INF 264 - Correction of exercice 1

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. • Inverse of A:

$$A \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 2 \end{pmatrix} \iff A \cdot \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 3 \end{pmatrix} \quad (C_2 \leftarrow C_2 + C_1)$$
$$\iff A \cdot \begin{pmatrix} 1 & \frac{1}{3} \\ 0 & \frac{3}{3} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \quad (C_2 \leftarrow \tfrac{1}{3}C_2)$$
$$\iff A \cdot \begin{pmatrix} \frac{2}{3} & \frac{1}{3} \\ -\frac{1}{3} & \frac{1}{3} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad (C_1 \leftarrow C_1 - C_2)$$

So the inverse of A is

$$A^{-1} = \frac{1}{3}\begin{pmatrix} 2 & 1 \\ -1 & 1 \end{pmatrix}.$$

• Inverse of B:
Similarly, apply the following linear transformations to the columns of B:

i $C_1 \leftarrow C_1 + \frac{1}{2}C_2$
ii $C_1 \leftarrow \frac{2}{5}C_1$
iii $C_2 \leftarrow C_2 - C_1$
iv $C_2 \leftarrow -\frac{1}{2}C_2$,

to find

$$B^{-1} = \frac{1}{5}\begin{pmatrix} 2 & 1 \\ 1 & -2 \end{pmatrix}.$$

2. Straight calculation yields:

$$B^{-1} \cdot A^{-1} = \frac{1}{5}\begin{pmatrix} 2 & 1 \\ 1 & -2 \end{pmatrix} \cdot \frac{1}{3}\begin{pmatrix} 2 & 1 \\ -1 & 1 \end{pmatrix} = \frac{1}{15}\begin{pmatrix} 3 & 3 \\ 4 & -1 \end{pmatrix}.$$

We also have

$$A \cdot B = \begin{pmatrix} 1 & -1 \\ 1 & 2 \end{pmatrix} \cdot \begin{pmatrix} 2 & 1 \\ 1 & -2 \end{pmatrix} = \begin{pmatrix} 1 & 3 \\ 4 & -3 \end{pmatrix}.$$

Apply the following linear transformations to the columns of $A \cdot B$:

i $C_2 \leftarrow C_2 - 3C_1$
ii $C_2 \leftarrow -\frac{1}{15}C_2$
iii $C_1 \leftarrow C_1 - 4C_2$,

to find again

$$(A \cdot B)^{-1} = \frac{1}{15}\begin{pmatrix} 3 & 3 \\ 4 & -1 \end{pmatrix}.$$

3. We have:

$$A^t = \begin{pmatrix} 1 & -1 \\ 1 & 2 \end{pmatrix}^t = \begin{pmatrix} 1 & 1 \\ -1 & 2 \end{pmatrix}, \quad B^t = \begin{pmatrix} 2 & 1 \\ 1 & -2 \end{pmatrix}^t = \begin{pmatrix} 2 & 1 \\ 1 & -2 \end{pmatrix}, \quad (A{\cdot}B)^t = \begin{pmatrix} 1 & 3 \\ 4 & -3 \end{pmatrix}^t = \begin{pmatrix} 1 & 4 \\ 3 & -3 \end{pmatrix}.$$

Direct calculation gives as well:

$$B^t \cdot A^t = \begin{pmatrix} 2 & 1 \\ 1 & -2 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 \\ -1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 4 \\ 3 & -3 \end{pmatrix}.$$

Now consider the matrix:

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

4. Most straightforward way to find the eigenvalues of D is simply to notice that D is in triangular form, thus its eigenvalues are those numbers in its diagonal: 1, 2 and 3. More generally, you could have solved the characteristic polynomial of D, that is solve $P(\lambda) = det(D - \lambda I) = 0$ with respect to the real value λ, where det is the determinant and I is the identity matrix.

