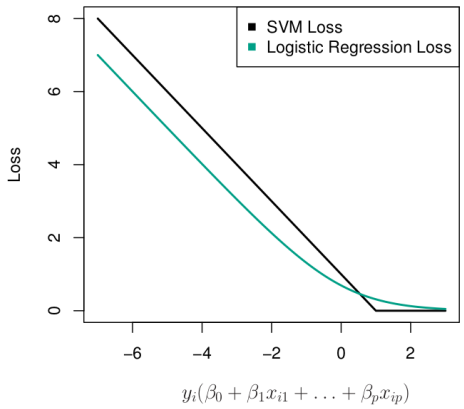

Figure 1: ISLR Figure 9.12: hinge loss - loss 0 for observations on the correct side of the margin

Show that the loss function

$$\log(1 + \exp(-y_i f(\boldsymbol{x}_i)))$$

is the negative log-likelihood for the $y_i = -1, 1$ encoding in a logistic regression model.

Problem 5 – Data analysis 1 (20P)

In this task we are again using the bodyfat example from the course, but this time we use a version with more covariates. The data set can be loaded and split into a training and test set as described below. Please look at the data set yourself, for example by using the `pairs()` and `str()` functions, before you start working on the analysis.

```
id <- "1dNLfx9Dbs2gYIooUxA6HMxK_MPFwE3Hn"  # google file ID
d.bodyfat <- read.csv(sprintf("https://docs.google.com/uc?id=%s&export=download",
    id), header = T)
```

The variable `bodyfat` is the reponse variable, all other variables are covariates.

We are splitting into a training set (180 observations) and a test set (63 observations):

```
set.seed(1234)

# Some students scaled the data: d.bodyfat <-
# data.frame(scale(d.bodyfat))

samples <- sample(1:234, 180, replace = F)
d.body.train <- d.bodyfat[samples, ]
d.body.test <- d.bodyfat[-samples, ]
```

a) (4P)

Carry out Lasso regression on the training set, and say how you choose λ. Report the MSE on the test data.

b) (4P)

(i) (2P) Fit a linear regression model with all covariates to the training data and calculate the MSE on the test data.
(ii) (1P) Compare the test MSE to the one you obtain from a) and interpret the difference.
(iii) (1P) Systematically compare the regression coefficients between linear regression and Lasso, and give a theoretical explanation for the pattern you see.

c) (4P)

Fit a GAM on the training data, including

- a polynomial of degree 2 for age,
- a natural cubic spline with 3 degrees of freedom for height,
- a natural cubic spline for abdomen with three knots at the 25%, 50% and 75% quantiles, respectively,
- a smoothing spline for hip,
- a linear term for weight and bmi.

Calculate the MSE for the test set.

d) (3P)

In the course you also heard about *partial least squares (PLS)* regression, which is a smart approach that uses the principal component regression idea, but finds the components that are most correlated with the response. For the bodyfat example do the following:

(i) (1P) Run a PLS regression on the training data (don't forget to scale the variables, `scale=TRUE`).
(ii) (1P) Choose the smallest number of components such that at least 95% of the covariate variance in the training data is explained.
(iii) (1P) Report the MSE of the test data when using the respective number of components.

e) (3P)

Finally, fit either a random forest or a boosted regression tree to the training data, and calculate the MSE on the test data. Explain your choic(es) that you make for the tuning parameter(s) in your selected method.

f) (2P)

Compare all the MSEs you found in a) to e) for the test set. Which of these methods seems to do best and worst for the given test data?

Problem 6 – Data analysis 2 (15P)

In this data analysis problem we are using data collected by people at the Centre for Biodiversity Dynamics at NTNU. The main study question was whether inbreeding is influencing the probability that young sparrows survive until the second year (denoted as recruitement). The list of covariates is as follows:

- `sex`: Sex of the animal.
- `lnrhday`: Day in the year when the bird hatched (enumerated from 1 to 365)
- `clsize`: The clutch size of the clutch the bird was born in
- `hyear`: Hatch year
- `f`: Inbreeding coefficient
- `hisl`: Hatch island (this is a categorical variable!)
- `H1`: Proportion of heterozygous loci
- `GTloci`: number of genotyped microsatellite loci
- `Hloci`: number of heterozygous loci
- `geno`: A variable where we do not know what it means
- `recruit`: The binary response variable for survival to the second year

