

# Clustering

## Exercise 1

Q1: What is the fundamental difference between a Gaussian mixture model (GMM) and k-means?

Q2: Keeping the previous question in mind design a dataset that you can cluster correctly using a GMM but not using k-means (3 clusters).

Q3: Implement the EM algorithm in R and perform the clustering of your dataset with a GMM. Also cluster the dataset using k-means. Visualize the results. If you cannot implement it, use the package mclust. The function to use is on the slides.

## Linear models

### Exercise 1

Generate the following data in R and visualize it with colours (points of class 1 in blue and points of class 2 in red for example).

---

---

```
x1=10+10*rnorm(200); # 1st dimension for points of class 1
y1=2*x1+20+10*rnorm(200); # 2nd dimension for points of class 1
x2=10+10*rnorm(300); # 1st dimension for points of class 2
y2=x2-15+10*rnorm(300); # 2nd dimension for points of class 2
outliers=matrix(0,nrow=60,ncol=2); # class 2 outliers
outliers[,1]=35+5*rnorm(60); # 1st dimension for outliers
outliers[,2]=-80+10*rnorm(60); # 2nd dimension for outliers
```

```
X=matrix(0,nrow=length(x1)+length(x2)+nrow(outliers),ncol=2)
X[,1]=c(x1,x2,outliers[,1]); # complete data – 1st dimension
X[,2]=c(y1,y2,outliers[,2]); # complete data – 2nd dimension
t=vector()
for(i in 1:length(x1)) t=c(t,1); # labels for class 1
for(i in 1:(length(x2)+nrow(outliers))) t=c(t,0); # labels for class 2
tls= vector()
for(i in 1:length(x1)) tls=c(tls,1); # labels for class 1
```

```
for(i in 1:(length(x2)+nrow(outliers))) tls= c(tls,-1); # labels for class 2
```

---

---

Q1: what would you expect to happen with Least squares? This is not an easy question... think about it carefully.

Q2: run Least squares and Logistic regression on this data and plot each of the decision boundaries.

## **Exercise 2**

How would you kernelise PCA?

## **Probability theory**

### **Exercise 1**

There is a small island somewhere in the Caribbean. Passenger ships (200 per year) travel from the continent to the island and continue their route to other islands afterwards. Only 2 routes are known through the dangerous sea. The danger of sinking is 10% with the short route, and 5% with the long route. About 20% of the ships decided to take the shorter route.

Q1: how many sinking ships do we expect in a year?

Q2: a sunken ship is reported but we don't know where it is. Which route should we check first?

### **Exercise 2**

After your yearly check-up, the doctor comes up with your test results. Unfortunately you tested positive for a serious disease and the test is 99% accurate (i.e. the probability of testing positive given that you have the disease is 0.99, as is the probability of testing negative given that you don't have the disease). This disease usually strikes only 1 in 10000 people.

Q1: should you be worried?

Q2: can you consider it good news that the test is 99% accurate?

### Exercise 3

The following questions ask you to prove more general versions of probability rules with respect to some background evidence  $E$ .

Q1: prove the conditional version of the product rule:

$$p(A, B | E) = p(A | B, E)p(B | E)$$

Q2: prove the conditional version of Bayes rule:

$$p(A | B, E) = \frac{p(B | A, E)p(A | E)}{p(B | E)}$$

### Exercise 4

This problem investigates the way in which conditional independence relationships affect the amount of information needed for probabilistic calculations.

Q1: we wish to calculate  $p(H | E_1, E_2)$ , and we have no conditional independence information. Which of the following sets of quantities is/are sufficient for the calculations?

1-  $p(E_1, E_2), p(H), p(E_1 | H), p(E_2 | H)$

2-  $p(E_1, E_2), p(H), p(E_1, E_2 | H)$

3-  $p(E_1 | H), p(E_2 | H), p(H)$

Q2: we now know that  $E_1$  and  $E_2$  are conditionally independent given  $H$ . Which of the 3 sets is/are sufficient?

## Bayesian networks

## Exercise 1

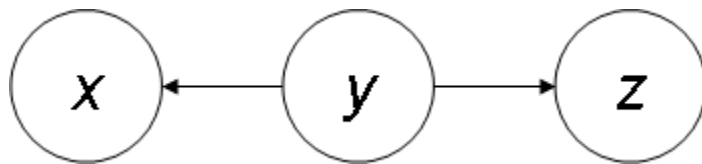
Q1: explain how to read the joint distribution in a Bayesian network.

Q2: what benefits can you see in this representation?

Q3: what is the difference between inference and learning?

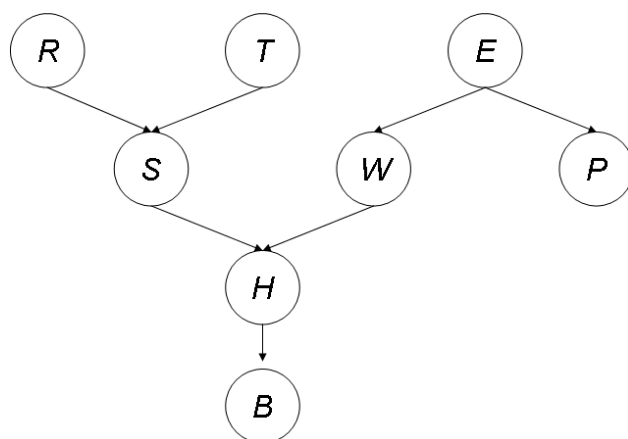
Q3: give the Bayesian network for Bayesian linear regression. And obviously, try not to look at the slides! 😊

## Exercise 2



Explain which arrows can be inverted and why. And obviously, try not to look at the slides! 😊

## Exercise 3



R: rain  $\in \{0,1\}$   
T: temperature  $\in \{\text{cold}, \text{warm}\}$   
S: sick  $\in \{0,1\}$   
E: exam  $\in \{0,1\}$   
P: party  $\in \{0,1\}$   
W: must work  $\in \{0,1\}$   
H: must stay home  $\in \{0,1\}$   
B: bored  $\in \{0,1\}$

Q1: what is the joint distribution of all these variables?

Q2: what can we say about  $R$ ,  $T$  and  $E$ ? About  $W$  and  $P$ ?

Q3: assuming that the appropriate conditional probabilities are known:

- find the probability to get bored when the temperature outside is cold
- find the probability that the temperature is cold when you are bored
- find the probability to have to work when you are sick
- what can you deduce from it?

## **Exercise 4**

Look at the graphical model for Gaussian mixtures on the slides. How would you make this model Bayesian? Draw the new graph.