Midterm III for STSCI 3740/5740

Nayel Bettache

Fall, 2024

- Read all the instructions carefully.
- They is one multiple choice question and two problems.
- This exam is on 34 points.
- This is a closed book exam. You cannot use any electronic devices (e.g., calculator, computer, cell phone) and the lecture notes.
- Please make your writing as clear as possible. If your writing is too hard to recognize, this will lead to reduced score.
- Throughout, we use the notation, models, estimators, etc. introduced in the lecture.
- Please write down your name and NetID on the first page. Good luck!

Time: **75 minutes.**

Name: _____ NetID: _____

Marks: _____

Problem 1 (20 points, 1 point for each subquestion)

Check the boxes next to the right answers. There can be **one to four** correct answers to each question. One point is assigned to a multiple choice question if and only if all boxes next to the correct answers to this question are checked and no box next to an incorrect answer to this question is checked.

1. What does the response value in a decision tree's leaf node represent?

   ☐ Median of responses
   ☐ Mode of responses
   ✓ Mean of responses
   ☐ Standard deviation of responses

2. What is the purpose of pruning a regression tree?

   ☐ Increase the tree depth
   ☐ Reduce the variance of predictions
   ✓ Prevent overfitting
   ☐ Remove outliers

3. What is a key disadvantage of decision trees compared to linear regression?

   ✓ Lower predictive accuracy
   ☐ Cannot handle qualitative variables
   ☐ Cannot be interpreted
   ☐ Requires dummy variables

4. Which method reduces the variance of predictions by averaging multiple models?

   ✓ Random forests
   ✓ Bagging
   ☐ Boosting
   ☐ LOOCV

5. In k-fold cross-validation, what does 'k' represent?

   ✓ The number of folds the data is split into
   ☐ The number of predictors in the model
   ☐ The number of trees in an ensemble
   ☐ The number of test data points

6. What does the shrinkage parameter λ control in ridge regression?

   ✓ The flexibility of the model
   ✓ The size of the coefficients
   ☐ The number of predictors selected
   ☐ The degree of the polynomial

7. What is the L1 penalty used in?

- ✓ Lasso regression
- ☐ Ridge regression
- ☐ Principal Component Regression
- ☐ Random forests

8. Which of the following methods explicitly performs variable selection?

   - ☐ Ridge regression
   - ✓ Lasso regression
   - ☐ Principal Component Regression
   - ☐ LOOCV

9. What does the shrinkage parameter in boosting control?

   - ☐ Number of trees
   - ✓ Learning rate
   - ☐ Tree depth
   - ☐ Number of predictors

10. What is the purpose of knots in regression splines?

    - ☐ Select the best model
    - ✓ Divide the predictor space into regions
    - ☐ Control model smoothness
    - ☐ Fit linear models to regions

11. How is the principal component direction determined in PCA?

    - ✓ It maximizes the variance of the data along that direction
    - ☐ It minimizes the residual sum of squares
    - ☐ It maximizes correlation with the response
    - ☐ It minimizes multicollinearity

12. How does cross-validation estimate test error?

    - ☐ By computing the error on the training data
    - ☐ By using bootstrapped samples
    - ✓ By holding out parts of the training data
    - ☐ By adding a penalty to the training error

13. Which of the following statements about bagging are true?

    - ☐ It reduces bias
    - ✓ It reduces variance
    - ☐ It improves interpretability
    - ✓ It uses bootstrapped samples

14. How does LOOCV differ from k-fold cross-validation?

    - ☐ LOOCV is computationally less expensive

- ✓ LOOCV uses the full dataset for training except for one observation
- ☐ LOOCV is less stable
- ☐ LOOCV always overestimates the error

15. How does LOOCV differ from k-fold cross-validation?

    - ☐ Uses fewer training samples
    - ☐ Uses a fixed split ratio
    - ✓ Uses one observation as a validation set in each iteration
    - ☐ Requires bootstrap samples

16. Why does random forests perform better than bagging in certain cases?

    - ☐ It uses more trees
    - ✓ It reduces correlation among trees
    - ☐ It uses pruned trees
    - ☐ It selects a single strong predictor at every split

