Started on	Sunday, 21 July 2024, 2:31 PM
State	Finished
Completed on	Sunday, 21 July 2024, 2:42 PM
Time taken	10 mins 11 secs
Marks	73/73
Grade	**10** out of 10 (**100**%)

Question 1

Correct

Mark 4 out of 4

The three main fields touched in this lecture are:

- **Data science** is about [gaining insights into data].
- **Machine learning** is about [making predictions from data].
- **Artificial Intelligence** is about [designing systems that exhibit intelligent behavior].

The relate as follows:

- [deep learning] is subfield of ML, which is a subfield of / technology used for [artificial intelligence]; and
- [data science] uses amongst others techniques from [machine learning], and data visualization.

[general artificial intelligence]

Your answer is correct.

Our focus will be on techniques at the intersection of data science and machine learning.

A note on AGI: Artificial general intelligence is currently rather far from current research and not part of this lecture; it uses all techniques from artificial intelligence combined with further insights into cognitive sciences (and many more, like psychology for goal derivation or philosophy for consciousness).

Correct

Marks for this submission: 4/4.

Question 2

Correct

Mark 8 out of 8

Data science uses techniques from ...

Select one or more:

- [ ] Artificial General Intelligence
- [x] Machine learning
- [ ] Requirements engineering
- [ ] Graphics design
- [x] Data analytics
- [x] Stochastic modelling This is a basic technique in machine learning.
- [x] Statistics
- [ ] Marketing

Your answer is correct.

Correct

Marks for this submission: 8/8.

Question 3

Correct

Mark 12 out of 12

In the following, typical data science and machine learning problems are given. Assign ML to those rather belonging to the field of machine learning, and DS to those rather closer to related to data science.

ML	DS		
◉	○	a. Predict the peak of flowering in the blooming woods of Altes Land from weather data, based on data from past year blooming times.	This may be at the boundary between ML (learn a model of the flowering periods) and DS (find a predictive model from data) again. What domain are the required techniques foremostly from? By the way, in case you are planning a visit: Check out https://www.bluetenbarometer.de/bluetenbarometer-altes-land in order to not miss the beatiful blossoming period.
○	◉	b. Identify chemicals that are relevant to the freshness taste of beer based on pairs of chemical analysis and taste data points.	This is a typical task of determining feature importance.
◉	○	c. Generate a recipe for Gin that tastes like Lübecker Marzipan based on recipe variations of the famous Lübecker KöniGin der Hanse drink.	This is classical generative ML.
○	◉	d. Visualize the tourism hot spots near Holstentor in Lübeck using visitor movement data.	Classic data visualization task for data introspection.
◉	○	e. Estimate the population development of Lübeck in the next two years based on past numbers of population, economic strength, and housing prices.	This predictive modelling task is at the boundary of ML and DS (learning a predictive model that generalizes to new---in this case future---data points). This can be modeled as a typical regression prediction task, potentially relying on stochastic modelling. Which fields do these techniques foremostly belong to?
○	◉	f. Estimate the most commonly ordered dish at Soul Sushi from past order data.	Don't get distracted by the "estimation" term: The main goal here is to identify certain properties (here: the mode) of a probability distribution.

Machine learning is about making predictions from data; whereas data science is foremostly about gaining insights into data. Both overlap, as ML techniques (e.g, unsupervised clustering) can also be used to gain insights into data, and modelling distributions in a way that allows to make predictions about future occurrences, as done in data science, is also a problem in machine learning. Thus, the boundary between the two fields is rather fuzzy.

Correct

Marks for this submission: 12/12.

Question 4

Correct

Mark 8 out of 8

Find the matches:

supervised learning	a label for every training sample
reinforcement learning	reward after some subsequent agent actions
semi-supervised learning	labels for some training samples
unsupervised learning	labels / reward for none of the training samples

Your answer is correct.

Correct

Marks for this submission: 8/8.

Question 5

Correct

Mark 4 out of 4

Comparability of values is an important feature of a domain, in particular because it provides a notion of value similarity to some degree. Often, such similarity should be maintained / respected by value mappings, like machine learning models. Hence, modeling the value representations in a way that allows to capture the comparability is crucial.
We saw some degrees of comparability in the lecture. Order them ascending by the richness of the comparability feature (e.g., not comparable < comparable):

1. nominal
2. ordinal
3. interval
4. relative interval

Your answer is correct.

Features of interest are:

- Does it allow to sort the values? (ordering)
- Does it allow to sort differences of values? (distance metric)
- Does it allow to set differences in relation to the absolute involved values? (zero point)

Correct

Marks for this submission: 4/4.

Question 6

Correct

Mark 8 out of 8

Which of the following properties describe instance-based learning (IBL) and which model-based learning (MBL)?

