COMS W4995 007 2018 3 Final Exam 12/10/2018

Instructions:

1. Please write your name at the top of each page.
2. Please check that your test has 40 questions.
3. Clearly indicate your choice by circling or writing in your answer. Unclear answers will be marked incorrect.

---

1. The defining difference 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

2. ElasticNet is a often used for regularization because it:

   A. Reduces the number of hyperparameters we need to tune
   B. Reduces the size of our Grid Search
   C. Allows us to tune a mix of Ridge and LASSO regression
   D. None of the above

3. When working with a time-series dataset, if we want 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

4. A regression model which always returns the mean of the training set targets will likely underfit on the training set due to high:

   A. Bias
   B. Variance
   C. Dimensionality
   D. All of the above

5. A gradient boost model is an ensemble of weak learners which learn by:

   A. iteratively removing features
   B. iteratively assigning additional weight to items based on error
   C. iteratively increasing the number of parameters in each model
   D. None of the above

6. We can use dimensionality reduction to:

   A. Plot high dimensional data in 2 or 3D
   B. Improve model performance
   C. Reduce the storage space required for a dataset
   D. All of the above

7. Increasing the depth of a decision tree increases:

   A. The bias of the model
   B. The number of questions the model can ask
   C. The likelihood of underfitting
   D. None of the above

8. A train/test split is done to create a held aside set in order to get an idea of:

   Note: A and B accepted

   A. Generalization of our model to new data
   B. Accuracy of the model during training
   C. If we can decrease training time of the model
   D. None of the above

9. A p-value tells you probability of:

   A. The null hypothesis being false
   B. The alternative hypothesis being true
   C. Observing the result, or a more extreme value, given that the null hypothesis is false
   D. None of the above

10. We shift a timeseries dataset in order to:

    A. Compare data at different points in time
    B. Reorder features (columns) in the dataset
    C. Change the frequency of observations
    D. None of the above

11. We can use a z-score transformation on a set of real-valued features to:

    A. Detect outliers
    B. Place features in the same scale
    C. Shift the features to have a mean of zero
    D. All of the above

12. Topic Modeling of documents allows us to:

    A. Compare similarity of documents in topic space instead of term space
    B. Reduce the dimensionality of dataset
    C. Determine what topics a corpus is composed of
    D. All of the above

13. When using R^2 as a performance metric in regression, we want to:

    A. Maximize the score
    B. Minimize the score
    C. Drive the score to zero
    D. None of the above

14. If our classes are {True,False}, in order to maximize Recall we can simply:

    A. Predict True for all items
    B. Randomly predict True or False for all items
    C. Predict True only when $P(\text{True} \mid x)$ is high
    D. None of the Above

15. We choose LASSO regularization for feature selection in a linear regression model because it attempts to:

    Note: A and C accepted

    A. Drive coefficients to a value near zero
    B. Drive coefficients below zero
    C. Drive coefficients to exactly zero
    D. None of the above

16. The purpose of tuning hyperparameters is to:

    A. Choose the features which are most predictive
    B. Find a setting that optimizes a performance metric
    C. Find an optimal train-test split
    D. None of the above

17. We perform feature selection in order to:

    A. Increase training and evaluation speed
    B. Remove noisy features
    C. Reduce model error
    D. All of the above

18. Mean Squared Error is a summarization of the difference between:

    Note: This should have specified "summarization of squared differences. B and D accepted.

    A. Predictions $\hat{y}$ and the mean of the training set targets
    B. Predictions $\hat{y}$ and targets y
    C. Predictions $\hat{y}$ and 0
    D. None of the above

19. When using regularization, for instance to keep coefficients small in a linear regression, we are attempting to:

    A. Reduce underfitting on the training set
    B. Reduce bias in the model
    C. Reduce overfitting on the training set
    D. None of the Above

20. An example of Unstructured data is:

    Note: this question was tossed as many people thought of images as being structured.

    A. a set of emails with header information
    B. a set of images
    C. a table of closing stock prices
    D. None of the above

21. Using a bag-of-words representation for documents removes context. We can retain some context by:

    A. **generating n-grams**
    B. removing stopwords
    C. tokenizing on whitespace
    D. None of the above

22. What are we avoiding when we use Cross Validation to tune parameters:

    A. evaluating too many settings of hyperparameters
    B. a long training time
    C. training and evaluating on the same items
    D. All of the above

23. If our classes are {True,False}, Precision refers to:

    A. The number of correctly predicted True out of all True
    B. The number of correctly predicted True out of everything we called True
    C. The number of correctly predicted False out of all False
    D. None of the above

24. Methods for avoiding overfitting a model include:

    Note: this question was tossed as too many people missed it

    A. Training and evaluating on different sets of items
    B. Using regularization to reduce model complexity
    C. Halting the change of a hyperparameter when a test set error exceeds training set error
    D. All of the above

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

    A. Hyperparamater setting
    B. Training set from several train test splits
    C. The number of folds to use in cross-validation
    D. All of the above

26. If we want to select the 3rd and 4th columns of all rows of a pandas dataframe X, we would call:

    Note: this question was tossed as too many people missed it

    A. X.loc[:,2:4]
    B. X.iloc[0:-1,3:4]
    C. X.iloc[:,2:5]
    D. None of the above

27. Latent Dirichlet Allocation (LDA) does NOT provide which of the following after training:

    A. Per topic word distributions
    B. Specific labels for topics
    C. Per document topic distributions
    D. None of the above

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

    Note: this question was tossed as too many people missed it

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

29. We often need to de-normalize a dataset pulled from a relational database using using JOINs because relational databases:

    A. only store structured data
    B. can't store data in columns
    C. store data in a way that reduces redundancy
    D. None of the above

30. When using a sliding window of size k over a timeseries dataset of size n, each datapoint will be used how many times?:

    Note: A and C accepted as step size was not specified

    A. 1
    B. n
    C. between 1 and k
    D. None of the above

31. We would use collaborative filtering to:

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

32. In our confusion matrix, a False Positive refers to an instance of:

    A. Predicting True when the target is True
    B. Predicting False when the target is True
    C. Predicting True when the target is False
    D. Predicting False when the target is False

33. A model with the flexibility to fit any training set very closely is said to have high:

    Note: this question was tossed as too many people missed it

    A. Bias
    B. Variance
    C. Error
    D. All of the Above

34. In general, when using Hierarchical Agglomerative Clustering (HAC) we must first define:

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

35. After plotting ROC curves to compare several models, the model closest to simply a random guess is the model with:

    A. An AUC close to 1
    B. An AUC close to 0.5
    C. An AUC close to 0
    D. None of the above

36. 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 centers
D. All of the above

37. Step-wise, as opposed to univariate feature selection, allows us to:

A. Evaluate different sets of features
B. Perform fewer tests
C. Use only some of our observations (rows)
D. All of the above

38. The first component of a PCA transformation gives us the direction of highest:

A. Error in the dataset
B. Number of features in the dataset
C. Variance in the dataset
D. None of the above

39. The key difference between clustering and classification/regression is that:

A. clustering is unsupervised
B. clustering uses distance between items
C. cluster assignment can be plotted
D. All of the above

40. We look for structure in our residual plot to indicate issues with our model or data. Residuals represent:

A. Difference between features in our dataset
B. Variance in the features
C. Differences between prediction and target
D. None of the above