

Information Retrieval in High Dimensional Data  
Assignment #1, 19.11.2020

Due date: 06.12.2020, 10 P.M.

Please hand in your solutions via Moodle as an IPYTHON (Jupyter) notebook.

Solutions can be handed in by the groups. Please state the names of your group members at the beginning of your provided notebook.

## Curse of Dimensionality

**Task 1:** [2 Points] Let  $\mathcal{C}_d = \{\mathbf{x} \in \mathbb{R}^p \mid \|\mathbf{x}\|_\infty \leq \frac{d}{2}\}$  denote the  $p$ -dimensional hypercube of edge length  $d$ , centered at the origin.

- Assume  $X$  to be uniformly distributed in  $\mathcal{C}_1$ . Determine  $d$  in dependence of  $p$  and  $q \in [0, 1]$ , such that

$$\Pr(X \in \mathcal{C}_d) = q$$

holds.

- Let the components of the  $p$ -dimensional random variable  $X^p$  be independent and have the standard normal distribution. It is known that  $\Pr(|X^1| \leq 2.576) = 0.99$ . For an arbitrary  $p$ , determine the probability  $\Pr(\|X^p\|_\infty > 2.576)$  for any of the components of  $X^p$  to lie outside of the interval  $[-2.576, 2.576]$ . Evaluate the value for  $p = 2$ ,  $p = 3$  and  $p = 500$ .

**Task 2:** [10 Points] Provide the PYTHON code to the following tasks (the code needs to be commented properly):

- Sample 100 uniformly distributed random vectors from the box  $[-1, 1]^d$  for  $d = 2$ .
- For each of the 100 vectors determine the minimum angle to all other vectors. Then compute the average of these minimum angles. Note that for two vectors  $x, y$  the cosine of the angle between the two vectors is defined as

$$\cos(\angle(x, y)) = \frac{\langle x, y \rangle}{\|x\| \|y\|}.$$

- Repeat the above for dimensions  $d = 1, \dots, 1000$  and use the results to plot the average minimum angle against the dimension.

- Give an interpretation of the result. What conclusions can you draw for 2 randomly sampled vectors in a  $d$ -dimensional space?
- Does the result change if the sample size increases?

## Statistical Decision Making

**Task 3:** [10 Points] Answer the following questions. All answers must be justified.

- The numbers in Figure 1 show the probability of the respective event to happen (e.g. the probability for the event  $X = 1$  and  $Y = 1$  is 0.02). Is this table a probability table? If so, why?
- Based on Figure 1 give the conditional expectation  $\mathbb{E}_{Y|X=2}[Y]$  and the probability of the event  $X = 1$  under the condition that  $Y = 3$ .
- Is the function  $p(x, y)$  given by

$$p(x, y) = \begin{cases} 1 & \text{for } 0 \leq x \leq 1, 0 \leq y \leq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

a joint density function for two random variables?

- For two random variables  $X$  and  $Y$  the joint density function is given by

$$p(x, y) = \begin{cases} 2e^{-(x+y)} & \text{for } 0 \leq x \leq y, 0 \leq y \\ 0 & \text{otherwise.} \end{cases}$$

What are the marginal density functions for  $X$  and  $Y$  respectively?

- Let the joint density function of two random variables  $X$  and  $Y$  be given by

$$p(x, y) = \begin{cases} \frac{1}{15}(2x + 4y) & \text{for } 0 < x < 3, 0 < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Determine the probability for  $X \leq 2$  under the condition that  $Y = \frac{1}{2}$ .

**Task 4:** [3 Points] Show that the covariance matrix  $\mathbf{C}$  of any random variable  $X \in \mathbb{R}^p$  is symmetric positive semidefinite, i.e.  $\mathbf{C} = \mathbf{C}^\top$  and  $\mathbf{x}^\top \mathbf{C} \mathbf{x} \geq 0$  for any covariance matrix  $\mathbf{C} \in \mathbb{R}^{p \times p}$  and any  $\mathbf{x} \in \mathbb{R}^p$ .

Figure 1: Task 3