5. • $\lambda = 1$: We have immediately

$$D - 1I = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \\ 2 & 0 & 2 \end{pmatrix},$$

so

$$\begin{aligned} (D - 1I) \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} &\iff \begin{cases} 0 = 0 \\ x + y = 0 \\ 2x + 2z = 0 \end{cases} \\ &\iff \begin{cases} x = -y \\ x = -z \end{cases} \\ &\iff (x, y, z) \in < (1, -1, -1) >_{\mathbb{R}} := \{\mu(1, -1, -1), \mu \in \mathbb{R}\} \end{aligned}$$

where $< v_1, \ldots, v_n >_{\mathbb{R}}$ denotes the linear span of the set of vectors $\{v_1, \ldots, v_n\}$ in a $\mathbb{R}$-vector space. This tells us that $(1, -1, -1)$ is an eigenvector of D associated to the eigenvalue $\lambda = 1$.

Remark. The following is a bit more technical, and it is mainly adressed to the students at ease with linear algebra:
What we did is in fact equivalent to finding a basis of the kernel of $D - \lambda I$, since by definition of the kernel of a vector space:

$$(x, y, z) \in ker(D - \lambda I) \iff (D - \lambda I) \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

thus on the above example, $ker(D - 1I) =< (1, -1, -1) >_{\mathbb{R}}$, i.e. the set $\{(1, -1, -1)\}$ forms a basis of $ker(D - 1I)$ since it is a linearly independant set of vectors (this set contains only one non-zero vector) that spans $ker(D - 1I)$. The interesting thing for us is that we can then find the eigenvectors of D only with linear transformations of the columns of $D - \lambda I$ to find a basis of $ker(D - \lambda I)$. For $\lambda = 1$, we can do:

$$(D{-}1I){\cdot}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \\ 2 & 0 & 2 \end{pmatrix} \overset{C_1 \leftarrow C_1 - C_2 - C_3}{\iff} (D{-}1I){\cdot}\begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix},$$

and we thus have that $\{(0, 1, 0), (0, 0, 2)\}$ forms a basis of the image of $D - 1I$ and more interestingly for us that $\{(1, -1, -1)\}$ forms a basis of the kernel of $D - 1I$, so $(1, -1, -1)$ is indeed an eigenvector of D associated to the eigenvalue $\lambda = 1$.

- $\lambda = 2$ and $\lambda = 3$: Using one of the two methods above, we easily find (with the second method it is in fact immediate) that $(0, 1, 0)$ is an eigenvector of D associated to the eigenvalue $\lambda = 2$ and that $(0, 0, 1)$ is an eigenvector of D associated to the eigenvalue $\lambda = 3$.

1.2 Calculus

Consider the bivariate function $f(x, y) = x^2 + y^2 - 4x + 6y + 13$.

1. We have

$$\nabla f(x, y) = \big(\partial_x f(x, y), \partial_y f(x, y)\big) = (2x - 4, 2y + 6) = \big(2(x - 2), 2(y + 3)\big) = 2(x - 2, y + 3)$$

2. Nullifying the gradient of f gives:

$$\nabla f(x, y) = (0, 0) \iff \begin{cases} x - 2 = 0 \\ y + 3 = 0 \end{cases} \iff \begin{cases} x = 2 \\ y = -3 \end{cases},$$

so the only critical point of f is $(2, -3)$, and

$$f(2, -3) = 2^2 + (-3)^2 - 4 \times 2 + 6 \times (-3) + 13 = 4 + 9 - 8 - 18 + 13 = 0.$$

3. We can factorize f as follows:

$$f(x, y) = x^2 + y^2 - 4x + 6y + 13 = x^2 - 4x + 4 + y^2 + 6y + 9 = (x - 2)^2 + (y + 3)^2.$$

