## Intro to Big Data Science — Spring 2023-2024

Name:	ID No.:
Quiz 2	

To receive credit, this worksheet MUST be handed in at the end of the class.

- 1. True or false:
  - 1) Decision tree is learned by minimizing information gain.
  - 2) No classifier can do better than a naive Bayes classifier if the distribution of the data is known.
- 2. You have trained a Naive Bayes classifier and plan to make predictions according to:
  - Predict y = 1 if  $h_{\theta}(x) \geqslant$  threshold
  - Predict y = 0 if  $h_{\theta}(x) < \text{threshold}$

For different values of the threshold parameters, you get different values of precision (P) and recall (R). Which of the following would be a reasonable way to pick the value to use for the threshold?

- (A) Measure precision (P) and recall (R) on the **test set** and choose the value of threshold which maximizes  $\frac{P+R}{2}$
- (B) Measure precision (P) and recall (R) on the **test set** and choose the value of threshold which maximizes  $2\frac{PR}{P+R}$
- (C) Measure precision (P) and recall (R) on the CV set and choose the value of threshold which maximizes  $\frac{P+R}{2}$
- (D) Measure precision (P) and recall (R) on the CV set and choose the value of threshold which maximizes  $2\frac{PR}{P+R}$
- 3. In which of the following circumstances is getting more training data likely to significantly help a learning algorithm's performance? (Select all correct choices.)
  - (A) Algorithm is suffering from high bias.
  - (B) Algorithm is suffering from high variance.
  - (C) CV error is much larger than training error.
  - (D) CV error is about the same as training error.
- 4. Which kinds of problems may the ordinary least square (OLS) suffer from ? (Select all correct choices.)
  - (A) Bad performance for nonlinear data
  - (B) Multicolinearity, thus resulting in the incorrect coefficients
  - (C) Underfitting for high dimensional problems
  - (D)  $(\mathbf{X}^T\mathbf{X})^{-1}$  may not be computed for high dimensional problems
- 5. Which is incorrect?
  - (A) Regularization is a process that adds a penalty term, which is usually the norm of model parameters, to the cost function.
  - (B) Regularization is to tradeoff between the training error and model complexity

- (C) In K-fold cross-validation, every sample could be used as training sample
- (D) K-fold cross-validation split the data into K subsets with different sizes.
- 6. Recall the kNN regression:  $\hat{f}(\mathbf{x}) = \frac{1}{k} \sum_{\mathbf{x}_{(i)} \in N_k(\mathbf{x})} y_{(i)}$ . Which of the following is correct?
  - (A) Small k may lead to large bias and small variance
  - (B) Large k may lead to small bias and large variance
  - (C) The model may be overfitted for too large value of k
  - (D) Appropriate selection of k by cross-validation could avoid overfitting
- 7. Consider fitting the linear regression model for the data

X	-1	0	2
У	1	-1	1

- (a) Fit  $y_i = w_0 + \epsilon_i$  (degenerated linear regression), find  $w_0$ .
- (b) Fit  $y_i = w_1 x_i + \epsilon_i$  (linear regression without constant term), find  $w_1$ .

8. Suppose you have the following training set with three boolean inputs x, y and z, and a boolean output U. Suppose you have to predict U using a naive Bayes classifier. Then after learning is complete what would be the predicted probability P(U=0|x=0,y=1,z=0)?

X	У	$\mathbf{z}$	U
1	0	0	0
0	1	1	0
0	0	1	0
1	0	0	1
0	0	1	1
0	1	0	1
1	1	0	1