**Exercise1 (0.5p)**: Given the covariance matrix for two variables: *X1* and *X2*

|  |  |
| --- | --- |
| 16 | 9 |
| 9 | 25 |

calculate the correlation coefficient between *X1* and *X2*. Is the linear correlation strong or weak?

**Exercise2 (1p):** For the following data

|  |  |  |  |
| --- | --- | --- | --- |
| X | 1 | 2 | 3 |
| Y | 1 | 3 | 8 |

one built two regression models:

1. Linear: *Y = 2X-1*
2. Non-Linear: *Y = X2-1*

Let’s evaluate these models using the following loss function *(MAE, mean absolute error)*

where *N* is the number of observations. Which model is better according to this criterion?

**Exercise3 (0.5):** Which of the following functions **can not** be used for modeling the probability of binary random variable *Y*

**Exercise4 (0.5p):** Which statement is **not true** regarding the loss function:

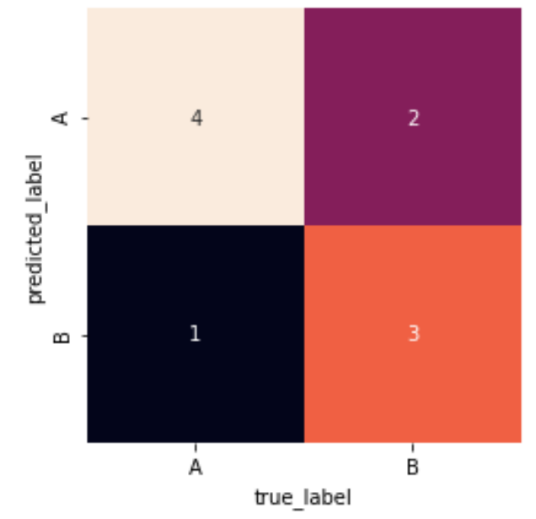
a) It measures the difference between actual values and predictions.

b) During the training step we want to minimize it.

c) For the logistic regression it is usually *MSE (mean squared error)*, for linear regression *BCE (binary cross entropy)*.

d) The smallest value correspond to situation where actual values and predictions equal.

**Execise5 (1p):** Given the confusion matrix

****

Calculate the precision and recall for the category **B**. Is the dataset balanced? Why?

**Exercise6 (0.5p):** One built a model based on the logistic regression. The aim was to predict whether a review is positive (1) or negative (0). A set of classifiers is presenteted on *ROC* curve (classifiers for different *pc* values).

Obraz zawierający szkic, Sztuka dziecięca, rysowanie, sztuka

Opis wygenerowany automatycznie

Which classifier does predict all negative reviews appropriately? Write a number \_\_\_\_\_

**Exercise7 (0.5p):** One built a model based on Bayesian networks to predict relations between some variables among Lung Cancer patients. The following features were considered:

- **Pollution** (P) which is High (H) or Low (L)

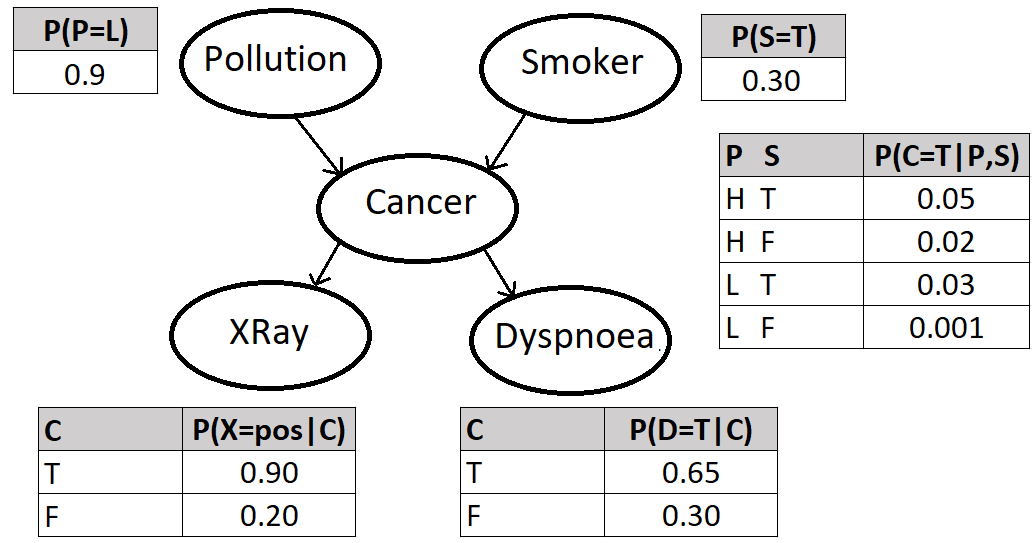
- **Smokers** (S) which is True (T) or False (F)