From this factorization we deduce that

$$\forall x, y \in \mathbb{R},\ f(x, y) \geq 0 = f(2, -3),$$

which proves that $(2, -3)$ is a global minimum of f (in particular it is also a local 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.
   - The probability that this student is in 2nd year is $\frac{\text{total number of students in 2nd year}}{\text{total number of students}} = \frac{300}{960}$.
   - The probability that this student is a girl in 1st year is $\frac{\text{total number of girls in 1st year}}{\text{total number of students}} = \frac{176}{960}$.
   - The probability that this student is not in 3rd year is $\frac{\text{total number of students not in 3rd year}}{\text{total number of students}} = \frac{\text{total number of students in 1st year+total number of students in 2nd year}}{\text{total number of students}} = \frac{320+300}{960} = \frac{620}{960}$.

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. $P(A|B) = P(\text{Girl}|\text{1st year}) = \frac{\text{total number of girls in 1st year}}{\text{total number of students in 1st year}} = \frac{176}{320}$.
3. $P(B|A) = P(\text{1st year}|\text{Girl}) = \frac{\text{total number of girls in 1st year}}{\text{total number of girls}} = \frac{176}{576}$.
4. The two previous questions prove that $P(A|B)$ and $P(B|A)$ are not equal in general.

2 Machine learning problems

1. Getting rich with Rema 1000:

   - T: a machine learning algorithm which gives a score to each type of sortiment, plus a global expected profit score. A type of sortiment with a score of 1 means that it should be sold no matter what, whereas a score of 0 means you should definitively not sell it in your shop. To predict the score of each type of sortiment and the global expected profit score, the algorithm is fed with features describing different things like the level of urbanism, the local geography, the nationality, the median income...
   - P: maximizing the global expected profit score, identifying type of sortiment with a high score i.e. the algorithm is highly confident in us making profit from this type of sortiment.
   - E: dataset with features related to the level of urbanism, the local geography, the nationality, the median income etc... and the targets are vectors of scores ranging from 0 to 1, where each score in the output vector is associated to a type of sortiment, plus a global score representing the expected profit.

2. Avoiding a bloodshed with Tesla's new self-driving car:

   - T: a machine learning algorithm which takes as input data coming from different sources in real time: radar, GPS and meteorologic informations, but also depth maps representing the distance between the car's camera and its surrounding. The algorithm outputs in real time a score representing the risk of collision, and this information is then sent to the car's auto-pilot mode to have the car react if necessary.
   - P: minimizing the risk of collision score at every moment in time, avoiding at all cost the false negative predicted collisions, limiting the false positive predicted collisions if possible, sending feedback to the auto-pilot mode so that the car reacts in such a way that the temporal evolution of the risk of collision score is smooth over time (small gradient with respect to time, except when the risk of collision score is very high ?).
   - E: dataset which includes:
     - videos captured from the car's camera sensors and converted into video depth maps measuring the distance between the camera sensors and every pixel in the scene
     - radar informations that associates in real time a (position,speed) pair of values to objects surrounding the car
     - GPS informations such as the type of road, the topography of the area, the intensity of the traffic etc...
     - meteorologic informations

