Assignment 1

Statistical Machine Learning, Summer term 2022

Moritz Haas / Ulrike von Luxburg

due on **Monday, April 25th 2022.**

On this sheet we are going to recap the basic maths that will be needed to follow this course (note that the contents of "Maths for ML" are a prerequisite to the course) and provide the instructions to install Python. There are basic recap documents and links on the course webpage, recap slides at the end of the slides of this class, and youtube videos in this playlist: `https://www.youtube.com/playlist?list=PL05umP7R6ij1a6KdEy8PVE9zoCv6SlHRS`

Exercise 1 (Joint probability, marginal probability, relative frequency, expected value and variance, 1+3 points)
Consider two random variables, the sex X and body height Y of a randomly drawn person. The sex X is binary (male or female) and the body height Y has three values (small, medium, large). The relation between X and Y can be visualized in a contingency table. In this table the JOINT PROBABILITIES are shown. For example we denote the probability to randomly sample a medium sized woman as $P(X = \text{female}, Y = \text{medium})$ and from the table we have, that it is 0.1.

	Y		
X	Small	Medium	Large
Male	0.1	0.15	0.25
Female	0.3	0.1	0.1

(a) The MARGINAL PROBABILITIES refer to the probability of only one variable, $P(X = x)$ or in short $P(x)$. Compute the marginal probabilities of X and Y. We have that $P(X = \text{Male}) + P(X = \text{Female}) = 1$, why?

(b) Calculate the EXPECTED VALUE of X, $E(X)$. Let x_i for $i = 1, \ldots, n$ be i.i.d samples from a distribution X. Then the EMPIRICAL MEAN is defined as

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$$

The empirical mean is an estimate for the expected value. In particular the WEAK LAW OF LARGE NUMBERS holds. It states that for all $\varepsilon > 0$

$$\lim_{n \to \infty} P\left(\left|\bar{X}_n - E(X)\right| > \varepsilon\right) = 0$$

Assuming that both $E(X)$ and $\text{Var}(X)$ are finite, prove the weak law of large numbers. You can use Chebyshev's inequality. It states that if X_i are n random variables i.i.d. as X, with expected value $E(X)$ and variance $\text{Var}(X)$, then for every $\varepsilon > 0$ it holds that

$$P\left(|X - E(X)| \geq \varepsilon\right) \leq \frac{\text{Var}(X)}{\varepsilon^2}.$$

You can also use the following facts. For all $a_i \in \mathbb{R}$,

$$E\left(\sum_i a_i X_i\right) = \sum_i a_i E(X_i)$$

and

$$Var\left(\sum_i a_i X_i\right) = \sum_i a_i^2 \text{Var}(X_i).$$

Exercise 2 (Conditional probability, independence and Bayes theorem, 1+1+2+2 points)

(a) CONDITIONAL PROBABILITIES refer to the probability distribution of one variable given another one. It is denoted by

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A, B)}{P(B)},$$

which reads as the probability of A given B. For example in Exercise 1 we have that $P(Y = \text{Large} \mid X = \text{Male}) = 0.5$. Calculate the probability $P(Y = \text{Medium} \mid X = \text{Female})$.

(b) When are two random variables X and Y INDEPENDENT? Name two characterizations.

(c) The BAYES THEOREM states that

$$P(B = b \mid A = a) = \frac{P(A = a \mid B = b)P(B = b)}{P(A = a)}$$

Let A be the test result for cancer screening, it can be negative or positive, and let B indicate whether the tested patient has cancer or not. The probability of having cancer is 1% and the test is accurate with 95% probability, which for this exercise means that

$$P(A = \text{positive} \mid B = \text{cancer}) = 0.95 \text{ and } P(A = \text{negative} \mid B = \text{no cancer}) = 0.95$$

Compute the probability of having cancer with a positive test result

$$P(B = \text{cancer} \mid A = \text{positive}).$$

Are you surprised by the result? Can you give an informal explanation of why we obtain such result?

(d) The ODDS of having cancer are given by

$$O(B = \text{cancer}) = \frac{P(B = \text{cancer})}{P(B = \text{no cancer})}.$$

This quantity states how many cancer patients you have to expect per person without the disease. The BAYES FACTOR is given by $\frac{P(A=\text{positive}|B=\text{cancer})}{P(A=\text{positive}|B=\text{no cancer})}$ and states how much more likely it is to get a positive test result given a person has cancer compared to when it has no cancer. Can you state the updated odds after a positive test result

$$O(B = \text{cancer} \mid A = \text{positive}) = \frac{P(B = \text{cancer}|A = \text{positive})}{P(B = \text{no cancer}|A = \text{positive})}$$

in terms of $O(B = \text{cancer})$ and the Bayes factor? Why is this view valuable?

Exercise 3 (Linear independence, basis and rank, 1+1+1+1 points)
Consider the following matrix A:

$$A = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 1 \\ 5 & 7 & 8 \end{pmatrix}.$$

We will use this particular 3x3 matrix when we refer to A through the rest of this exercise unless otherwise specified. Also, a vector $x \in \mathbb{R}^d$ always refers to a column vector. We use x^T to refer to a row vector in $\mathbb{R}^d$.