17. What is the main advantage of Generalized Additive Models (GAMs)?

    - ☐ Models interactions between predictors
    - ✓ Allows nonlinear modeling while maintaining additivity
    - ☐ Improves interpretability over linear models
    - ☐ Incorporates polynomial regression

18. What does a high number of knots in a spline indicate?

    - ✓ Increased flexibility
    - ☐ Reduced overfitting
    - ☐ Improved interpretability
    - ☐ Higher bias

19. How is the optimal number of principal components in PCR determined?

    - ☐ By minimizing RSS
    - ☐ By using AIC
    - ✓ By cross-validation
    - ☐ By maximizing adjusted R^2

20. What distinguishes boosting from bagging?

    - ☐ Bagging uses bootstrapping
    - ✓ Boosting grows trees sequentially to learn residuals
    - ☐ Bagging reduces bias more effectively
    - ☐ Boosting uses random predictor selection

Problem 2: Understanding Random Forests (9 points)

Answer the following theoretical questions to demonstrate your understanding of random forests. Each question requires a detailed explanation.

1. **Structure of Random Forests**

   (a) Explain how random forests differ from a single decision tree.
   Random forests construct an ensemble of decision trees rather than relying on a single tree. Each tree is trained on a bootstrapped sample of the data, and predictions are aggregated. This aggregation reduces overfitting and improves the model's robustness and accuracy compared to a single decision tree, which may overfit the training data.

   (b) What role does bootstrapping play in the construction of random forests?
   Bootstrapping involves sampling the training data with replacement to create multiple datasets for training individual trees. This introduces diversity among the trees, which reduces variance when their predictions are aggregated.

   (c) At each split in a decision tree within a random forest, only a random subset of predictors is considered for the split. Why is this done, and how does it affect the performance of the model?
   Considering a random subset of predictors at each split (instead of all predictors) prevents strong predictors from dominating every tree's structure. This decorrelates the trees, making the ensemble more robust and reducing variance without significantly increasing bias. By decorrelating the trees, random forests ensure that the aggregated predictions are less prone to overfitting and capture more nuanced patterns in the data.

2. **Out-of-Bag Error**

   (a) Define out-of-bag (OOB) error in the context of random forests.
   OOB error is an estimate of the prediction error calculated using observations not included in the bootstrapped sample for a particular tree. These observations are referred to as out-of-bag because they were not used to train that tree.

   (b) How can OOB error be used to estimate the performance of a random forest without using a separate validation set?
   Since each observation serves as a test point for the trees that did not include it in their training sample, the OOB error is an unbiased estimate of the model's test error without needing a separate validation set.

3. **Feature Importance**

   (a) Describe how random forests compute the importance of each predictor.
   Random forests calculate feature importance by measuring how much each predictor reduces a specified metric (e.g., Gini impurity for classification or residual sum of squares for regression) when used for splitting. This reduction is averaged over all trees where the feature is used.

   (b) Explain one way to interpret feature importance values in practice.
   Features with higher importance scores contribute more to the model's predictions. In practice, these scores can guide feature selection by identifying which predictors are most influential.

4. **Advantages and Limitations**

(a) List two advantages of random forests over single decision trees.
1 - Reduced Overfitting: Aggregating predictions from multiple trees reduces the risk of overfitting to the training data.
2 - Robustness to Noise: By averaging across many models, random forests are less sensitive to noisy data or outliers compared to a single decision tree.

(b) Describe one scenario where random forests might not perform well and explain why.
Random forests may struggle with datasets containing a large number of irrelevant features or sparse data because splitting decisions and bootstrapped sampling could rely on noisy or uninformative predictors, reducing model efficiency.

Problem 3: Resampling Methods and Non-Linear Models (15 points)

Answer the following theoretical questions to demonstrate your understanding of resampling methods and non-linear modeling techniques. Provide detailed explanations for each question.