The data can be loaded as follows:

```
id <- "1cSVIJv-OoAwkhUAuun2qQyOfiuZzkmo3"  # google file ID
d.sparrows <- read.csv(sprintf("https://docs.google.com/uc?id=%s&export=download",
    id), header = T)
```

It is advisable to look at the dataset, for example by using `pairs(d.sparrows)` and `str(d.sparrows)` before you do the following analyses:

a) (4P)

(i) (2P) Fit a logistic regression model on the full data set with `recruit` as response variable, using all the covariates plus an interaction term between `sex` and `f`. Remember that hatch island (`hisl`) is a categorical variable.

(ii) (1P) Fit a second model, but this time without hatch island as covariate. Use the anova function to compare to the model in (i).

(iii) (1P) Is there evidence that survival probabilities differed between hatch islands? Explain your response.

R-hints

- To compare two models using an anova table, use `anova(model1,model2, test="Chisq")`.

b) (3P)

Now split the dataset into a training and a test sample for prediction (assuming our aim is to predict survival). Split the dataset as in the code below, using the same seed. Then

(i) (1P) Fit logistic regression for the training data, but without hatch island and without interaction between f and sex.

(ii) (1P) Use the fitted model to predict survival in the test set using a probability cutoff of $p = 0.5$.

(iii) (1P) Generate the confusion table and calculate sensitivity and specificity for the prediction on the test set.

```
set.seed(123456)
samples <- sample(1:169, 120, replace = F)
d.sparrows.train <- d.sparrows[samples, ]
d.sparrows.test <- d.sparrows[-samples, ]
```

c) (3P)

Repeat the task from b), but now use a quadratic discriminant analysis (QDA) instead of logistic regression. Calculate sensitivity and specificity for QDA. Calculate sensitivity and specificity for the predictions on the test set.

d) (5P)

Finally we are using a neural network approach for our classification task. Prepare the data as indicated in the R-hints

(i) Fit a neural network with two hidden layers to the training set, where

- the first hidden layer has 32 units and the second hidden layer 64.
- the hidden layers have ReLU and the output layer a sigmoid activation function.
- you add 20% dropout in both hidden layers.
- you use RMSprop optimization.
- you use a batch size of 16.
- you use a validation split of 0.5 (**R-hint:** fit(..., validation_split=0.5)).
- you train the model for 25 epochs.
- you use `set.seed(1234)` before you train the network.

(ii) Calculate sensitivity and specificity for the predictions (probability cut-off 0.5) for the test set, and compare to the values you got from logistic regression and QDA above.

R-hints

```
library(keras)
library(caret)
x_train <- d.sparrows.train[, -c(6, 11)]
x_test = d.sparrows.test[, -c(6, 11)]

mean = apply(x_train, 2, mean)
std = apply(x_train, 2, sd)
x_train = scale(x_train, center = mean, scale = std)
x_test = scale(x_test, center = mean, scale = std)

y_train = as.numeric(d.sparrows.train$recruit)
y_test = as.numeric(d.sparrows.test$recruit)
```

To receive predictions from the neural network output, use

```
predictionsNN <- model %>% predict_classes(x_test)
```

Multiple and single choice questions

Problem 7 (4P, single choice, 1P each)

a)

Look at the estimated coefficients from regression model $y_i = \beta_0 + \beta_x x_i + \beta_z z_i + \beta_w w_i + \epsilon_i$ with continuous covariate x and binary covariates z and w.

```
##
## Call:
## lm(formula = y ~ x + z + w)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6167 -3.6347  0.5217  3.0458 12.0439
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.6311     2.4346   0.670   0.5046
## x             0.9826     0.1091   9.008 3.05e-14 ***
## z            -2.4697     1.0861  -2.274   0.0253 *
## w             2.4402     1.4355   1.700   0.0926 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.222 on 91 degrees of freedom
## Multiple R-squared:  0.491,  Adjusted R-squared:  0.4742
## F-statistic: 29.26 on 3 and 91 DF,  p-value: 2.463e-13
```

What is the predicted outcome ($\hat{y}_i$) for an individual i with $x_i = 20$, $z_i = 0$ and $w_i = 1$?