(a) Note that the product Ax for any arbitrary matrix $A \in \mathbb{R}^{3\times3}$ and any $x \in \mathbb{R}^3$ can always be written as a linear combination of the column vectors of A with the elements of x as coefficients. Let $x^T = (x_1, x_2, x_3) \in \mathbb{R}^3$. Write down the explicit form of Ax as a linear combination of the column vectors of A.

(b) Do the columns of A form a basis of $\mathbb{R}^3$? Answer the same question for the rows of A.

(c) Now consider the system of linear equations $Ax = b$; where $x \in \mathbb{R}^3$, $b \in \mathbb{R}^3$. Try to find a $\tilde{b} \in \mathbb{R}^3$ (if such a $\tilde{b}$ exists) so that there is no real valued solution x to the linear system $Ax = \tilde{b}$. If such a $\tilde{b}$ does not exist then explain why. Try to understand the answer to this question in terms of the expression in part (a) of the exercise.

Now consider $b^T = (2, 3, 12)$. Find x that satisfies the relation $Ax = b$.

(d) Just for the sake of completeness, what is the column rank, row rank and rank of the matrix A? Write one line justifying/explaining your answers.

Exercise 4 (Symmetric matrices, eigenvectors, eigenvalues and SVD, 1+2+2+0 points)

(a) *Geometric Visualization of eigenvectors*: Consider the following 2x2 matrices.

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

You can find a java applet to visualize the eigenvectors for 2x2 matrices in this link.

`https://www.geogebra.org/m/KuMAuEnd`. Enable java in your browser in order to access the applet. Using the applet scale and rotate the vector x in order to identify the INDEPENDENT eigenvectors and eigenvalues of the three matrices A, B, C. Explain briefly what you observe in the case of matrix C.

(b) $A \in \mathbb{R}^{n \times n}$ is a symmetric matrix, with a set of eigenvectors $u_1, \ldots, u_n$ with corresponding eigenvalues $\lambda_1, \ldots, \lambda_n$. Derive the eigenvectors and the eigenvalues of the following matrices in terms of eigenvectors and eigenvalues of A.

(1) $A + \alpha I$, where I is the identity matrix of size n and $\alpha \in \mathbb{R}$

(2) $A^T A$

(3) AA^T

(4) If in addition A is a non-singular matrix, then find the eigenvectors and eigenvalues of A^{-1}.

(c) Let $S \in \mathbb{R}^{m \times n}$ with $m \neq n$. Identify the components of the SINGULAR VALUE DECOMPOSITION (SVD) of S given that we have the eigendecomposition of the square symmetric matrices $S^T S$ and SS^T.

(d) To have a better understanding of eigenvalues/vectors and SVD we recommend the following video `https://www.youtube.com/watch?v=PFDu9oVAE-g`. The whole series is worth watching to gain a better geometric understanding of linear algebra.

Exercise 5 (Setting up python, 0 point)

During this course you will be required to implement some of the algorithms presented in class. We will use Python and in particular Jupiter notebooks. In order to save time you should install Python and all the required packages on your laptop. Here you will find the instruction of how to do so. We will present two methods, the first one is the easiest and recommended if you do not have any previous experience with Python. Second one is more suitable for those who know `pip`.

Installation

1) Anaconda. All you need to do is to follow the instructions that you find at the following links. Select the correct one for your operating system and follow the instructions. When you need to decide what to download please download ANACONDA and not MINICONDA. Furthermore download the "Python 3.X version" NOT the "Python 2.X version".

- Windows: `https://conda.io/projects/conda/en/latest/user-guide/install/windows.html`
- MacOS: `https://conda.io/projects/conda/en/latest/user-guide/install/macos.html`
- Linux: `https://conda.io/projects/conda/en/latest/user-guide/install/linux.html`

2) Pip. We will use the following packages: `numpy`, `scikit-learn`, `pandas`, `matplotlib`, `jupyter`.

For example, on Ubuntu, Debian and derivate
`sudo pip3 install numpy scikit-learn pandas matplotlib jupyter`
or
`pip3 --user install numpy scikit-learn pandas matplotlib jupyter`

Test

Now it is time to see if everything we need is installed. Together with this sheet you should have a file named `Assignment 1.ipynb`. We will use it to test that everything is correctly installed.

First thing we need to launch Jupyter. This depends on your operating system

- Windows: Start → "Jupyter Notebook"
- MacOS/Linux: Open a terminal in the folder that contains the `Assignment 1.ipnyb` file and run `jupyter notebook`

Once Jupyter is running, navigate your folder structure until you find the `Assignment 1.ipnyb` file and click on it. Once it is open, please click on `cells` → `Run all`. If it says that you are ready to go then you are ready to go. Otherwise ask for help at the tutorial.