Review #1

DS 6030 | Fall 2024

review-1.pdf

Contents

1 Supervised Learning

1.1 HW 1

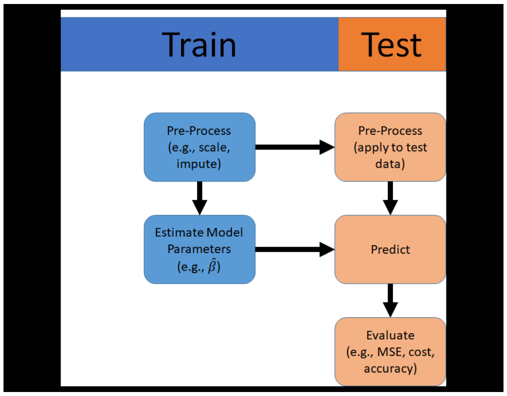
- The best predictive model is not always the true model.
  - Quadratic didn't always make best predictions. Why not?

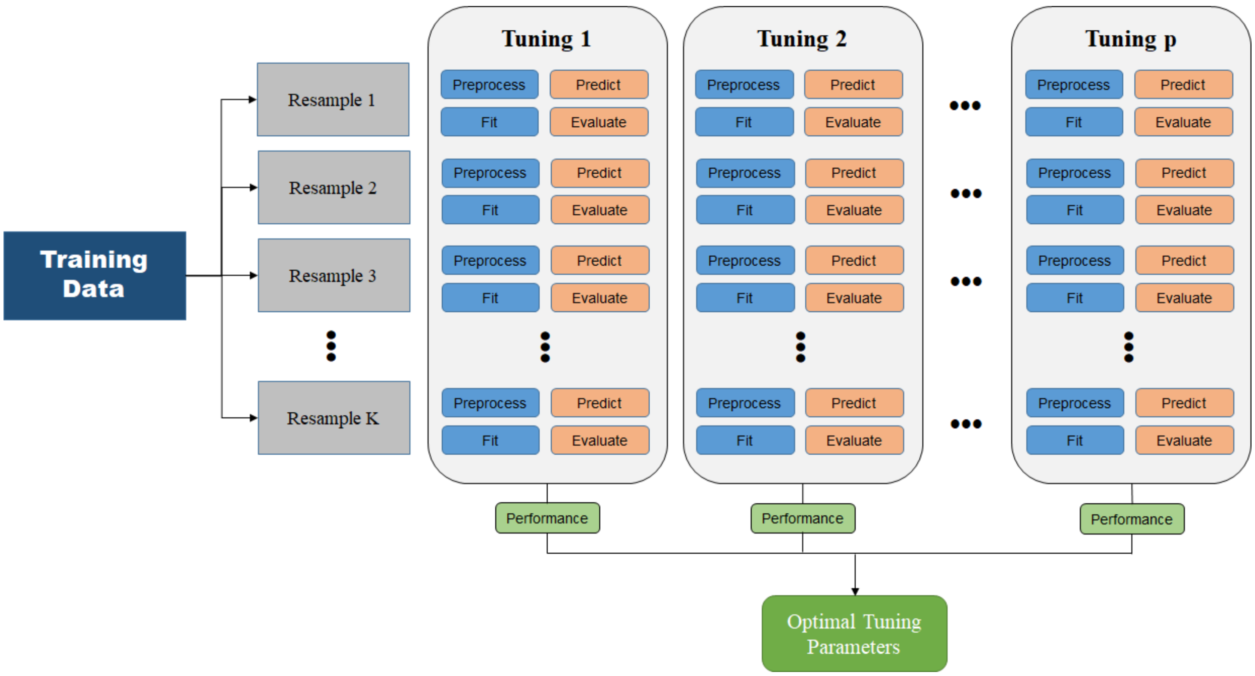
1.1.1 Questions

1. What is the Expected Prediction Error (EPE) (also known as Risk) and why do we care about it?
2. How is the EPE different from the training error (also known as empirical Risk)?
3. Under the squared error loss function, what is the optimal prediction? What about for the absolute error? Log loss?
4. How does model complexity/flexibility relate to the bias and variance of a predictive model?
5. What is overfitting? What is underfitting? How can they be prevented?
6. What are some ways to *increase* the complexity/flexibility of a predictive model? Ways to *decrease*?

