# Examination

**Linköping University, Department of Computer and Information Science, Statistics**

| | |
|---|---|
| Course code and name | 732A99 Machine Learning |
| Date and time | 2020-01-16, 14.00-19.00 |
| Assisting teacher | Oleg Sysoev |
| Allowed aids | "Pattern recognition and Machine Learning" by Bishop and "The Elements of Statistical learning" by Hastie |
| Grades: | A=19-20 points<br>B=16-18 points<br>C=11-15 points<br>D=9-10 points<br>E=7-8 points<br>F=0-6 points |

**Provide a detailed report that includes plots, conclusions and interpretations. Give motivated answers to the questions. If an answer is not motivated, the points are reduced. Provide all necessary codes in the appendix.**

**Use seed 12345 when randomness is present unless specified otherwise.**

## Assignment 1 (3p)

The data file **glass.csv** contains information about the chemical components of two different glass types. The Type of glass is represented by variable Class (0/1), data also contains identification number ID. Import data to R and divide it into training, validation and test sets (40/30/30) by using this kind of code:

```
n=dim(data)[1]
set.seed(12345)
id=sample(1:n, floor(n*percent1))
train=data[id,]
id1=setdiff(1:n, id)
set.seed(12345)
id2=sample(id1, floor(n*percent2))
```

```
valid=data[id2,]
id3=setdiff(id1,id2)
test=data[id3,]
```

1. Use combined training and validation data to train a logistic regression model in which all chemical components are features and Class variable is the target. Report confusion matrix for the test data and comment on the quality of prediction. Report the fitted probabilistic model and the equation of the decision boundary.

## Assignment 2 (4p)

A total of 1,203,646 fruit flies were studied and the number of flies found dead each day was recorded. The data set **mortality_rate.csv** contains the mortality rate (*Rate*) of the flies for each day (*Day*)

1. Use basis function approach to fit an order-5 spline with a single knot $\zeta = 75$ such that Day is the feature and Rate is the target variable. Use only basic R functions in your implementation, and you may also use function lm(). Present the original and predicted data in the same plot and comment on the quality of fit. Comment whether the third- and fourth-degree terms in the model were necessary and report the degrees of freedom of the model. Finally, answer whether the model is parametric or not (provide a well-motivated answer!)

## Assignment 3 (3p)

Data file **geneexp.csv** contains information about gene expression of three different cell types (column Cell Type).

1. Fit a Nearest Shrunken Centroid model to these data in such a way that the data are not scaled and the threshold value is selected by the cross-validation. Present the optimal value of the threshold and the centroid plot for the optimal model. Which 5 genes are most important according to the plot and how do you decide on that? What meaning do positive and negative values have in the centroid plot? Can it happen that all values in the centroid plot are positive for some gene? Report the total number of genes selected by the model.

## Assignment 4 (10p)

MIXTURE MODELS - 7 POINTS

You are asked to implement the EM algorithm for mixtures of multivariate Gaussian distri- butions. You should use the following equations in the E-step

.

$$p(z_{nk}|\boldsymbol{x}_n, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \frac{\pi_k f(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_k \pi_k f(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}$$

and in the M-step

$$\begin{aligned}
\pi_k^{ML} &= \frac{\sum_n p(z_{nk}|\boldsymbol{x}_n, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})}{N} \\
\boldsymbol{\mu}_k^{ML} &= \frac{\sum_n \boldsymbol{x}_n p(z_{nk}|\boldsymbol{x}_n, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})}{\sum_n p(z_{nk}|\boldsymbol{x}_n, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})} \\
\boldsymbol{\Sigma}_k^{ML} &= \frac{\sum_n (\boldsymbol{x}_n - \boldsymbol{\mu}_k^{ML})(\boldsymbol{x}_n - \boldsymbol{\mu}_k^{ML})^T p(z_{nk}|\boldsymbol{x}_n, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})}{\sum_n p(z_{nk}|\boldsymbol{x}_n, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})}
\end{aligned}$$

where $f$ is the density function of a Gaussian distribution, which is implemented by the function `dmvnorm` in the R package `mvtnorm`. You may want to reuse your solution for the lab on mixture models. Use the following code for sampling the learning data and initializing the parameters. Note that the learning data consists of 300 points sampled from a mixture model with three equally likely components, and each component is a bivariate Gaussian distribution.

```
library(mvtnorm)
set.seed(1234567890)
max_it <- 100 # max number of EM iterations
min_change <- 0.1 # min change in log likelihood between two consecutive EM iterations
N=300 # number of training points
D=2 # number of dimensions
x <- matrix(nrow=N, ncol=D) # training data
# Sampling the training data
mu1<-c(0,0) # component 1
Sigma1 <- matrix(c(5,3,3,5),D,D)
dat1<-rmvnorm(n = 100, mu1, Sigma1)
mu2<-c(5,7) # component 2
Sigma2 <- matrix(c(5,-3,-3,5),D,D)
dat2<-rmvnorm(n = 100, mu2, Sigma2)
mu3<-c(8,3) # component 3
Sigma3 <- matrix(c(3,2,2,3),D,D)
dat3<-rmvnorm(n = 100, mu3, Sigma3)
plot(dat1,xlim=c(-10,15),ylim=c(-10,15))
points(dat2,col="red")
points(dat3,col="blue")
x[1:100,]<-dat1
x[101:200,]<-dat2
x[201:300,]<-dat3
plot(x,xlim=c(-10,15),ylim=c(-10,15))
K=3 # number of guessed components

z <- matrix(nrow=N, ncol=K) # fractional component assignments

pi <- vector(length = K) # mixing coefficients
mu <- matrix(nrow=K, ncol=D) # conditional means
Sigma <- array(dim=c(D,D,K)) # conditional covariances
llik <- vector(length = max_it) # log likelihood of the EM iterations
# Random initialization of the parameters
pi <- runif(K,0,1)
pi <- pi / sum(pi)
for(k in 1:K) {
mu[k,] <- runif(D,0,5)
```

```
Sigma[,,k]<-c(1,0,0,1)
}
```

(3 p) Implement the EM algorithm as described above.

(1 p) Run your code on the data provided with the true number of components, i.e. three.

Show that the log likelihood increases with the number of iterations. Show also that the final parameters are close to the true ones.

(1 p) Use the Bayesian information criterion (BIC) to select among two, three or four components. The BIC is defined as

$$LL - \frac{M}{2} \log N$$

where $LL$ is the log likelihood of the training data given the parameters returned by the EM algorithm, $M$ is the numbers of parameters in the mixture model at hand, and $N$ is the number of points in the learning data.

(1 p) Use the code provided to sample 3000 additional points, which will conform your validation data. Use the log likelihood of the validation data to select among two, three or four components.

(1 p) Discuss in one or two lines if it makes sense to use the log likelihood of the validation data to select the number of components.

## NEURAL NETWORKS - 3 POINTS

Run the code below to train a neural network (NN) for summing two numbers from the interval [-1,1]. Look at the plot of the learned NN and explain why the weights learned make sense. Hints:

- Note that the activation function is `tanh`.
- The two intercepts are so small that you can disregard them.
- Note that the weights in the first layer are the inverse of the weight in the second layer, i.e. $0.13 \approx 1/7.75$.

```
library(neuralnet)
set.seed(1234567890)
x1 <- runif(1000, -1, 1)
x2 <- runif(1000, -1, 1)
tr <- data.frame(x1,x2, y=x1 + x2)
winit <- runif(9, -1, 1)
nn<-neuralnet(formula = y ~ x1 + x2, data = tr, hidden = c(1), act.fct = "tanh")
plot(nn)
```