INF 264 **- Exercice** 1

Instructions

- Deadline: 30/08/2019, 23.59
- Submission place: `https://mitt.uib.no/courses/19532/assignments`
- Format: Your answers are to be returned in a single pdf report. You can also return scanned pages for your calculations.

The present document is the first of a series of eight weekly exercises. The deadline for weekly exercises is on Friday evening. You will be awarded points for solving weekly exercises. Each week you can earn at most 1.5 raw points (depending on how many tasks you solve). Your solutions do not need to be perfect; we award points for reasonable effort. Weekly exercises give at most 10 points towards the final grade; if you get more than 10 raw points, it will count as 10 points. To pass the course, you will need to get at least 5/10 points from the weekly exercises. Note that you do not have to get points every week, just at least 5 points in total.

In this exercise we will refresh our memory in relevant maths topics, practice formulating machine learning problems and learn about the k-NN algorithm.

1 Mathematical warm up

1.1 Matrices

Consider the following two matrices:

$$A = \begin{pmatrix} 1 & -1 \\ 1 & 2 \end{pmatrix} \qquad B = \begin{pmatrix} 2 & 1 \\ 1 & -2 \end{pmatrix} \tag{1}$$

1. Compute their respective inverse matrix.
2. Considering two invertible matrices M and N, it holds that MN is invertible and its inverse matrix is $(MN)^{-1} = N^{-1}M^{-1}$. Verify this equality for A and B.
3. It also holds for two real matrices M and N that $(MN)^t = N^tM^t$, where M^t represents the transposed matrix of M. Verify this equality for A and B.

Now consider the matrix:

$$D = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 2 & 0 \\ 2 & 0 & 3 \end{pmatrix} \tag{2}$$

4. Compute the eigenvalues of D. You can for instance solve the equation $det(D - \lambda I) = 0$ with respect to λ.
5. For every eigenvalue λ of D, compute an associated eigenvector v_λ by solving the linear system $(D - \lambda I)v_\lambda = 0$.

1.2 Calculus

Consider the bivariate function $f(x, y) = x^2 + y^2 - 4x + 6y + 13$, whose graph is shown below:

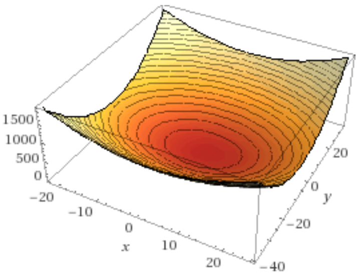

Figure 1: Contour of $f(x,y) = x^2 + y^2 - 4x + 6y + 13$

1. Compute the gradient of f, defined as the vector of the first order derivatives:
$$\nabla f(x,y) = \big(\partial_x f(x,y), \partial_y f(x,y)\big) \tag{3}$$
2. The function f has a unique critical point, that is a point (x,y) such that $\nabla f(x,y) = (0,0)$. Compute this critical point. What value does f take at this critical point ?
3. Factorize f into a sum of two square terms: identify two linear functions $a(x)$ and $b(y)$ such that $f(x,y) = \big(a(x)\big)^2 + \big(b(y)\big)^2$. Deduce from this that the critical point found previously is a local minimum, and even a global minimum.

1.3 Probabilities

The following table sums up the repartition of the students in a high school:

	1st year	2nd year	3rd year	Total
Girls	176	200	200	576
Boys	144	100	140	384
Total	320	300	340	960

1. We randomly choose a student from the high school.
    - What is the probability that this student is in 2nd year ?
    - What is the probability that this student is a girl in 1st year ?
    - What is the probability that this student is not in 3rd year ?

Denote by A and B the events "the randomly chosen student is a girl" and "the randomly chosen student is in 1st year" respectively. Also denote by $P(A|B)$ the probability of A knowing B.