IBL	MBL		
◉	○	a. Allows **updates** on sample level without changing a model.	Updates on sample level here refers to adding or removing the influence of single training samples to inference.
◉	○	b. The learning is done **lazily**, i.e., no training phase is required prior to making a prediction on new data points.	
○	◉	c. Outputs are **hard to trace back** to specific training samples.	Note: This can pose issues in updatability or verifiability.
○	◉	d. Does not accurately memorize the data, but builds a **parameterized** model thereof.	
◉	○	e. All or a subset of the training data samples are **memorized**.	Memorized here refers to storing the samples unchanged in memory.
○	◉	f. Puts efforts into a **training phase preliminary to inference** for modeling the data.	
○	◉	g. Properties of the dataset that are relevant for the prediction are captured in a **compressed** manner.	Compressed here is in contrast to memorizing the samples directly and reconstructably.
◉	○	h. **High computational costs** can occur for inference; in particular does the inference cost depend on the number of training samples.	In the example algorithm we discussed, inference can

Correct

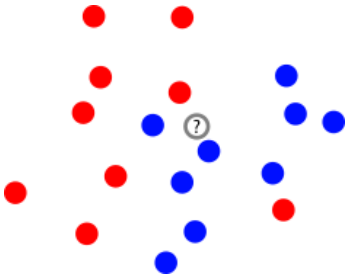
Marks for this submission: 8/8.

Question 7

Correct

Mark 5 out of 5

Remember the k-NN example data of 2D points with label values red or blue from the lecture. What color does the k-nearest neighbor algorithm predict for the white dot if k=18?



Select one:

- undefined This can, e.g., be caused by a tie in the majority vote.
- blue
- cannot be calculated
- red

Your answer is correct.

In this specific case, k-NN is used for a classification of categorical values. Hence, in order to make a decision, a majority vote amongst the 18 nearest neighbors is made (see the definition of the k-NN algorithm). What are the 18 nearest neighbors? What is the outcome of the majority vote? In case this can be calculated, the possible outcomes are:

- majority red = red
- majority blue = blue
- tie = undefined

Correct

Marks for this submission: 5/5.

Question 8

Correct

Mark 6 out of 6

Which are good positive indicators that a k-nearest neighbor algorithm might be appropriate?

Select one or more:

- little **computation** capacity **during inference** available
- ML model decisions must be **traceable** to training samples
- **real-time** capability needed
- little available storage or **memory**
- little **computation** capacity **during training** available
- ML model must be easily **updateable**

Your answer is correct.

Correct

Marks for this submission: 6/6.

Question 9

Correct

Mark 6 out of 6

What metric matches the following intuition:

metric intuitions

Metric	Intuition
Mean squared error (MSE)	How close were predicted continuous-valued outputs to expected ones?
Accuracy	What percentage of predicted classification outputs was correct?
Precision	What percentage of alarms (=positive predictions) by a binary classifier was correctly raised?
Recall	What percentage of positive class items was discovered by positive predictions of the classifier?
F1 Score	How good are precision and recall?
Area under precision-recall curve (AUC PR)	Can the binary classifier reach good trade-offs between precision and recall for respective choices of the decision threshold?

Your answer is correct.

Correct

Marks for this submission: 6/6.

Question 10

Correct

Mark 8 out of 8

The basic classification metrics are one of the fundamentals of evaluating machine learning systems. Since you will come across them (or at least the underlying ideas for evaluation) quite often when modelling, let's recap the concrete formulas again. Make sure to memorize them!

Fill in the formulas to match the definition of the respective metric, using

- TP / TN = true positive / true negative
- FP / FN = false positive / false negative
- P / N = ground-truth positive / negative class samples

To ensure uniqueness of the solution: For commutative operators, sort the entries according to occurrence in above list.

Note that P = (TP + FN) and N = (TN + FP).

- **Accuracy** = (TP + TN) / (P + N)
- **Precision** = TP / (TP + FP)
- **Recall** = TP / (TP + FN)
- **F_1 Score** = 2 * (Precision * Recall) / (Precision + Recall)
  = 2 * TP / (2* TP + FP + FN)

TP | TN | FP | FN | P | N

Your answer is correct.

Correct

Marks for this submission: 8/8.

Question 11

Correct

Mark 2 out of 2

Which of the following is symmetric with respect to the choice of which class is the positive and which the negative one? I.e., will the metric still produce the same outcome, if all labels are swapped from positive to negative and vice versa?

Select one or more:

- [ ] F_1 score
- [x] Accuracy
- [ ] Precision
- [ ] Recall

Your answer is correct.

Correct

Marks for this submission: 2/2.

Question 12

Correct

Mark 2 out of 2

What is the accuracy of the model f for the following pairs (y, f(x)) of ground-truth label y and model output f(x)?

Ground-truth vs. Prediction

Ground-truth label	Model output f(x)
true	false
true	true
false	true
false	false
false	true
false	*true**
true	true
true	false

Note: This classifier provides worse answers than random choice (assuming there are just as many false as true samples in the data). If one encounters something like this in practice, this is a hint that there might be a sign flip somewhere, or that the data is unbalanced (i.e., much more positive than negative samples or vice versa).

** ERRATA CORRECTION: There was an error in the originally provided dataset. This italic sample marked with * was corrected.*

Answer: 0.375

Correct

Marks for this submission: 2/2.

◄ Questions & Discussion

Jump to...

01. Quiz - Deep Neural Networks and Gradient Descent ►