Last Name: _____ First Name: _____ Student ID: _____

AIDI 1000: AI Algorithms and Mathematics – Final Exam Fall 2022

Due Date : December 16, 2022, 6:00 PM - 8:30 PM

- Part I: True/False and Reasoning Questions (2 points each)

1. TRUE or FALSE? K-means clustering algorithms can find clusters of arbitrary shape

No. K-means pre-defines k centers and calculate all the sample distances to its nearest center. Therefore, K-means will work better when Features are close to Gaussian Distribution. Not all kinds of cluster shapes.

2. TRUE or FALSE? A binary classifier having accuracy 0.8 is considered to be more useful than a binary classifier having accuracy 0.1.

For a classifier, the higher the accuracy score is, the better its performance will be on test data.

3. TRUE or FALSE? Both LDA and PCA are linear transformation techniques. The main difference is LDA is supervised whereas PCA is unsupervised.

Yes. PCA is for dimensionality reduction, while LDA finds linear combinations of features for multi-target classification. So LDA is a supervise learning algorithm.

4. TRUE or FALSE? One can copy a training set 10 times to form a larger training set in order to learn a better classifier.

Classifiers learns the patterns of training data by using the distribution of training set. Upsampling the whole training data set will not be helpful to improve the classification model. Generally speaking, it does not provide any new information to the model.

But it will work when training data target are unbalanced. Upsampling samples of the rare target will help balance the training data in order to prevent over-fitting.

5. TRUE or FALSE? PCA maximize the variance of the data, whereas LDA maximize the separation between different classes.

Yes. As the question describe. This is exactly what PCA and LDA did to the training data. And the purpose of these two algorithm is different. PCA is for dimensionality reduction. Top N singular values explains most of the information of raw data. LDA is for classification. It needs target to do supervise learning. And the result of it is to maximize the separation between different classes.

- Part II: Short Answer Questions (10 point each) (Attempt any 4 Questions)

1. (Dimensionality Reduction)

Consider PCA over 3D dataset (in R^3), which produces the first two principal components as $u^1 = (\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, 0)$ and $u_2 = (-\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, 0)$. If one of the original data points was x=(1,2,3), what will be its representation in the projected space?

y = [y1, y2]

y1=u1·x=$\frac{1}{\sqrt{3}}+\frac{2}{\sqrt{3}}+0=\frac{3}{\sqrt{3}}$

y2=u2·x=$\frac{-1}{\sqrt{3}}+\frac{2}{\sqrt{3}}+0=\frac{1}{\sqrt{3}}$

Therefore: y = $[\frac{3}{\sqrt{3}},\frac{1}{\sqrt{3}},0]$

2. (Hierarchical Clustering)

Based on Euclidean distance in R, we obtained following three clusters: A=(0,5,6),B=(3,9),C=(12). In the next iteration of the clustering, which two clusters will be merged by complete linkage and single linkage approaches, respectively?

- For complete linkage: (A, B) or (B, C) will be merged first because Distance(A,B)=9 is equal to Distance(B,C)=9. Distance(A,C)=12 is bigger than the others.
- For single linkage: A and B will be merged fist because C = (12) is obviously further than these to clusters.

3. (Regression)

Consider the covariance matrix $\Sigma=Cov\begin{bmatrix}X_1\\X_2\end{bmatrix}=\begin{bmatrix}9&-2\\-2&1\end{bmatrix}$

3.1 What are the standard deviations for X_1 and X_2?

- $\begin{cases}cov(x1,x1)=var(x1)=9\\cov(x2,x2)=var(x2)=1\\cov(x1,x2)=cov(x2,x1)=-2\end{cases}$

Therefore:

- stand_var(x1) =3
- stand_var(x2) =1

3.2 Assume that means of X_1 and X_2 are zero. Find the regression equation to predict X_1 as a function of X_2.

Because: mean(X1) = mean(X2) = 0

y=b+wX

4. (Logistic Regression)

Consider a Logistic Regression model with ReLU activation function, which has input $x\in R$ and a bias term, and the output value of $y=max(0,w_0+w_1x)$. What would be the input value x that produces output $y > 0$?

