TEST TOPIC DM Answers

2023 - Q1, test 4/10/23

Q1: [5 points, each item 1 point]

Indicate for the following applications whether it is a classification, regression, or clustering task:

		Classification	Regression	Clustering
a	Based on a set of data on the top 500 firms of the world on profit, number of employees, industry, and CEO salary someone wants to understand which factors affect CEO salary.			
b	Netflix provides you an advice on which series to watch based on your history and similar clients.			
c	Based on data from RNA measurements in urine from healthy controls and lung cancer patients a new test is developed to predict whether someone has a high probability of having lung cancer.			
d	Based on historical data on household energy use, a prediction is made for the coming winter.			
e	Detection of fractures on X-ray images based on a large database of X-ray images that were annotated by radiologists.			

Q2: [2 points]

Which model below cannot be used for classification?

a. Support vector machine
b. Logistic regression
c. Linear regression
d. Random Forest

Q3: [2 points]

What performance measure is most useful for a linear regression model?

a. Accuracy
b. Sensitivity
c. RMSE (Root mean squared error)
d. F1-measure

Q4: [2 points]

What is the right order for the following steps when developing a machine learning model:

1. Impute missing data
2. Train the model
3. Test the model
4. Split the data into a train and test set
5. Feature engineering
6. Feature selection

a. 5 – 6 – 1 – 4 – 2 – 3
b. 5 – 1 – 4 – 6 – 2 – 3
c. 4 – 5 – 1 – 6 – 2 – 3
d. 1 – 5 – 6 – 4 – 2 – 3

Q5:

A physician wants to predict the number of patients over 50 years that will develop COPD (Chronic Obstructive Pulmonary Disease) in his general practice for the coming years.

The following table summarizes data from his general practice which is available. The attributes are gender (male / female), education (low / high), and smoking (yes / no). The physician wants to use Data Mining techniques, and in particular decision trees to classify patients in the class “Yes COPD”, denoted by Y, and “No COPD”, denoted by N.

Attributes			Number of COPD	
Gender	Education	Smoking	Y	N
Male	Low	Yes	65	35
Male	Low	No	7	93
Male	High	Yes	38	62
Male	High	No	2	98
Female	Low	Yes	70	30
Female	Low	No	8	92
Female	High	Yes	35	65
Female	High	No	17	83
			242	558

a. [2 points] Assume a new patient is registered in the general practice but the physician has no other information at all about this patient. How will this new patient be classified based on the above data?
    a. COPD
    b. No COPD
    c. Indecisive, you should first gather more information

b. [2 points] What is the probability that a low-educated smoking male patient will be classified as developing COPD? Give your answer as a fraction with 3 decimals.

c. [2 points] What is the classification error for the attribute 'Smoking'? Give your answer as a fraction with 3 decimals.

d. [2 points] What will be the splitting attribute in the top root of the Decision Tree if one uses the classification error rate?

    a. Education, as its classification error rate is smaller than the classification error rates of Smoking and Gender.
    b. Education, as its classification error rate is larger than the classification error rates of Smoking and Gender.
    c. Smoking, as its classification error rate is smaller than the classification error rate of Gender and Education.
    d. Smoking, as its classification error rate is larger than the classification error rate of Gender and Education.

Q6:

The table below provides a training data set containing 7 observations, two predictors (attributes) X1 and X2, and a qualitative response variable Y (Y=Yes, N=No).

Observation	X1	X2	Y
1	25	30	Y
2	30	40	Y
3	18	20	N
4	35	25	Y
5	24	17	N
6	34	23	N

a. [3 points, subtract -1 for each error] Compute the Euclidian distance between each observation and the test point X1 = X2 = 25. Give your answer with 1 decimal.

Observation	X1	X2	Y	X1 - 25	X2 – 25	(X1 – 25)^2	(X2 – 25)^2	Sum	Distance
1	25	30	Y	0	5	0	25	25	
2	30	40	Y	5	15	25	225	250	
3	18	20	N	-7	-5	49	25	74	
4	35	25	Y	10	0	100	0	100	
5	24	17	N	-1	-8	1	64	65	
6	34	23	N	9	-2	81	4	85	

b. [3 points] What is the prediction for the test point X1 = X2 = 25 with K-nearest neighbours with K = 1 and K = 3?

	K = 1	K = 3
a	Y	Y
B	Y	N
C	N	Y
d	N	N

Q7: [4 points, each question 2 points]

In supervised learning we generally have observations on a quantitative response Y and p different predictors (features, attributes) $X_1, X_2, \ldots, X_p$. We assume that there is some relationship between the response and the predictors:

$Y = f(X_1, X_2, \ldots, X_p) + \varepsilon.$

The goal is to estimate the unknown function f from the observations.

Suppose the **unknown** function f is linear, what can you say about the **bias** and **variance** of a highly flexible estimator $\hat{f}$ of f?

	f is linear, $\hat{f}$ highly flexible estimator	
	Bias	Variance
a	Small	Small
b	High	High
c	Small	High
d	High	Small

Suppose the **unknown** function *f* is highly nonlinear, what can you see about the **bias** and **variance** of a highly flexible estimator $\hat{f}$ of *f*?

	f is highly nonlinear, $\hat{f}$ highly flexible estimator	
	Bias	Variance
a	Small	Small
b	High	High
c	Small	High
d	High	Small

Q8: [2points]

In order to obtain a model with a is developed and is trained on some training data. and some performance measures of the model are subsequently calculated on the test data. Which of the following might be a reason to think an error is made in the coding of the outcome compared to the predicted outcomes?

a. The sensitivity is below 0.5
b. The specificity is below 0.5
c. The accuracy is below 0.5
d. The positive predictive value is below 0.5