2. What is $P(A|B)$?
3. What is $P(B|A)$?
4. What can you say about the probabilities $P(A|B)$ and $P(B|A)$?

2 Machine learning problems

Recall Mitchell's definition of a well-defined machine learning problem:

> "A computer program is said to learn from experience **E** with respect to some class of tasks **T** and performance measure **P** if its performance at tasks in **T**, as measured by **P**, improves with experience **E**"

Consider the three following scenarios:

1. The grocery store Rema1000 is considering opening a new store in Marineholmen. They know that popularity of different products varies depending on the location of a store. Now, they want to know what kind of sortiment they should choose for the new store. The managers at Rema1000 have heard that machine learning can help solve this kind of problems.
2. Tesla wants to build reliable self-driving cars that don't crash. Cars use different sensors (cameras, radars, GPS) to gather knowledge of their surroundings. Now Tesla is building a steering system to avoid collisions, that involves machine learning.
3. A company is specialized in the destruction of industrial waste. Workers are employed to control mechanic arms that sort the different types of waste in a chain for proper destruction. A recent security investigation showed that the workers were exposed to dangerous substances and the company is expected to fully automate the waste destruction chain. They heard machine learning could help them automate the chain efficiently.

In both situations, you are a consultant specialized in machine learning solutions. Formulate a well-defined machine learning problem for each of the three above tasks. To do this, shortly explain what **T**, **P** and **E** could be in those contexts. Note that there is no single correct answer: a problem may be formulated in several reasonable ways.

- Task **T** is what the machine learning solution does. It may be helpful to think of it as a computational problem that has an input and an output which you should identify. You need not think how that program would work.
- Performance measure **P** tells how well the program works. The performance measure should reflect your customer's goal and it should be measurable in practice.
- Experience **E** specifies the data that is used for learning. Think about what kind of data you would collect.

3 *k*-NN for a classification problem on the Iris dataset

Iris is a small dataset consisting of 150 vectors describing iris flowers, split into three different classes representing three species of the iris family. Each vector comes with a label (the name of the species) and a set of four features which are measurements of different parts of the flower.

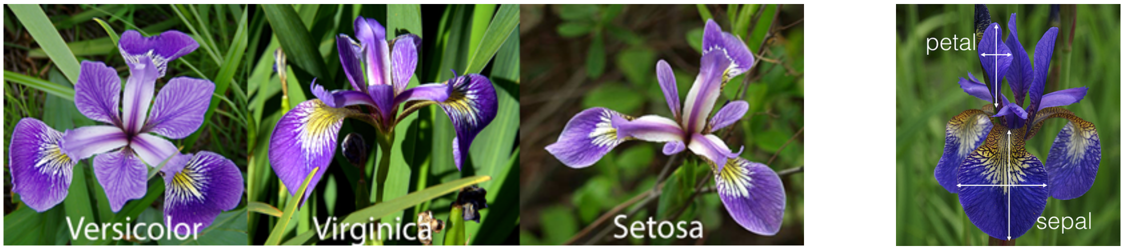

Figure 2: Left: The three species in the Iris dataset
Right: The four features in the Iris dataset (petal and sepal width and length)

Those measurements tend to differ between the different species, thus it is possible for us to learn and evaluate a classifier from this dataset whose task will be to predict the species of an iris flower represented by aforementioned set of features. This last section will have you test a *k*-NN classifier:

1. Load the Iris dataset directly from sklearn. You can alternatively download the dataset here: `https://archive.ics.uci.edu/ml/datasets/iris`. Store the two first features (sepal length and sepal width) in a matrix X. Also store the labels in a vector Y.
2. Split the dataset into a train set and a validation set, i.e. split X and Y into X_{train}, X_{val} and Y_{train}, Y_{val} respectively. You can for instance use a train/validation ratio of 0.7/0.3.
3. Perform a k-NN classification of your dataset for each k in $1, 5, 10, 20, 30$. You can for instance use the KNeighborsClassifier class from sklearn. Plot the decision boundaries with the training points overlayed for every k (since there are three classes, you will need three different colors); the axis of the two selected features must be apparent in your decision boundaries plots. Finally, plot a curve representing the training accuracy as a function of k and same for the validation accuracy.
4. From your observations, for which values of k does k-NN tend to overfit ?
5. For $k = 1$, k-NN train accuracy should be equal to 1 (100% correct predictions). Explain why this is not the case here.