2 Resampling

2.1 HW 2





2.2 Questions

1. In a train/test split, what proportion of observations should go in test set? Why?
2. What is the primary purpose of the bootstrap method?

3. How much training data does K-fold CV use to estimate the model parameters?
4. How does the bootstrap method simulate new data?
5. What is the expected proportion of observations that will not appear in a bootstrap sample (out-of-bag)?
6. How can out-of-bag samples be used in model evaluation?
7. What are the advantages of using the bootstrap over traditional methods like deriving confidence intervals from normal distribution assumptions?
8. Explain the bias-variance tradeoff in the context of bootstrap aggregating (bagging)? How does the bootstrap help in reducing variance?
9. What is cross-validation? What is it used for?
10. What is difference between k-fold and monte-carlo cross-validation? What are the advantages of each?
11. What is difference between OOB and cross-validation?
12. What is stratified cross-validation and why is it useful?
13. What is nested cross-validation and how does it compare to train-validate-test splits?
14. What is the optimal K in K-fold cross-validation?
15. In comparing predictive models using cross-validation, is it OK if each model uses a different cross-validation folds?

3 Penalized Regression

3.1 HW 3

3.2 Questions

1. Compare the lambda min and one-standard error rule in penalized regression.
2. What is one way to compare predictions of two models on a test set?
3. What is regularization (or penalization) in regression? What are some examples? What are the advantages of each example? How do you choose?

4 Tree-based methods

4.1 HW 4

4.2 Questions

1. Explain how CART (classification and regression trees) work?
2. In a classification tree, how are splits made?
3. In a classification tree, what are the predictions made in the leaf nodes?
4. How are trees similar to nearest neighbor models?
5. What are the tuning parameters in CART?
6. How does the OOB error work in Random Forest? How does the number of trees impact the uncertainty in this estimate? What is an advantage of OOB over cross-validation in RF?
7. Why do I not suggest tuning the number of trees in Random Forest?

5 SVM

1. How are SVMs similar to Logistic Regression?
2. What are the “kernels” in SVM?
3. What are “support vectors” in SVM?
4. What is the loss function used by SVMs? What is the penalty?
5. How does the output from SVM get converted to a probability? What are other ways?
6. Wny does probability calibration for SVM not expected to work well?
7. How does the Radial Basis Function (RBF) kernel work in SVM?
8. Suppose you have a large dataset with millions of features. How would you optimize SVM to handle this efficiently?
9. How would you choose the best kernel for your SVM model?
10. What are some advantages and disadvantages of using SVM compared to other classifiers like logistic regression or random forests?

6 Classification

6.1 HW 5

6.1.1 Contest Part 1 Results

6.1.2 Contest Part 2 Results

6.2 Questions

1. What is the logit?
2. What is the common loss function used in logistic regression?
3. Explain how logistic regression make probability outputs?
4. How can hard classifications be made in logistic regression?
5. How do you assess the performance of a logistic regression model?
6. What are some methods to handle class imbalance in logistic regression?
7. What is the maximum likelihood estimation, and how is it used in logistic regression?
8. What is the difference between accuracy, precision, recall, and F1 score
9. What is a confusion matrix, and how is it used in the evaluation of classification models?
10. Suppose your logistic regression model has high accuracy but poor recall. How would you improve it?
11. How would undersampling influence the predictive performance of a classification model?
12. Can ROC curves and AUC tell you which observations are predicted poorly?
13. How can you tell which types of observations are predicted poorly?
14. How should you choose the classification threshold if you have to make a hard decision?
15. Why do I say it may be unethical for a predictive model to make a hard classification?
16. How does class unbalance influence the quality of a predictive model? Which types of models are most impacted by class unbalance?

17. Should anything be done if there is class unbalance?