ACT4311 Homework 2

Deadline: 2020-04-19 23:59:59

Q1. [6 points] How many binary classification models we need to train if we want to classify a training data with 5 different labels using One-versus-rest method?

D

A) 1
B) 2
C) 3
D) 5

Q2. [6 points] Among the three loss measures of classification problem, i.e., hinge loss, zero-one loss and Logistic loss, which one will punish those extreme errors (errors that are quite far away from the decision boundary) most severely?

C

A) Hinge loss
B) Zero-one loss
C) Logistic loss
D) All the same

Q3. [6 points] Which of the following approaches is not an approach for reducing/avoiding overfitting in machine learning?

D

A) Collecting more data
B) Pruning tree when adopting decision tree models
C) Feature selection
D) Adding nonlinear features

Q4. [6 points] If we try to fit a linear model with parameter *w* to training data by optimizing the following objective function, where *penalty(w)* is a penalty of model complexity (more complex model will have

higher penalty). Then in general, if the value of *C(C>0)* become smaller, the fitted model become:

B

$$\min_{\mathbf{w}} TrainlLoss(\mathbf{w}) + C \cdot penalty(\mathbf{w})$$

A) Less complex
B) More complex
C) The same
D) It depends

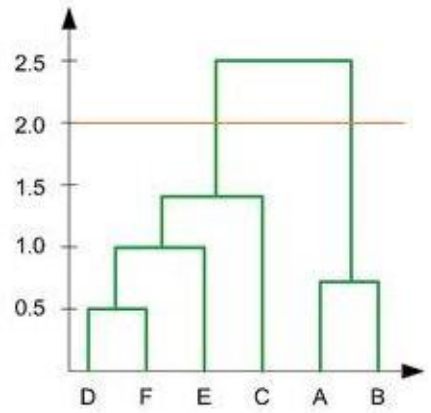
Q5. [6 points] In general, when we increase the complexity of adopted model, the model performance on training data will be (1), and the generalization performance of model will be (2).

A

A) 1-better, 2-worse
B) 1-worse, 2-better
C) 1-not sure, 2-better
D) 1-better, 2- not sure

Q6. [6 points] In the figure below, if you draw a horizontal line on y-axis for y=2, what will be the number of clusters formed?

B



A) 1
B) 2
C) 3
D) 4

Q7. [6 points] If we are allowed to use a chosen model as complex as we can, which of the following model can always get perfect zero-one loss (i.e., loss=0) among arbitrary training data without feature engineering (no duplication):

B

A) SVM
B) Decision tree
C) Logistic regression
D) Linear classifier with squared loss

Q8. [6 points] K-means clustering algorithms are usually used with the ___ distance measure for better convergence.

C

A) Manhattan
B) Cosine
C) Euclidean
D) All above

Q9. [6 points] Which of the following statement about fitting graph and learning curve is TRUE?

C

A) Fitting graph generalization performance vs. size of training set
B) Learning curve can help us to determine the optimal model complexity
C) Fitting graph can tell us when the overfitting happened
D) Learning curve is usually flat initially, and then become steep

Q10. [6 points] Which of the following statement about K-means model is False?

B

A) Clusters' distortions in a good measure to pick up the best model (that with minimum distortions) for different K-means models with fixed K
B) Clusters' distortions in a good measure to pick up the best value of K (that with minimum distortions) for K-means models
C) K-means model is sensitive to the initialization of centroids
D) K-means model is sensitive to outliers

Q11. [10 points] What does the number *K* in K-NN or K-means model determine in corresponding algorithm, respectively? And in general, how the model performance metric (i.e. clusters' distortion for K-means and classification accuracy on the training data for K-NN) will change (such as increase, decrease or other) when we increase the value of *K*?

The number of K in K-NN means the number of neighbors in the model, the predictions are based on these neighbors to be made. The K in K-means means the model tries to extract K clusters on the given datasets. In general, the performance metric will improve as K increases.

