Algorithmic Foundations of Data Science

Exam 1, 14.08.2020

Write your name and student ID on every sheet of paper you submit.

Instructions

- **This is a closed-book exam.** Only use your electronic device to scroll through the exam or (but only if asked to) to operate Zoom. Close all other tabs and programs. Remove all other material from your surroundings. If you use two different devices for connecting to Zoom and for displaying the PDF, this applies to both of them. **Turn off all other devices. Calculators** are also **not** allowed.
- **Your camera and microphone have to be enabled at all times!**
- **You may use your own (empty!) paper.** Do not print the exam. Only use **black** or **blue** document-proof pens that are good to read on scan. Do **not** use pencils.
- You are also allowed to write down your solutions digitally (iPad notes / graphics tablet) instead. (You are then additionally allowed to have the necessary tool open to do so.) Please let your supervisor know via chat that you are doing so.
- If you need to use the restroom, contact a supervisor via Zoom (Chat) about it.
- You may answer in either English or German. Do not mix languages though.
- Clearly mark your solutions and results as such. If you provide **multiple** solutions to a question, the **worst** of those counts.
- You have **120 minutes** time on the exam. Your supervisors will tell you when it officially starts.
- When the exam ends, you will be notified by your supervisors. There are up to 30 minutes to scan and upload your solutions. During this time you have to stay connected to the room and should leave camera and microphone on. If this is not possible, ask your supervisors how to proceed.
- With **60 points** or more, you will have passed this exam.

Cheating attempts are taken very seriously. You declared in lieu of oath that you do not attempt to cheat.

If your connection drops, immediately try to rejoin and notify the supervisor of your room via chat. If no internet communication is possible, call +49 241 80 21716 (Germany, charges may apply).

Do not write below this line.

	1	2	3	4	5	6
Points	/ 20	/ 20	/ 19	/ 22	/ 19	/ 20
Sign.						

$\sum$	/ 120

Problem 1 (General Questions) 4+4+4+4+4 = 20 points

a) What are the hypothesis spaces in the following algorithms / learning scenarios?

- Perceptron algorithm in n dimensions
- decicision tree learning for Boolean classification with k input features where every feature has at most d values

Be as precise as possible.

b) Sketch the k-Nearest-Neighbour algorithm and answer the following questions:

- What is the underlying assumption on the target function?
- What happens if k is chosen too small or too large?

c) What is the informal statement of the *Occam's razor* principle with respect to learning hypotheses? Roughly describe the formal result justifying this principle, and comment on the impact of the particular description scheme that is chosen for representing hypotheses.

d) In the scenario of playing on different slot machines, briefly explain the concepts of exploration and exploitation. Why is either extreme case (putting too much weight on one or the other) problematic?

e) Verbally define the notions of *principal components*, and the *best-fit k-dimensional subspace* of a data matrix $A \in \mathbb{R}^{n \times \ell}$. What is the connection between these two?

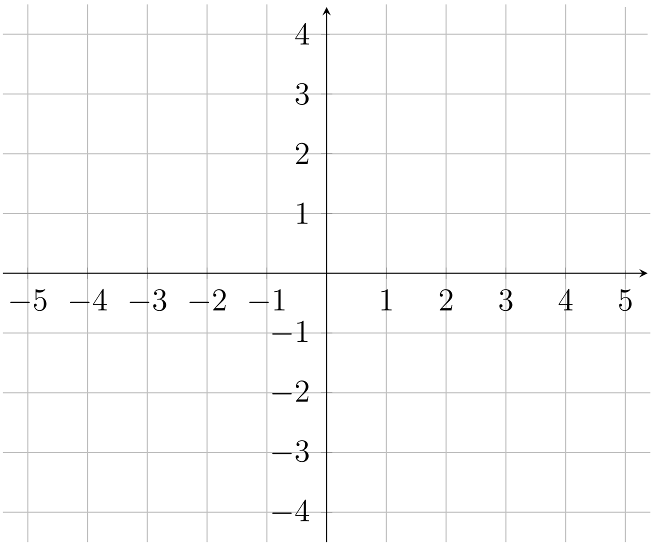
Problem 2 (Linear Separators) 1+6+5+3+5 = 20 points

We consider a Boolean linear classification problem with data points in $\mathbb{R}^2$ (and class labels -1 and 1). Given is the following training sequence:

$$S = \left\{ \left(\left(\begin{smallmatrix} 4 \\ -1 \end{smallmatrix}\right), -1\right), \left(\left(\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right), -1\right), \left(\left(\begin{smallmatrix} 0 \\ 3 \end{smallmatrix}\right), -1\right), \left(\left(\begin{smallmatrix} -1 \\ -1 \end{smallmatrix}\right), 1\right), \left(\left(\begin{smallmatrix} 0 \\ -3 \end{smallmatrix}\right), 1\right) \right\}$$

(The colors are only used to increase the readibility of the class labels.)

a) Plot the data points from the training sequence in a coordinate system such as the one below. Indicate the class labels by annotating the points with $+1$ and -1.



b) Run the Perceptron algorithm on the training sequence in order to algorithmically find a homogeneous linear separator for the classification problem. **Caution:** Assume that the algorithm considers the points exactly in the above order.