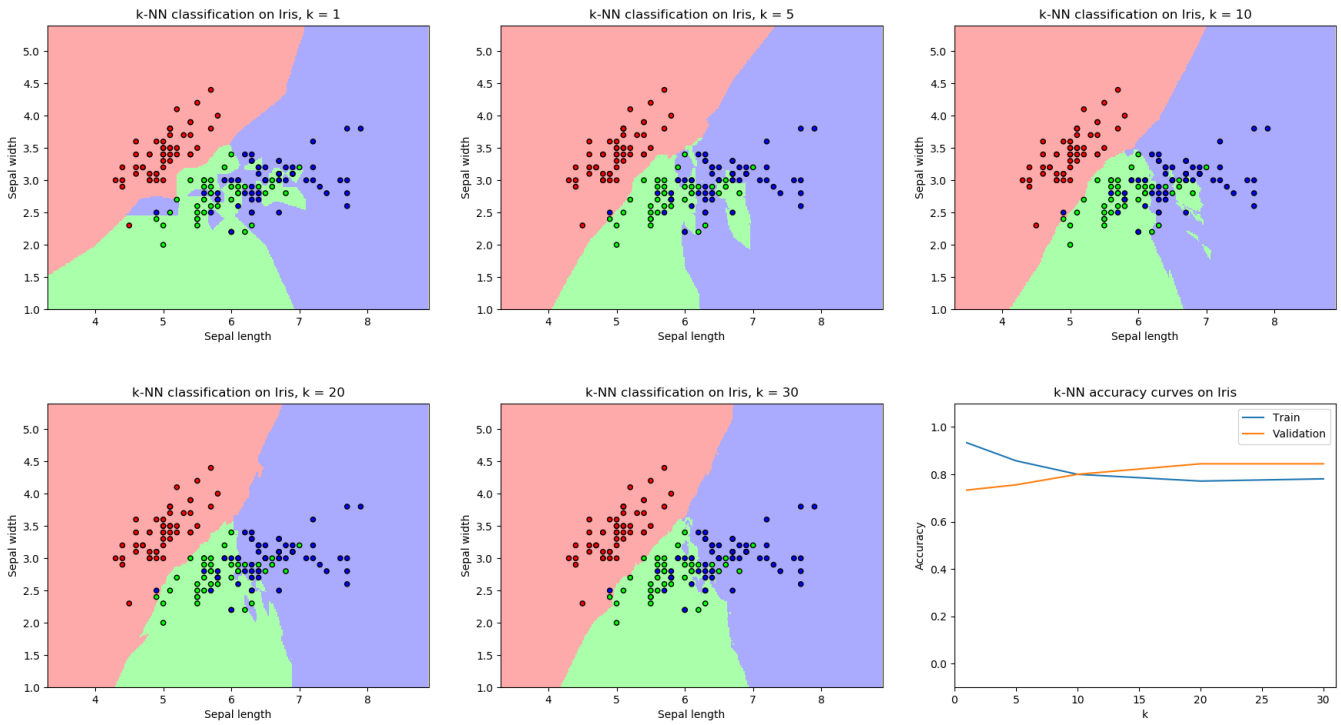
     The target is a score representing the probability of collision if the car does not react.

3. How to ~~lay off~~ protect workers in modern industry:

   - T: a machine learning algorithm which is able from an input video of the wastes moving in the chain to identify the class of each wastes and locate it with high precision in real time. Those informations could be sent to automated mechanical arms which will make sure each type of wastes is sent at the right place.
   - P: maximizing the precision of wastes localization in space, correctly labelling as many wastes as possible, making sure data imbalance is properly handled (difficult problem: many labels and some type of wastes are potentially way more common than others).
   - E: real time videos of wastes moving in the chain. Groundtruth consists of annotated videos in which each pixel in every image from the video is given an integer value associated to its corresponding type of wastes.

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

3. Decision boudaries and train/validation accuracies for $k \in \{1, 5, 10, 20, 30\}$:



4. When k is small, the decision boundaries are not smooth, they follow the training points distribution (notice for $k = 1$ all the small blues regions inside the big green region and conversely). On the contrary, as k increases the decision boudaries become smoother and cleanly defined. This tells us that a small k leads to overfitting as the model learns the noise within the data, and that a higher value of k tends to regularize the decision boundaries, in the sense that it fills the "anomaly regions" by the dominant region.

   An other way to notice the overfitting of k-NN when k is small is to look at the train and validation accuracies with respect to k: When k is small, train accuracy is very high whereas train validation is comparatively significantly smaller. As k increases, this tendency is reversed: train accuracy decreases and validation accuracy on the contrary increases.

5. The reason for imperfect accuracy when $k = 1$ is that in the Iris dataset, there are samples with the exact same 2 first features but having different label.