Intro to Big Data Science — Spring 2021-2022

Name: 刘乐奇 **ID No.:** 12011327

Quiz 2

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

1. True or false:

True 1) Given m i.i.d. data points drawn from some distribution, then the training error converges to the true error as $m \to \infty$.

True 2) Both Manhattan distance and Jaccard distance satisfy the three properties: positive definiteness, symmetry, and triangle inequality.

False 3) Decision tree is learned by minimizing information gain.

False 4) No classifier can do better than a naive Bayes classifier if the distribution of the data is known.

D 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 $h_\theta(x) <$ 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 $2\frac{PR}{P+R}$

BD 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.

ABD 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

D 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. What is the difference between maximum likelihood (MLE) and maximum a posterior (MAP) approaches? Give some examples where these two approaches apply.

MLE thinks the parameter θ is fixed while the parameter θ in MAP obeys a distribution.

MLE can be used to find the parameters of linear model.

MAP can be used to derive the terms of reidge regression and LASSO regression.

7. Consider fitting the linear regression model for the data

x	-1	0	2
y	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.

(a) $RSS(w_0) = \sum_{i=1}^{3} (y_i - w_0)^2$

$$\frac{\partial RSS}{\partial w_0} = -2 \sum_{i=1}^{3} (y_i - w_0) = 0$$

$$\Rightarrow w_0 = \frac{1}{3} \sum_{i=1}^{3} y_i = \frac{1}{3}$$

(b) $RSS(w_1) = \sum_{i=1}^{3} (y_i - w_1 x_i)^2$

$$\frac{\partial RSS}{\partial w_1} = -2 \sum_{i=1}^{3} (y_i - w_1 x_i) x_i = 0$$

$$\Rightarrow w_1 = \frac{\sum_{i=1}^{3} x_i y_i}{\sum_{i=1}^{3} x_i^2} = \frac{1}{5}$$