Specifically, do the following:

i) For each round of the algorithm, explicitly indicate the considered point, and explicitly calculate the updated weight vector.

ii) Plot the hypotheses from the intermediate weight vectors after each round of the algorithm into the plot from a) (indicate which round they belong to, and indicate which one is the final one).

c) Calculate the margin of S with respect to the linear separator you found in part b). For this, calculate *all* necessary values that need to be taken into consideration.

d) What is the maximum possible margin of a homogeneous linear separator for the given training sequence? (You may find the answer by reading it off your plot from part a) of this problem.)

Explain why your answer is correct.

e) For S, give the SVM formulation for the problem of finding a homogeneous linear separator. Specifically, solve the following tasks / answer the following questions:

- Formulate the SVM as a quadratic program, with plugging in the explicit values from S. Your final program should not contain any scalar products or norms. You may omit duplicate inequalities.
- What specific property does the hypothesis returned by the SVM exhibit?

Name: Student ID:

Problem 3 (Entropy) 4+4+5+6 = 19 points

We are given a set of examples regarding the weekend exercise activities of a person. Based on an activity type (**A**), a day (**D**), and the expenditure of time (**T**), the person makes a yes/no choice (encoded as 1, resp. 0) as indicated below.

Ex. no.	**A**ctivity	**D**ay	**T**ime	Choice
1	Running	Sat	Short	0
2	Cycling	Sun	Long	1
3	Swimming	Sat	Long	1
4	Running	Sat	Long	1
5	Cycling	Sat	Short	0
6	Swimming	Sun	Long	1

a) Give a decision tree for the Boolean classifcation problem above such that

- every feature appears at least once as a node label in your tree; and
- your tree is consistent with the examples given above.

b) Lat X be a random variable with finite range $\mathrm{rg}(X)$ (that is, $\mathrm{rg}(X)$ is the finite set of values that X can attain). Solve the following tasks / answer the following questions:

- Give the formal definition of the entropy $\mathrm{H}(X)$ of the random variable X.
- What is the intuitive meaning of the entropy of a random variable with respect to the information content of events?

c) Solve the following tasks.

i) What are the minimal and maximal possible values for $\mathrm{H}(X)$?

ii) For which distributions of X is $\mathrm{H}(X)$ minimal? Prove this.

iii) For which distributions of X is $\mathrm{H}(X)$ maximal? (You do not need to justify this.)

d) Recall that in the context of constructing decision trees, the *information gain* $G(S, A)$ of a feature A is defined as

$$G(S, A) := \mathrm{H}(\mathcal{P}) - \sum_{x \in \mathbb{D}_A} \frac{|S_{A=x}|}{|S|} \cdot \mathrm{H}(\mathcal{P}_{A=x})$$

where $\mathbb{D}_A$ is the set of possible values of the feature A, for suitably defined $\mathcal{P}$, $\mathcal{P}_{A=x}$ and $S_{A=x}$.

Solve the following tasks.

i) Describe what $\mathcal{P}$, $\mathcal{P}_{A=x}$ and $S_{A=x}$ are in the concrete example from a)

ii) What is the information gain of the feature $\mathbf{T}$ (given S)? (You can leave your answer in terms of logarithms.)

iii) Deduce that $\mathbf{T}$ must be a feature of maximum information gain.

Name: Student ID:

Problem 4 (Markov Chains) 3+6+4+6+3 = 22 points

In this problem we consider two Markov chains $\mathcal{Q}_1$ and $\mathcal{Q}_2$ that are given by their transition matrices Q_1 and Q_2, respecively.

$$Q_1 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \qquad Q_2 = \frac{1}{3}\begin{pmatrix} 1 & 1 & 1 \\ 2 & 0 & 1 \\ 2 & 1 & 0 \end{pmatrix}$$

a) For both $\mathcal{Q}_1$ and $\mathcal{Q}_2$, give the graphical representation of the Markov chain. In your drawing, **number the states** such that they fit the transition matrix; and **label the edges** with their respective transition probability. (The absence of an edge indicates probability 0.)

b) Determine the stationary distributions $\boldsymbol{\pi}^{(1)}$ of $\mathcal{Q}_1$ and $\boldsymbol{\pi}^{(2)}$ of $\mathcal{Q}_2$. Justify the correctness of your solution.

