Machine Learning I

Final Exam

Prof. Dr. Dr. Lars Schmidt-Thieme
Hadi S. Jomaa

February 12th, 2021
ISMLL Universität Hildesheim

Problem	A	B	C	Σ
1. Linear Models				/10
2. Neural Networks				/10
3. Decision Trees				/10
4. Unsupervised Learning				/10

First Name:		Bonus:	/04
Last Name:		Total:	/40
Matrikel:		Grade:	

Note:

- Time: 120 minutes
- Add your name and matrikel on top of every **odd numbered page**.
- No electronic devices besides scientific calculators and clocks are allowed! If we catch you using any other devices you automatically failed the exam.
- Write with a **non-erasable pen**,do not use red color.
- Use the **reserve pages** at the end if you run out of space. If you still need more, ask staff for rextra pages.
- **Always** explain your reasoning unless asked otherwise.

Reserve Page – clearly indicate which problem you are working on!

1. Linear Models

1A. Linear Regression Use Case (3 points)

You plan on driving across the country with your car but you are not sure how much you will need to *pay* for gas? Describe how you would build a linear regression model to plan your gas budget, clearly stating the target and each variable (Name at least 3 variables). Describe a process where you can verify if weather plays a role in how much you will pay.

1B. Spam or Ham (5+2 points)

You are sick of spam emails popping up in your inbox so you decide to build a spam classification model using logistic regression given the data from Table 1, where x_1 and x_2 are some character frequencies.

x_1	x_2	y
0	1	SPAM
0.06	0.75	SPAM
0.37	0	HAM
1	0.72	HAM

Table 1

1. Write down the optimization objective. Is it a mimization or a maximization objective ?
2. Starting with an initial value of $\beta^0 = (0.5, 0.5)$ and a learning rate $\eta = 1$:
   a. Perform 1 iteration of gradient ascent (descent?) and note down the loss values before and after the update.
   - **Bonus** Perform 1 iteration of Newton's method starting and note down the loss values after the update.
3. You receive afterwards a new email: $(0.15, 0.25)$. Is it SPAM or HAM based on your two models ?
4. Can we use Linear Discriminant Analysis instead of logistic regression ? Explain.

1C. Learning Problems (2 points)

Name 2 methods you would use when the performance of your model on the training data is great, and poor on the validation data. Explain the difference between both.

2. Neural Networks

2A. Neural Networks (3 points)

1. Provide the scheme of a feed-forward neural network model that is equivalent to a logistic binary linear regression model.
2. What is an activation function? Give two examples with their formulas.

2B. Backpropagation (5 points)

Consider the neural network shown in Figure 1. Here, the first hidden layers uses a ReLU activation and the output layer uses a linear activation (i.e. identity) function. The weights are initialized as

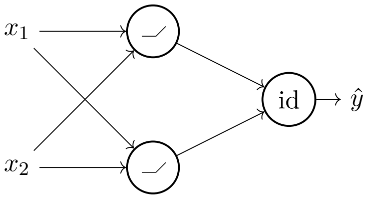


Figure 1: Neural Network

$$\beta_1 = \begin{bmatrix} 1 & -\frac{5}{2} \\ -1 & 1 \end{bmatrix} \qquad \beta_2 = \begin{bmatrix} -\frac{1}{2} & \frac{1}{5} \end{bmatrix}$$

Given $x = (\frac{1}{3})$, $y = \frac{3}{2}$, perform one forward pass and report the absolute error $L(\hat{y}, y) = (\hat{y} - y)^2$. Then perform one backward pass w.r.t. to this loss and update all weights once. Use learning rate $\eta = \frac{1}{10}$, and then report the new prediction (i.e. one more forward pass).

2C. Linear Separability (2+2 points)

1. Create a sketch of the data. Is it linearly seperable? If so, draw a separating hyperplane.

x	y
-3	-1
-2	-1
-1	1
0	1
1	1
2	-1
3	-1

Table 2

2. Apply the mapping $g : \mathbb{R} \longrightarrow \mathbb{R}^2$ defined by

$$g(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix}$$

on all the data points to create the transformed data set and create a plot of it. Is the transformed data set linearly seperable? If yes, find a seperating hyperplane H_β, compute its parameters β and plot it.

3. **Bonus** Find all $x \in \mathbb{R}$ such that $g(x) \in H_\beta$, i.e. all one dimensional points that map to this hyperplane using g. Does the resulting set seperate the initial data points?

3. Decision Trees

3A. Decision Tree From 2D Data (3 points)

Construct a decision tree (without explicitly training) which realizes this partition.