- 17.2
- 3.4
- 19.6
- 18.8
- 23.7

b)

Using again the same example as in a). Which of the following statements is true?

- The 95% confidence interval for $\hat{\beta}_x$ ranges from 0.87 to 1.20.
- The 95% confidence interval for $\hat{\beta}_z$ ranges from -3.56 to -1.38.
- The R-squared is the proportion of residual variability on the total variability of the response variable.
- The total number of data points in the above analysis was 94.
- None of the other statements are correct.

c)

We have 9 covariates, X_1 to X_9 , each of them uniformly distributed in the interval $[0, 1]$. To predict a new test observation $(X_1, \ldots, X_9)$ in a K-nearest neighbor (KNN) approach, we use all observations within 15% of the range closest to each of the covariates (that is, in each dimension). Which proportion of available (training) observations can you expect to use for prediction?

- $3.8 \cdot 10^{-8}$
- $1.5 \cdot 10^{-9}$
- $0.15 \cdot 10^{-9}$
- 0.15
- 10^{-9}

d)

Let us look at a neural network with one single output node. By which of the following users can this network potentially be used for the specified task?

- By the Norwegian post to read zip codes on letters.
- By a medical institution to discriminate gene expression patterns into three different disease statuses.
- By the criminal police to assign handwriting to different persons.
- By a hospital to classify X-ray pictures into healthy and unhealthy.
- Non of the other alternatives.

Problem 8 (6P, multiple choice, 2P each)

a)

In a study it was investigated how yearly income (in 1000 Eur) of adults varies with age (years) and education status (low, medium or high education). The result from the analysis is given in the following tables:

```
##                       Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)         33.3820487 3.46209745  9.6421459 2.404054e-20
## age                  1.0217459 0.07639833 13.3739291 2.977750e-35
## educationhigh        7.9713187 4.77037713  1.6710039 9.532491e-02
## educationmedium      3.6151331 4.70229197  0.7688023 4.423610e-01
## age:educationhigh    0.3496907 0.10748204  3.2534806 1.214376e-03
## age:educationmedium  0.1468468 0.10516250  1.3963800 1.631981e-01
```

```
## Analysis of Variance Table
##
## Response: income
##                Df Sum Sq Mean Sq  F value    Pr(>F)
## age             1 108000  108000 712.5665 < 2.2e-16 ***
## education       2  43378   21689 143.0993 < 2.2e-16 ***
## age:education   2   1621     810   5.3474  0.005027 **
## Residuals     518  78510     152
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Please select all statements that are true, according to the summary and anova output:

(i) People with high education tend to earn more than people with medium or low education.
(ii) Someone with a medium education earns, on average, EUR 102927 at age 50.
(iii) Income seems to increase with age, and the rate of the increase depends on the education.
(iv) Education does not seem to play a role for people at higher ages.

b)

Which of the following statements are true, which false?

(i) The maximal margin hyperplane approach is equivalent to a linear discriminant analysis when the covariates are normally distributed with identical covariance matrix in the different groups.
(ii) The support vector classifier is equivalent to quadratic discriminant analysis when the covariates are normally distributed with group-specific covariance matrices.
(iii) Logistic regression, LDA and support vector machines tend to perform similar when decision boundaries are linear, unless classes are linearly separable.
(iv) An advantage of logistic regression over SVMs is that it is easier to do feature selection and to interpret the results.

c)

We are looking at the `mtcars` dataset given in R. This dataset consists of data on 32 models of cars, taken from an American motoring magazine (1974 Motor Trend magazine). For each car, you have 11 features, expressed in varying units (US units). You can check in R via ?mtcars to see what the different variables mean.

Here we carried out a principal component analysis and give the biplot and the scree plot below. In the biplot we also color the cars according to their origin. Which of the following statements are correct?

(i) The first two components explain 84.2% of the variability of the response variable.
(ii) Japanese cars tend to have lower miles per gallon (mpg) and more cylinders (cyl variable) than American cars, which indicates that Japanese cars are more environment friendly.
(iii) mpg has a very small loading for the second principal component.
(iv) The first three PCs explain most of the variability in the data.