Q12. Given that we have six labelled training examples formatted as $((x_1,x_2),y)$, where (x_1,x_2) are 2-dimension input feature vector, and y is the target label. In detail, these 6 training examples are as follows:

((1,3),1), ((2,5),1), ((3,7),1), ((2,0),-1), ((3,1),-1), ((7,3),-1)

And now we have following three optional linear decision boundaries to classify above examples:

$$f_1(\mathbf{x}) = x_2 - x_1$$
$$f_2(\mathbf{x}) = x_2 - 2x_1$$
$$f_3(\mathbf{x}) = 2x_2 - 1x_1$$

Then please answer following questions:

Q12.1 [9 points] Please compute the total loss of classifier when using different loss function (zero-one loss, Logistic loss, and hinge loss) to evaluate above linear classifiers and fill the results (3 decimal digits) into the following table (please use natural logarithm (with base *e*) to compute Logistic loss):

	Zero-one loss	Logistic loss	Hinge loss
$f_1(x)$	0	0.466	0
$f_2(x)$	0	0.965	0
$f_3(x)$	0	0.761	0

Q12.2 [3 points] Now if we decide to select best classifier based on above three types of total loss (sum of individual loss) accordingly, which classifier should we choose (we can choose multiple best classifiers), respectively?

Loss function	Best classifier among $f_1(x)$, $f_2(x)$ or $f_3(x)$ (could choose multiple)
Zero-one loss	f1(x), f2(x) and f3(x)
Logistic loss	f1(x)
Hinge loss	f1(x), f2(x) and f3(x)

Q12.3 [3 points] Now we want to punish a linear classifier $f(\mathbf{x}) = w_1 x_1 + w2x_2$ **by the L2-norm of its weight, i.e., choose the best classifier by minimizing the total loss as well as penalty:**

$$\min_{\mathbf{w}} TrainlLoss(\mathbf{w}) + C \cdot penalty(\mathbf{w})$$

Where C=1, and $penalty(\mathbf{w}) = \sqrt{w_1^2 + w_2^2}$

Then based on the above regularity, which classifier should we choose (we can choose multiple best classifiers) based on above three types of total loss, respectively?

Loss function	Best classifier among $f_1(x)$, $f_2(x)$ or $f_3(x)$ (could choose multiple)
Zero-one loss	f1(x)
Logistic loss	f1(x)
Hinge loss	f1(x)

Q13. Now we have 5 (A to E) 2-dimensional points as follows:

A- (2,4), B-(5,8), C-(5,4), D-(10,16), E-(15,12)

Q13.1 [5 points] Please compute the pairwise Euclidean distance between the 5 points, and because dist(a,b) = dist(b,a), please just fill the upper triangle parts of the following table (distance matrix) accordingly?

	A	B	C	D	E
A	**0**	**dist(A,B)=?** 5	**dist(A,C)=?** 3	**dist(A,D)=?** 14.422	**dist(A,E)=?** 15.264
B	**Skipped**	**0**	**dist(B,C)=?** 4	**dist(B,D)=?** 9.434	**dist(B,E)=?** 10.77
C	**Skipped**	**Skipped**	**0**	**dist(C,D)=?** 13	**dist(C,E)=?** 12.806
D	**Skipped**	**Skipped**	**Skipped**	**0**	**dist(D,E)=?** 6.403
E	**Skipped**	**Skipped**	**Skipped**	**Skipped**	**0**

Q13.2 [3 points] If we want to groups above 5 points into two clusters using hierarchical clustering based on Euclidean distance, which two points will be grouped into one cluster first?

A and C

Q13.3 [7 points] Now we want to apply the hierarchical clustering algorithm to cluster the 5 points using <u>cosine distance</u>. Please plot the dendrogram of the clustering results. Please note that the linkage function among clusters is computed on the nearest points. (You may skip representing the distance between clusters in the dendrogram.)