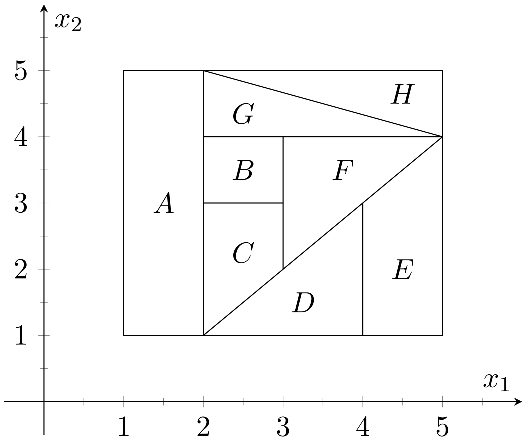


Figure 2: Partition

3B. Decision Tree Training (5 points)

We want to predict if a car is stolen based on some of its attributes. Given the data in Table 3, train and draw a binary decision tree (including labels at the leaves and decisions at the nodes) for predicting the target using the misclassification rate as the splitting criterion. (Let the True leaf be to the left.)

Color	Type	Origin	Stolen
Red	Sports	Domestic	Yes
Red	Sports	Domestic	No
Blue	Sports	Domestic	Yes
Blue	Sports	Domestic	No
Blue	Sports	Imported	Yes
Blue	Grand tourer	Imported	No
Blue	Grand tourer	Imported	Yes
Blue	Grand tourer	Domestic	No
Red	Grand tourer	Imported	Yes
Red	Sports	Imported	Yes

Table 3

3C. Cost-Complexity Pruning (2 points)

We define the training error of a subtree T_t with node t as its root as $R(T_t) = \sum_t^{f(T_t)} R(t)$, where $f(T_t)$ is a function that returns the leaves in the subtree, and $R(t) = r(t)p(t)$ is the misclassification error $r(t)$ at node t (without considering the leaves) multiplied by the proportion $p(t)$ of instances at node t with respect to the data.
Cost-Complexity pruning is the process of iteratively removing the subtree with the least value $g(t)$, representing the reduction in error, until there are no subtrees left. The final subtree is selected based on a pre-defined condition, i.e. $x < g(t) < y$.
Given the tree in Figure 3, and $g(t) = \frac{R(t)-R(T_t)}{|f(T_t)|-1}$, prune the tree for $g(t) > \frac{1}{8}$. (Hint: We have 3 inner nodes that we can prune.)

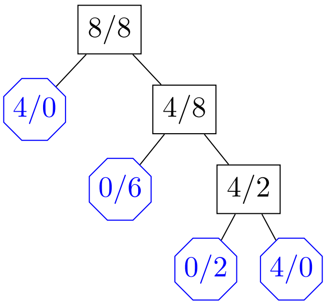


Figure 3: Decision tree for 3C

4. Unsupervised Learning

4A. Clustering (3 points)

1. What is the difference between soft clustering and hard clustering ?
2. Can we draw a dendogram to analyze K-means clustering? Why? Why not ?

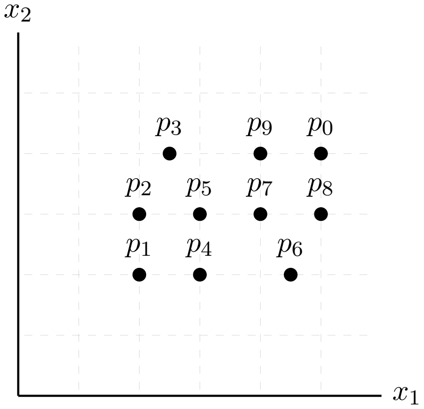
4B. Hierarchical Clustering (5 points)

Apply Hierarchical Clustering to the data from Table 4. Use the Manhattan distance and single (i.e. minimum) linkage as the distortion measure. Perform **agglomerative Hierarchical Clustering** using **single linkage** as the cluster distance measure. Draw the associated dendogram.

Table 4

id	x_1	x_2
p_0	4	3
p_1	1	1
p_2	1	2
p_3	1.5	3
p_4	2	1
p_5	2	2
p_6	3.5	1
p_7	3	2
p_8	4	2
p_9	3	3

Figure 4



4C. Depth First Search (2 points)

You work at a pharmacy where you keep track of the transactions made, Table 5. Create a list of items to recommend based on what the customers purchase using the ECLAT algorithm. Note: The customers should purchase at least one product before you can make a recommendation, i.e. minimum support = 2.

Id	Mask	Sanitizer	Panadol	Mouth Wash	Gloves
T1	1	1	0	0	1
T2	0	1	0	1	0
T3	0	1	1	0	0
T4	1	1	0	1	0
T5	1	0	1	0	0
T6	0	1	1	0	0
T7	1	0	1	0	0
T8	1	1	1	0	1
T9	1	1	1	0	0

Table 5