- **Cancer** (C) which is True (T) or False (F)

- **Dyspnoea** (short breath, D) which is True (T) or False (F)

- **X-ray** (X) which is Positive (pos) or Negative (neg)

and the final model is



According to the model, what is the probability of the following situation: Xray test is positive, Dysponea is present, Cancer is present, Pollution is on low level and a person is non-smoker.

**Exercise8 (1p):** The dataset contains 5 features. After the PCA algorithm the following eigenvalues and corresponding eigenvectors were obtained:

1.60, PC1=[-0.16, 0.08, -0.36, -0.91, -0.10]

0.20, PC2=[-0.47, 0.34, 0.51, -0.21, 0.60]

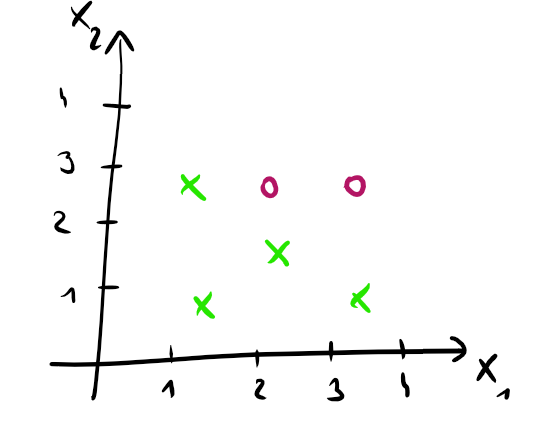
0.10, PC3=[-0.71, 0.01, 0.18, 0.12, -0.66]

0.05, PC4=[-0.34, 0.85, -0.11, 0.13, 0.36]

0.05, PC5=[-0.37, -0.38, -0.75, 0.31, 0.24]

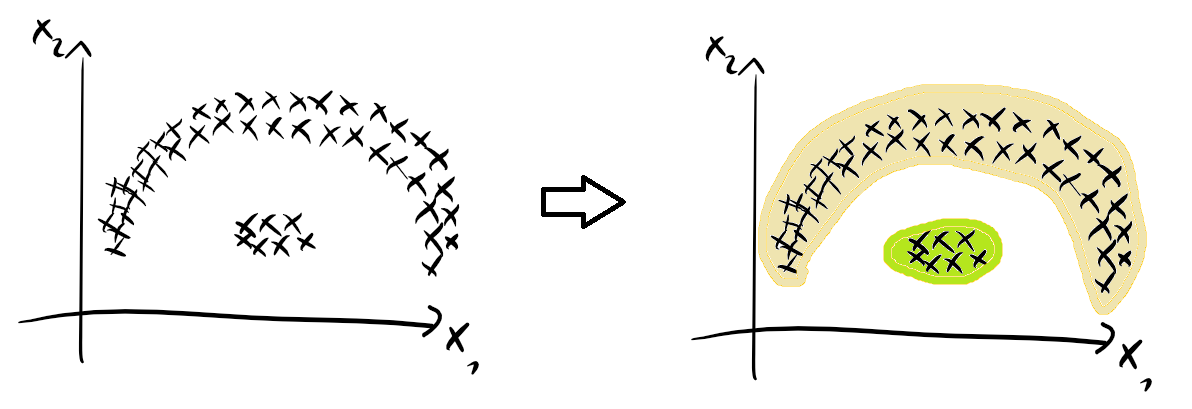
For the further analysis, we want to use the first two components. What is the explained variance of them? Is it sufficient to save the character of the original data? Which original feature seems to have the highest impact?

