CSP 571 Data Preparation and Analysis

Final Exam

1) **A node contains 217 records of class C1 and 694 records of class C2 in a binary decision (classification) tree. What is the Gini score at this node that will be used for performing a split?**

A. 0.36

2) **When performing hierarchical agglomerative clustering, we obtain a cluster containing two points: P1=(50,48) and P2=(662,885) - we wish to merge in a third point P3=(56,58). At what dissimilarity value (using L1-norm distance) would the merge be performed, assuming single linkage?**

A. 16

3) **A node contains 969 records of class C1 and 857 records of class C2 in a binary decision (classification) tree. What is the information gain assuming a perfect split?**

A. 1.00

4) **The results from a K-Means clustering with K=2 produces the following 1-dimensional clusters: C1={71.23,58.15,91.55}; C2={40.83,71.57,22.46}. What is the L1-norm distance between the two clusters?**

A. 28.69

5) **We perform dimensionality reduction on an 837 x 76 data matrix via PCA using an eigenvalue decomposition - we obtain eigenvalues of (32,46,97,94). What is the proportion of variance explained if we choose the first 2 principal components? Please provide your answer in decimal form (e.g., 25% is entered as 0.25).**

A. 0.29

6) **Assuming a sufficiently large bootstrap of size 942 in an ensemble of decision trees, approximately how many total Out-of-Bag (OOB) estimates would we take an average of given an 9204 x 10 training/validation data matrix? Please submit your answer in an appropriately rounded whole-number value.**

A. 2,890,056

7) Given a sample data matrix of size 7034 x 1793, we perform a PCA for dimensionality reduction and select the top 58 principal components which explain 80% of the variance in the original data. We transform/project the original sample data into the reduced dimension space - what are the total number of values in the resulting data matrix?
A. 407,972

8) Given a multi-layer feed-forward neural network with fully-connected hidden layer L1 with 1574 hidden units and fully-connected hiddent layer L2 with 624 hidden units, what is the number of shared parameters stored in the weight matrix W2 for this system?
A. 982,800

9) When fitting a support vector machine (SVM) with a linear kernel, we allow for a soft-margin budget of 32 - given this constraint, what is the maximum percentage of slack variables that will have a value greater than 1, assuming an 202 x 95 training data matrix?
A. 0.16

10) When fitting a SVM with a polynomial kernel of degree 5 with 924 support vectors, what is the maximum number of dot-product operations that will be needed for classifying a new out-of-sample data point, assuming a 5864 x 25 training data matrix?
A. 924

11) Given a convolutional neural network that takes as input 3-channel color images of size 86 x 18, what is the size of the weight matrix W1 assuming the fully-connected input layer L1 has 419 units?
A. 1,947,093

12) When limiting the interaction depth to 80 in a boosted decision tree ensemble, what is the maximum number of variables that can be included in any tree assuming an 9255 x 575 training data matrix?
A. 80