c) For both of $\mathcal{Q}_1$ and $\mathcal{Q}_2$ determine whether or not the Markov chain is ergodic. Give a brief explanation to justify your answer.

d) Consider the probability vector $\boldsymbol{p}_0 = \left(\frac{2}{7}, \frac{1}{7}, \frac{4}{7}\right)$ and let

$$\boldsymbol{p}_t^{(1)} := \boldsymbol{p}_0 Q_1 \qquad \text{and} \qquad \boldsymbol{p}_t^{(2)} := \boldsymbol{p}_0 Q_2$$

Give the (approximate) numeric value of both $\boldsymbol{p}_{10000}^{(1)}$ and $\boldsymbol{p}_{10000}^{(2)}$. Explain why your solutions are equal (or at least expected to be very close) to the true value.

e) The chains $\mathcal{Q}_1$ and $\mathcal{Q}_2$ have a unique stationary distribution, but not every Markov chain has this property. Give the graphical representation of a Markov chain $\mathcal{Q}$ that does not have a unique stationary distribution.

Provide two different probability vectors $\boldsymbol{\pi}$ satisfying $\boldsymbol{\pi} Q = \boldsymbol{\pi}$ where Q is the transition matrix of your chain.

Problem 5 (Map-Reduce) 5+10+4 = 19 points

In this problem we consider data records about radio stations. The data is given in key-value pairs of the shape $(s, (g, c, f))$ where

- s is a radio station,
- g is the music genre s is playing,
- c is the country where s can be listened to (we assume that stations are always broadcasting throughout the whole country and that every station is located in at most one country)
- f is the frequency range that s broadcasts on.

We work with this data in the Map-Reduce programming model.

a) Answer the following questions / solve the following tasks.

i) Name all phases of Map-Reduce algorithms (as taught in the lecture) *in the correct order.*

ii) In which shape is data handled in Map-Reduce algorithms?

iii) In the Map-Reduce programming model, when, and how is data grouped?

b) Given the setup from the introduction to this problem, specify *single-round* Map-Reduce algorithms for the following problems in pseudocode.

i) *Genre popularity per country.* Output all key-value pairs $((c, g), n)$ where c is a country, g is a genre and n is the number of stations in country c that play music of genre g.

ii) *Similar stations.* For every station s, output all key-value pairs $(s, (s', f'))$ such that s' is a radio station in the same country as s, and there exists a genre g such that both s and s' play music of genre g, and f' is the frequency range of s'. (You are allowed to output duplicates.)

Additional notes: Use the "on input . . . , [do some computation,] emit . . . " format for specifying your pseudocode. Make sure your that your algorithms balance are not unnecessarily inefficient.

c) In the lecture you learned about the cost measure *communication cost* for the analysis of Map-Reduce algorithms: The communication cost of a Map-Reduce algorithm is the sum of the input sizes of all tasks.

Answer the following questions / solve the following tasks.

i) Why is the measure called *communication* cost?

ii) Why doesn't the definition take the output sizes of the tasks into account?

iii) Sketch a scenario in which a Map-Reduce algorithm is inefficient despite having low communication cost. Explain, why the algorithm will perform badly.

Problem 6 (Streaming) 4+4+4+3+5 = 20 points

a) Consider the stream $\boldsymbol{a} = 7, 5, 5, 1, 4, 3, 8, 6, 6, 5$ over the universe $\mathbb{U} = \{1, \ldots, 9\}$. Compute the p-th frequency moment for $p = 0, 1, 2$, i.e. give $F_0(\boldsymbol{a})$, $F_1(\boldsymbol{a})$ and $F_2(\boldsymbol{a})$.

Hint: If it helps, you may count frequencies in the following table.

u	1	2	3	4	5	6	7	8	9
$f_u(\boldsymbol{a})$									

b) Apply the Tug-Of-War estimator to the stream of part a) using the hash function $h \colon \mathbb{U} \to \{-1, 1\}$ given below. Give a closed form expression for the result of the algorithm in terms of $h(a_i)$ (without plugging in the concrete values), then give the numeric result of the algorithm.

u	1	2	3	4	5	6	7	8	9
$h(u)$	1	−1	−1	−1	1	1	−1	1	1

(Colors only used to increase readibility.)

c) What are the requirements on the family of hash functions for the Tug-Of-War estimator? What guarantees on the output does the Tug-Of-War estimator give?

d) The Tug-Of-War estimator itself typically does not give very good results. What is the problem with the Tug-Of-War estimator and what does this mean for the output?

e) In the lecture we have discussed an extension of the Tug-Of-War estimator which avoids the problems discussed in part d). How does this extension work? What additional guarantee can we get for the extended version?