5. (K-Means Clustering)

Perform one iteration of k-means on the 1D dataset X ={2,4,7,9} with k=2 and initial centers at c1=0 and c2=10 using Manhattan distance.

5.1 Show initial cluster assignments

- The initial decision boundary is 5. When k=2, the result of 2 clusters are A={2,4} and B={7,9}

5.2 Show the new resulting cluster centers.

- For cluster A, the new center for next iteration is Center(A)=3

- For cluster B, the new center for next iteration is Center(B)=8

- Part III: Long Answer Questions

1. (Decision Tree and KNN) (25 points) Consider a training set provided below which contains two boolean features and a continuous feature.

	A	B	C	Class
row 1	F	T	115	-
row 2	T	F	890	-
row 3	T	T	257	+
row 4	F	F	509	+
row 5	T	T	733	+

1.1 How much information about the class would be gained by knowing whether or not the value of feature *C* is less than 400?

- For C<400, there are two samples, row1 and row3, in the subset.
- Therefore: $\mathrm{H}(\mathrm{Class}|\mathrm{C}<400) = \big(-0.5\times\log(0.5)\big)+(-0.5\times\log(0.5))\approx 0.301$

1.2 What is the information gain for feature A and B?

- For A:
  - H(Class) = -2/5*log(2/5)-3/5*log(3/5)=0.292
  - H (Class, A=T) = -1/3*log(1/3)-2/3*log(2/3)=0.276
  - H(Class, A=F) = -1/2*log(1/2) -1/2*log(1/2)=0.301
  - IG(Class, A, T) = 0.292-0.276=0.016
  - IG(Class, A, F) = 0.292-0.301=-0.009
- For B:
  - H(Class) = -2/5*log(2/5)-3/5*log(3/5)=0.292
  - H (Class, B=T) = -1/3*log(1/3)-2/3*log(2/3)=0.276
  - H(Class, B=F) = -1/2*log(1/2) -1/2*log(1/2)=0.301
  - IG(Class, B, T) = 0.292-0.276=0.016
  - IG(Class, B, F) = 0.292-0.301=-0.009

1.3 Is the above data set preprocessed well for kNN algorithms?

- Applying One-Hot encoding on feature A and B
- Applying normalization on feature C
- Applying standardization on feature C

As a conclusion, KNN uses feature values to calculate sample distance, so these preprocessing steps is helpful for kNN to perform well on training.

2. (KNN Classification) (25 points)

Given a dataset with binary labels $(x,y)=\{(2,+),(3,+),(5,-),(7,+),(11,-)\}$, compute

a. find the kNN training set error with k=5, using 0-1 loss. If there is a tie, always favor the positive class.

- 0-1 error: 0. This is a complete classification. Each sample for a cluster.

b. Similar to above but with *k*=4

- For k=4, distance(2,3) is the smallest. Then the clusters are A={(2,3),+}, B={5,-},C={7,+},D={11,-}
- For this clustering result. 0-1 error is still 0.

c. Similar to above but with *k*=3

- For k=3, then the clusters are A={(2,3),+}, B={(5,7),+}, C={11,-}
- 0-1 error for this clustering result is 1/3

d. Similar to above but with *k*=2

- For k=2, then the clusters are A={(2,3,5),+}, B={(7,11),+}
- 0-1 error for this clustering result is $\frac{\text{error(A)+error(B)}}{2} = \frac{\frac{1}{3}+\frac{1}{3}}{2} = \frac{1}{3}$

e. Similar to above but with *k*=1

- For k=2, then the clusters are A={(2,3,5,7,11),+}
- 0-1 error for this clustering result is 2/5

f. Should we choose the k with the smallest training set error? Why? The 0-1 Loss formulas is as follows:

- No, a complete splitting of samples always cause a zero training error.
- A clustering problem should also consider other factors for example silhouette coefficient.

$$\mathcal{L}_{0/1}(h) = \frac{1}{n}\sum_{i=1}^{n} \delta_{h(\mathbf{x}_i)\neq y_i}, \text{ where } \delta_{h(\mathbf{x}_i)\neq y_i} = \begin{cases} 1, & \text{if } h(\mathbf{x}_i) \neq y_i \\ 0, & \text{o.w.} \end{cases}$$