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.		**x**	
b	Netflix provides you an advice on which series to watch based on your history and similar clients.			**x**
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.	**x**		
d	Based on historical data on household energy use, a prediction is made for the coming winter.		**x**	
e	Detection of fractures on X-ray images based on a large database of X-ray images that were annotated by radiologists.	**x**		

a: Supervised learning problem; CEO salary, the outcome variable, is a continuous variable, so this is a regression problem.

b: This is an unsupervised learning problem that can be solved by clustering of series based on characteristics of clients

c: Supervised learning problem; outcome is having lung cancer or not; with the probability of having lung cancer you can classify someone as having lung cancer or not; Usually a classification model estimates the probability of being in a certain class.

d: Supervised learning problem; outcome variable is the household energy which is a continuous variable, so this is a regression problem.

e: Supervised learning problem; outcome is having a fracture or not, so a classification problem.

Q2: [2 points]

Which model below cannot be used for classification?

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

All mentioned models are supervised learning models. For linear regression, the outcome should be a continuous variable. Other models can be used for binary outcomes (to classify into two categories).

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

Accuracy is the same as the percentage of correctly classified cases. Sensitivity (which is the same as recall) is the fraction of true positives divided by the total number of positives, so a measure relevant for classification; F1-measure is a combination of the precision (the same as positive predictive value) and recall. Linear regression is a model where the aim is to minimize the mean squared error. Therefore RMSE is the most relevant outcome for linear regression. The other mentioned measures cannot be calculated for a linear model.

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

To prevent data leakage from the training set into the test set you should always start with splitting the data (4) and then do steps like feature selection, and imputing of missing data. Feature engineering, Imputing of missing data, and Feature selection (5-1-6) can only be applied before the train / test split (2-3) when you do not use any information from different rows of your dataset to impute or engineer a feature. E.g., to calculate a transformation of the data (such as a log transform) is allowed before splitting. But imputation with the mean of a column is not allowed as in that case information from the training set is leaked into the test set before splitting. Therefore the correct order is 4-5-1-6-2-3.

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

    P(Y) = 242 / 800, P(N) = 558 / 800; P(N) > P(Y) so classified as No COPD.

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.

In this case data of all combinations of the attributes is available. Therefore, there is no need to use the rule of Bayes, the calculation is:

$$\boldsymbol{P(Y\,|Male, Low, Smoking) = \frac{65}{(65+35)} = 0.650}$$

When using the rule of Bayes one would have:

$$\boldsymbol{P(Y\,|Male, Low, Smoking) = \frac{\frac{65}{242}*\frac{242}{800}}{\left(\frac{65}{242}*\frac{242}{800}+\frac{35}{558}*\frac{558}{800}\right)} = 0.650}$$

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

(192 + 34)/800 = 0.2825 [correct is 0.280 – 0.283]

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

Classification error rate Education: (150 + 92)/ 800 = 0.3025
Classification error rate Gender: (112 + 130) / 800 = 242 / 800 = 0.3025
So, smoking as an attribute gives the smallest 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	**5.0**
2	30	40	Y	5	15	25	225	250	**15.8**
3	18	20	N	-7	-5	49	25	74	**8.6**
4	35	25	Y	10	0	100	0	100	**10.0**
5	24	17	N	-1	-8	1	64	65	**8.1**
6	34	23	N	9	-2	81	4	85	**9.2**

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

The test point X1 = X2 = 25 is closest to observation 1, which has the label Y. The prediction is therefore Y.

For K = 3: Points 1, 3, and 5 are closest to the test point. Points 3 and 5 have the label N and point 1 has the label Y, therefore, the prediction for K = 3 is equal to 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

This is explained in the book, see Figure 2.12 page 36. In general, the more flexible a method, the variance will increase and the bias will decrease. A highly flexible estimator usually overfits an unknown linear function. The bias will be small then, but the variance will be high compared to the bias. When the unknown function is highly nonlinear, the more flexible the method will give a dramatic decline in bias and a small increase in variance, but much smaller compared to a very inflexible model.

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

When the accuracy (percentage of correctly classified cases) is smaller than 0.5 we could change all predicted labels the other way around and then the accuracy would be larger than 0.5 (is better). All other performance measures can be below 0.5 depending on the chosen cut-off level on the probability that is used to predict the labels (recall that usually the cut-off level on the probability is 0.5, but if you lower that, the sensitivity becomes larger, the specificity smaller simultaneously).