**Exercise9 (1p):** Consider the following two-classes data



Calculate the Gini index for this data. Build a decision tree with max\_depth = 1 (propose a condition) in such a way that Gini index decreases. Calculate it once again.

**Exercise10 (0.5p):** Consider the following 2D data. One wants to group them in the following way



Which algorithm(s) should we apply for this task?

a) k-means

b) DBSCAN

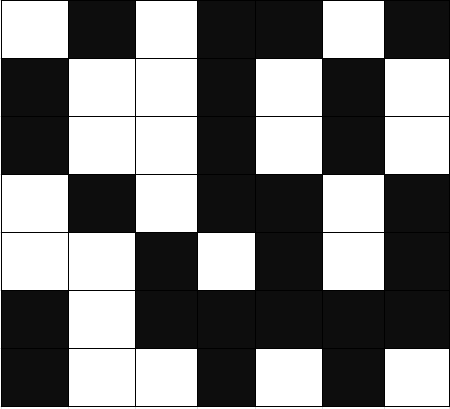
c) mountain grouping

d) Gustafson-Kessel approach

e) all a,b,c,d

f) all except a

**Exercise11 (1p):** Consider the following black and white image



What is the number of independent features (columns) in this case? Justify it.

**Exercise12 (0.5):** In the ICA algorithm, we assume the source (original) signals are

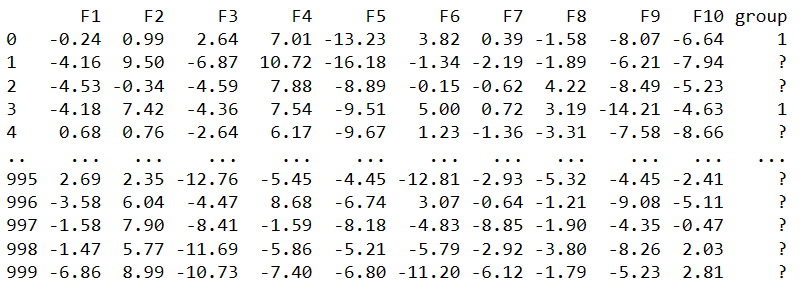
1. gaussian
2. non-gaussian

Underline an appropriate answer and propose an example approach (criterion) to find them.

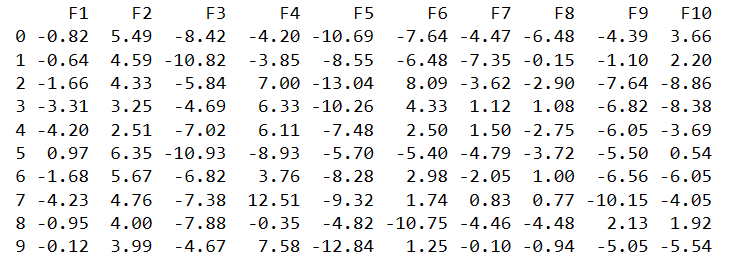
**Exercie13 (0.5):** What is the aim of using the kernel in SVM method? Write 1-2 sentence(s).

**Python exercises**

**Exercise1 (3p):** Consider the *data1* file. There are 10 columns (features) and 1000 rows (patients). For some patients, we know they are infected (1 value in 11th column - group), for the remaining ones we don’t know (denoted as ?). In the second case, patients may be infected or not.

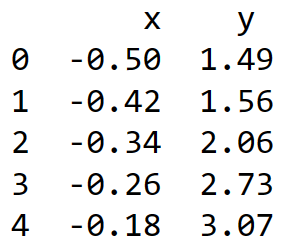


The aim is to recognize patterns in the dataset and finally indicate which of the following patients are infected (*data1test* file).

****

Write indexes of potentially infected people: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Exercise2 (3p):** Consider the *data2* file. There are two columns: x and y. Below, the first four rows



We assume the following relation between these two variables

where **a, b** and **c** are unknown parameters. The aim is to estimate them.

a = \_\_\_

b = \_\_\_

c = \_\_\_