1. **Resampling Methods**

   (a) What is the purpose of resampling methods in statistical modeling?
   Resampling methods are used to estimate the accuracy of a model or a statistic by generating multiple samples from the data. They help evaluate model performance (e.g., test error) when a separate test set is unavailable and can aid in model selection and hyperparameter tuning.

   (b) Differentiate between the validation set approach and k-fold cross-validation.
   1 - Validation Set Approach: The data is split into a training set and a validation set. The model is trained on the training set and evaluated on the validation set. This approach is simple but can lead to high variance due to random splits.
   2 - k-Fold Cross-Validation: The data is split into k subsets (folds). The model is trained on k-1 folds and tested on the remaining fold, iterating through all folds. This reduces variance compared to the validation set approach.

   (c) Explain the concept of leave-one-out cross-validation (LOOCV). What are its advantages and disadvantages compared to k-fold cross-validation?
   LOOCV: In LOOCV, each observation is used as the validation set exactly once, while the remaining n-1 observations form the training set. The error is averaged over all iterations.
   Advantages: It uses the maximum amount of data for training, leading to a low-bias estimate.
   Disadvantages: It is computationally expensive and can have high variance since each training set is very similar.

2. **Bootstrap**

   (a) Describe the bootstrap method and its main use cases.
   The bootstrap is a resampling method where multiple datasets are created by sampling with replacement from the original data. It is used to estimate the variability of a statistic (e.g., mean, regression coefficients) and construct confidence intervals when theoretical approaches are infeasible.

   (b) How does the bootstrap estimate the variability of a statistic?
   By repeatedly calculating the statistic of interest on each bootstrap sample, the variability is estimated using the sample variance of these statistics.

(c) Why is it necessary to sample with replacement in the bootstrap method?
Sampling with replacement ensures variability among the bootstrap samples, allowing the method to mimic the process of generating new datasets from the population.

3. **Non-Linear Models**

(a) Explain why linear models may not always be sufficient for capturing relationships in data.
Linear models assume a linear relationship between predictors and the response. This assumption can fail for complex, real-world data where relationships are often non-linear, leading to poor model fit and biased predictions.

(b) List three examples of non-linear modeling techniques and briefly describe their key characteristics.
Polynomial Regression: Extends linear regression by adding polynomial terms of predictors, allowing for curved relationships.
Regression Splines: Divides the predictor space into intervals and fits piecewise polynomials, ensuring smoothness at interval boundaries (knots).
Generalized Additive Models (GAMs): Combines non-linear functions for individual predictors while maintaining additivity for interpretability.

(c) What is the main trade-off when using more flexible non-linear models? Discuss in terms of bias-variance trade-off.
Flexible non-linear models reduce bias by better fitting the data but can increase variance due to overfitting, especially when the model is overly complex or the dataset is small.

4. **Generalized Additive Models (GAMs)**

(a) How do Generalized Additive Models (GAMs) extend linear models?
GAMs replace the linear relationship between predictors and the response with smooth, non-linear functions for each predictor while maintaining additivity.

(b) What are the advantages of using GAMs compared to polynomial regression?
GAMs provide flexibility for each predictor independently, avoiding the instability of high-degree polynomial terms and allowing for better interpretability.

(c) Why do GAMs still maintain interpretability despite allowing non-linear relationships?
GAMs are additive, meaning the contribution of each predictor to the response can be visualized and interpreted independently.

5. **Local Regression**

(a) What is local regression, and how does it differ from global regression techniques?
Local regression fits a model to a small subset of data near the target point, whereas global regression fits a single model over the entire dataset. Local regression captures localized patterns.

(b) Explain the role of the bandwidth parameter in local regression. What happens if the bandwidth is too small or too large?
The bandwidth controls the size of the neighborhood used for fitting the model:
Too small: Overfits the data, leading to high variance.
Too large: Oversmooths the data, leading to high bias.

(c) What is the curse of dimensionality, and why does it pose a challenge for local regression when applied to datasets with many predictors?

The curse of dimensionality refers to the exponential increase in data sparsity as the number of dimensions increases. In local regression, finding enough nearby data points in high dimensions becomes difficult, reducing the model's effectiveness.