COMS W4995 008 2017 3 Final Exam 12/07/2017

Instructions:

1. Write your name at the top of each page.
2. For multiple choice & true/false, clearly indicate your choice by circling or writing in your answer.
3. Unclear answers will be marked incorrect.
4. For all other questions, please indicate your answer in the space provided below the question.

---

1. We set aside a hold out set (after performing a train/test split) to provide an indication of:

   A. Model generalization to unseen data
   B. Overfitting on the training data
   C. A final performance metric for reporting
   D. All of the above

2. When evaluating a classifier that produces probablities, it is true that:

   A. We can increase precision by increasing our decision threshold
   B. We can increase recall by increasing our decision threshold
   C. Increase both precision and recall by increasing our decision threshold
   D. None of the above

3. Using a chi-squared measure for univariate feature selection tells us:

   A. Which combinations of features are predictive
   B. Which features are correlated with the target
   C. The performance of a single-feature model on the test set
   D. None of the above

4. Which of the following types of regression is best used for feature selection where we select non-zero coefficients in a linear model:

   A. Ridge (l2)
   B. LASSO (l1)
   C. ElasticNet (with mixture value set to 0.5)
   D. All of the above

5. When working in a time-series setting, if we wanted to change the sampling frequency from months to days, we could:

   A. Shift the dataset backward in time 29 days
   B. Pass a 30 day rolling mean window over the dataset
   C. Upsample the data and forward fill missing values
   D. None of the above

6. A set of ROC curves for different models can provide us with:

   A. An indication of model performance via Area Under the Curve (AUC)
   B. A comparison of model performances over different False Positive Rates
   C. A comparison of model performances over different True Postive Rates
   D. All of the above

7. When performing a permutation test to test a hypothesis, we permute over:

   A. The test-statistic being calculated
   B. The alpha level of the hypothesis test
   C. The assignment of observations to classes or groups
   D. None of the above

8. In order to calculate a p-value for a test statistic (generated from experimental data) using permutation test samples, we need to know:

   A. Number of permutation samples equal to 0
   B. Number of permutation samples equal to or more extreme than the observation
   C. Variance of underlying experimental distributions
   D. All of the above

9. In order to deal with imbalanced classes we can:

   A. Oversample the minority class
   B. Undersample the majority class
   C. Generate synthetic data for the minority class
   D. Any of the above

10. Which of the following are valid imputation techniques:

    A. Replace with the mean
    B. Replace with a randomly selected observed value
    C. Replace with a value seen in the prior record
    D. All of the above

11. Which is an example of using dummy variables?

    A. Adding a column to indicate where a feature is missing
    B. Adding a column to indicate the assignment of a category
    C. Adding a column to indicate where a value has been imputed
    D. All of the above

12. The defining characteristic between supervised and unsupervised learning is:

    A. Whether labels are provided
    B. Whether categorical features are allowed
    C. Whether the predicted value is categorical or numerical
    D. None of the above

13. We would use collaborative filtering to:

    A. Measure performance of a set of classifiers
    B. Select features based on several metrics
    C. Rank items based on item similarity
    D. None of the above

14. We use Grid Search to find the best performing:

    A. Model type and hyperparamater setting
    B. Training set
    C. Evaluation metric
    D. All of the above

15. When performing Cross Validation we are interested in finding, over folds on the training set:

    A. The best model performance
    B. The worst model performance
    C. The average model performance
    D. None of the above

16. Latent Dirichlet Allocation (LDA) provides which of the following after training:

    A. Cluster assignments for documents
    B. Labels for topics
    C. Per document topic distributions/mixtures
    D. All of the above

17. Multi-Armed Bandit algorithms provide an easy way to:

    A. Sample from unknown distributions
    B. Perform an experiment with more more than 2 populations
    C. Perform an experiment that can be stopped early
    D. All of the above

18. When using Hierarchical Agglomerative Clustering (HAC) we must define:

    A. The number of clusters
    B. The linkage method for determining which clusters to join
    C. A dendrogram defining the linkage structure
    D. All of the above

19. After training, a K-Means cluster model provides us with:

    A. Cluster assignments for the training set
    B. Ability to predict cluster assignments for new datapoints
    C. Locations of cluster centroids
    D. All of the above

20. We might use dimensionality reduction to:

    A. Plot high dimensional data in 2 or 3D
    B. Improve classification performance
    C. Calculate directions of highest variance in the dataset
    D. All of the above