Machine Learning I

Final Exam

Prof. Dr. Dr. Lars Schmidt-Thieme
M.Sc. Randolf Scholz

February 15th, 2019
Information Systems and Machine Learning Lab (ISMLL)
Universität Hildesheim

Problem	Points
1	/10
2	/10
3	/10
4	/10
Bonus:	/4
Total:	/40
Grade:	
Name:	
Surname:	
Matrikel:	

Note:

- Time: 120 minutes
- Write your name and matriculation number at the bottom of each sheet!
- Please only write your answers into the designated boxes. If you run out of space, continue on the "Extra Space" sheet that belongs to the problem. Indicate which part of the problem you are working on. If you still need extra paper, ask a supervisor.
- A couple of problems ask you to draw into already given plots. If you need to correct your initial drawing, you will have to erase your previous one or sketch the plot again.

Reserve Space

Do not write here unless you run out of space.

1. Linear Classification

1A. LDA and LR (2 points)

Explain the differences and commonalities between Linear Discriminant Analysis and Logistic Regression.

1B. Binary Classification (5 points)

Train a logistic regression classifier on the training data from table 1 by performing 1 iteration of Gradient Ascent with step-size $\alpha = 0.5$. Initialize with $\beta_1 = 1$ and $\beta_2 = 1$ and bias $\beta_0 = 0$.

Sketch the decision boundary before and after the gradient update (use figures 1 and 2). Which points are correctly classified in each case?

x_1	x_2	y
0	+2	1
−1	−2	0
+3	0	0

Table 1

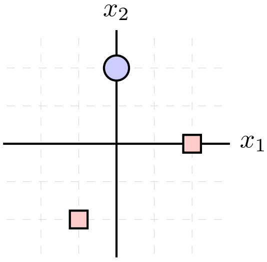

Figure 1: before update

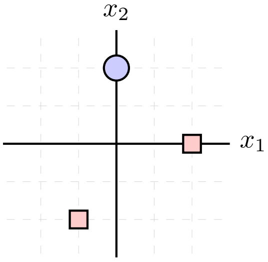

Figure 2: after update

1C. Newton's Method (3 points)

At which point $(x^*, f(x^*))$ does the function $f(x) = xe^{-x}$ attain its maximum value? Secondly, estimate x^* by performing 3 iterations of Newton's Method, starting at $x_0 = 0$. (Learning rate $\alpha = 1$)

2. K-Nearest-Neighbors

2A. KNN basics (3 points)

Explain in 2-3 sentences how the K-Nearest-Neighbor method works. Secondly, point out its main advantages and disadvantages.

2B. KNN prediction (4 points)

We want to predict whether or not a given text is related to Machine Learning or not. To do this we use the set of a all capitalized acronyms appearing in the text. Given the training data from table 2, use a K-Nearest-Neighbor classifier with $K = 3$ to predict whether the following text is related to Machine Learning :

...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.

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 2

Note: Use the **Hamming distance** as the metric!

$$\mathrm{dist}_{\mathrm{Ham}}(X,Y) = \big|(X \setminus Y) \cup (Y \setminus X)\big| \tag{1}$$

2C. Distance Measures (3 points)

In problem 2B the Hamming-distance was used. This has the disadvantage that two texts are automatically seen as dissimilar if one of them uses a lot of acronyms while the other does not, even if they are about exactly the same topic. To improve the model, your colleague suggests to use the the **Jaccard-distance** instead of the Hamming distance.

$$\mathrm{dist}_{\mathrm{Jac}}(X,Y) = 1 - \frac{|X \cap Y|}{|X \cup Y|} \tag{2}$$

Explain qualitatively why the Jaccard distance is better suited for this task, and demonstrate your argument on a simple example.

3. Decision Trees

3A. Hyperparameters (2 points)

Give a short explanation for what a hyperparameter is and how one can find good hyperparameters. Secondly, provide 3 examples of hyperparameters in regards to decision trees.

3B. Decision Tree Learning Algorithm (5 points)

We want to predict gender of an elephant given its weight and species (Asian elephant `Elephas maximus` or African elephant `Loxodonta africana`).

Train a Decision Tree of depth 2 using the Gini Index splitting criterion on the data-set provided by table 3. Draw the learned tree.

Weight	Species	Gender
5500kg	African	male
3500kg	African	female
3400kg	Asian	male
2700kg	Asian	female

Table 3

3C. Decision Trees II (3 points)

Draw a minimal depth Decision tree that solves the classification problem from 3B. Why is the greedy search algorithm not able to find this solution? How could one modify the learning procedure or the model, such that it does learn a decision tree of depth ≤ 2 for this task?

4. Unsupervised Learning

4A. Expectation-Maximization (3 points)

Sketch the EM-Algorithm for Gaussian Mixture Models as pseudo-code. You do not need to provide the full formulas, it is sufficient to explain which quantities are computed in each step.

4B. K-Means Clustering (5 points)

Perform K-means clustering ($K = 3$) on the data-set provided by table 4. Perform 1 iteration, using P3, P6 and P8 as the initial clusters. You can use the data from table 5 to speed up the computation. Draw a loop around each of the resulting clusters in figure 3. Mark the points that changed their cluster affiliation.

ID	x_1	x_2
P1	0	0
P2	1	2
P3	−1	3
P4	0	3
P5	1	0
P6	2	0
P7	0	1
P8	1	2

Table 4

ID	P1	P2	P3	P4	P5	P6	P7	P8
P1	0	5	10	9	1	4	1	5
P2		0	5	2	4	5	10	16
P3			0	1	12	17	17	28
P4				0	10	12	16	25
P5					0	1	2	4
P6						0	5	5
P7							0	2
P8								0

Table 5: **squared** euclidean distances

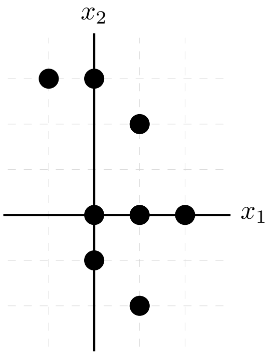

Figure 3

4C. Frequent Pattern Mining (2 points)

Define frequency and support in the context of frequent pattern mining and explain in your own words the difference. Moreover, explain what a frequent subset and a maximal frequent subset is.