model not enough complex
model do not include important
overfitting: small error on Dtrain big on Dtest.
model learns the data not problems
its model too complicated

Degree: CSA
Professor: Prof. Dr. Torsten Schön
Date: 05.07.2022
Exam time: 90 Minutes

Room number: 611
Seat number: 044

Allowed resources: non-programmable calculator

Task 1: General Understanding (~10%) 7P

a) Explain in your own words what underfitting and overfitting means
b) Why do we need to normalize numeric variables when using linear regression?
c) Give two reasons why you prefer cross-validation over holdout split
d) Name 4 things that you check on a classification dataset before you start using the dataset for model training
e) You are working with a dataset that has missing values in different features and instances. Explain how the MICE algorithm can be used to impute the missing values

Task 2: Fuzzy Logic (~20%) 18P

You are developing an automatic breaking system for a car using a fuzzy controller. Depending on the speed of the car and the distance to the car in front, the car should break more or less. The speed input is limited to a range of 0-130 km/h and the distance is limited to a range of 10-100 meters. The breaks can be applied with a force in the range of 0-4 kN.

a) Define the fuzzy subsets for the input and output variables using the following set of terms:
   speed: [slow, medium, fast]
   distance: [close, medium, far]
   breaks: [slightly, medium, harsh]
   Make sure that the terms are equally distributed over the ranges of the variables.
b) Calculate the force that is applied to the breaks if the speed is 110 km/h at and the distance is 35 meters where the following set of rules apply to the problem:

speed	distance	breaks
medium	close	Harsh
medium	medium	medium
medium	far	slightly
fast	close	harsh
fast	medium	harsh
fast	far	medium

c) Why are you preferring a fuzzy controller over Boolean logic for such a problem?

Task 3: Linear Regression (~20%)

You are faced with the following dataset and your task is to train a linear

X1	Y
2	6
4	10
-2	-4
-1	-1

a) Define the linear regression equation and visualize the data in a 2D diagram.
b) Initialize the weights with $w_0=0$ and $w_1=1$. Perform one iteration of Gradient Descent to update the weights. Use a learning rate of 0.01 and Root Mean Squared Error as loss function.

Task 4: True or False (~10%)

For the following statements, decide if it is true or not. If it is not true, give a reason why e.g. correct it:

a) In propositional logic, the syntax describes the way in which the truth of a sentence is determined
True F semantics

b) The discriminative approach in supervised learning is using probability estimations to predict the target variable.
False, because the discriminative approach divide data to predict target variable

c) kNN is an algorithm that can be used for classification and regression
True

d) If we have more than one feature in our data table as input to the regression, we call it Polynomial regression
False, we call it multinomial regression

e) For stochastic gradient descent with minibatch of 5, we evaluate the loss function on only 5 randomly picked data points in each iteration of the algorithm
True

f) As it is used for binary classification, the result of the cross-entropy loss can only be exactly 0 or 1

False, depending on value x and y, it can be between 0 and 1 ✓

Task 5: Random Forest (~20%)

Given the following dataset:

Mood	Age	Fur Color	Animal
Good	2	black	Dog
Bad	4	brown	Cat
Good	4	black	Dog
Good	12	black	Cat
Bad	2	brown	Cat
Good	4	brown	Dog
Good	4	Black	Dog
Bad	4	brown	Cat

a) Use the dataset to construct a decision tree for the given classification problem using Gini Impurity. Please make sure to provide all necessary calculations.

b) Use the tree to predict the class for the following unseen datapoint

Mood	Age	Fur Color
Good	14	brown

Task 6: Machine Learning Engineer (~20%)

You are working for a company that is selling toys. Frequently, new toys are added to the repository of the company. To get an estimate of how many samples of the new toys should be produced, your management wants to get a prediction of selling numbers for each toy. You plan to train a machine learning model that can predict the selling number of toys based on past experience from the stock repository.

a) What data do you collect to solve the given problem and how would you preprocess it?

b) Which model are you using for this problem and how do you evaluate its performance?

c) Draw a schematic view of the machine learning workflow used to solve this problem.

d) How close will the model be to the real selling numbers and how sure can you be about it?