Deadline: Th. December 5th , 10:00 Drop your printed or legible handwritten submissions into the boxes at Samelsonplatz. Alternatively upload a `.pdf` file via LearnWeb. (e.g. exported Jupyter notebook)

1. Distance Measures (14 points)

A. [4p] Compute the **Levenshtein distance** between the two strings MACHINE and LEARNING

		M	A	C	H	I	N	E
	0	1	2	3	4	5	6	7
L	1							
E	2							
A	3							
R	4							
N	5							
I	6							
N	7							
G	8							

B. [3p] Show that the **Hamming distance** between two sets

$$\mathrm{dist}_{\mathrm{Ham}}(X,Y) = |(X \setminus Y) \cup (Y \setminus X)| = |(X \cup Y) \setminus (Y \cap X)|$$

satisfies the 3 properties (positive definiteness, symmetry and triangle inequality) a of distance measures.

Hint: Draw some Venn diagrams!

C. [4p] In $\mathbb{R}^2$, draw all points that are distance 1 away from the origin with respect to

1. The taxicab distance $\mathrm{dist}(x,y) = \|x-y\|_1$
2. The euclidean distance $\mathrm{dist}(x,y) = \|x-y\|_2$
3. The maximum distance $\mathrm{dist}(x,y) = \|x-y\|_\infty$
4. The Mahalanobis distance with matrix $\Sigma^{-1} = \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)$

D. [3p] Given a symmetric, positive definite covariance matrix Σ, show that the **Mahalanobis distance**

$$\mathrm{dist}_{\mathrm{Maha}}(x,y) = \sqrt{(x-y)^{\mathsf{T}}\Sigma^{-1}(x-y)}$$

satisfies the 3 defining properties of a distance measure.

Hint: Try to make use of the following facts from linear algebra.

- If $\langle x|y\rangle$ is an **inner product**, then $\|x\| \overset{\mathrm{def}}{=} \sqrt{\langle x|x\rangle}$ is a **norm**.
- If $\|x\|$ is a **norm**, then $\mathrm{dist}(x,y) \overset{\mathrm{def}}{=} \|x-y\|$ is a **distance measure**.

2. K Nearest Neighbors (10 points)

A. [3p] We want to predict whether or not the topic of a given text is machine learning (document classification task). To do this we use the set (not a multi-set!) of a all capitalized acronyms appearing in the text as features. Given the training data from Table 1, apply a KNN classifier with $K = 3$, using the **Hamming distance**

$$\mathrm{dist}_{\mathrm{Ham}}(X,Y) = |(X \setminus Y) \cup (Y \setminus X)| = |(X \cup Y) \setminus (Y \cap X)|$$

to predict whether the following text is related to machine learning ($y = 0 \leftrightsquigarrow$ no, $y = 1 \leftrightsquigarrow$ yes):

> ...Logistic regression and probit regression are more similar to LDA than ANOVA is, as they also explain a categorical variable by the values of continuous independent variables. These other methods are preferable in applications where it is not reasonable to assume that the independent variables are normally distributed, which is a fundamental assumption of the LDA method. LDA is also closely related to principal component analysis (PCA) and factor analysis in that they both look for linear combinations of variables which best explain the data.

Figure 1: Excerpt from the Wikipedia article on Linear Discriminant Analysis

X	Y
{KNN, SBS, PSB}	0
{ANOVA, DOE}	1
{DUI, LDA, USA}	0
{SVD, ONB, GSL, PCA}	1
{ML, XML, HTML}	0
{LDA, QDA, PCA}	1

Table 1: training data

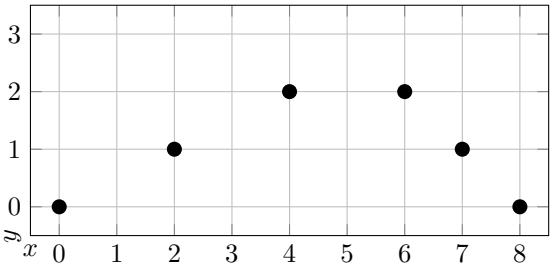
B. [2p] One of the problems of the distance metric used in 2A is that two text will automatically be seen as dissimilar if one of them contains a lot of acronyms and the other does not, irrespective of their topics. Explain qualitatively how the use of the **Jaccard distance**

$$\mathrm{dist}_{\mathrm{Jac}}(X,Y) = 1 - \frac{|X \cap Y|}{|X \cup Y|} = \frac{\mathrm{dist}_{\mathrm{Ham}}(X,Y)}{|X \cup Y|}$$

instead of the **Hamming distance** would circumvent this problem.

C. [3p] Given the dataset from Figure 2, use KNN Regression to predict the value at $x = 5$. Use the taxicab metric and try with $K = 2$, $K = 3$ and $K = 4$. Comment on the result.

Figure 2: toy dataset



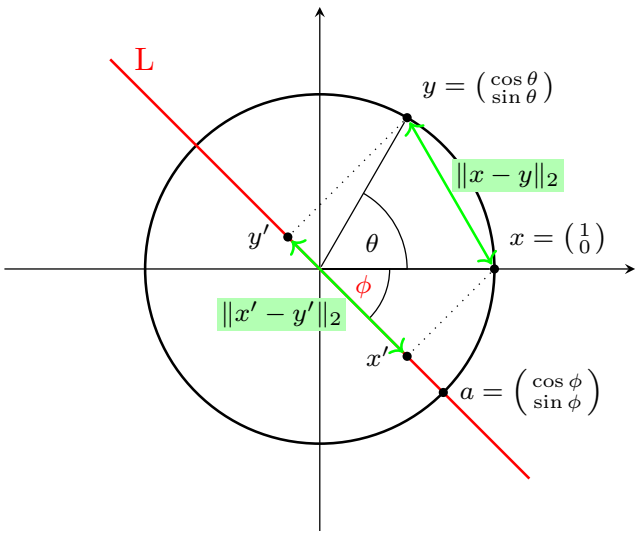
D. [2p] Given that the ground truth is

$$y(x) = \begin{cases} 0.5x: & x \leq 5 \\ 7.5 - x: & x > 5 \end{cases}$$

Explain why a KNN model would never be able to predict the true value $y(5) = 2.5$, irrespective of the choice of K or distance metric.

3⋆ Locality Sensitive Hashing (4 points)

The idea of LSH is that, in some sense, if x and y are close points in a high dimensional space, then – on average – they stay close when projecting them onto a randomly selected low dimensional space; and when they are far away they – on average – stay far away. Given the two datapoints on the unit circle $x = \binom{1}{0}$ and $y = \binom{\cos(\theta)}{\sin(\theta)}$ and a randomly selected line through the origin with parallel vector $a = \binom{\cos(\phi)}{\sin(\phi)}$. We are interested in comparing the distance of the two points x, y in $\mathbb{R}^2$ against the distance of their projections x', y' onto the random line spanned by a.



- Show that $\|x - y\|_2^2 = 2(1 - \cos(\theta))$
- Show that $\mathbb{E}_{\phi \sim [0,2\pi]}[\|x' - y'\|_2^2] = 1 - \cos(\theta)$

In particular, on average the error made by estimating $\|x - y\|_2$ with $\|x' - y'\|_2$ is around $\sqrt{2} - 1 \approx 40\%$. This value gets reduced more when one moves to higher dimensional spaces.