

Homework 3: Midterm Review

New Attempt

- Due Oct 29, 2024 by 6:10pm
- Points 22
- Submitting a file upload


Instructions: Please submit your answers as 2 files uploaded to Courseworks: a Jupyter Notebook (.ipynb) file & a pdf export. Please double check that all pages exported properly, sometimes they get cut off! In answering each of the following questions please include (a) the question as a markdown header in your Jupyter notebook, (b) the raw code that you used to generate any results, tables, or figures, and (c) the top ten or fewer rows of the dataframe (do not include more than ten rows for any table in your report). Include any plots or figures generated from your code as well.

For the following questions, use the spam dataset and variable descriptions located in the Homework 3 Data folder in this course's Files on Courseworks.

Part A:

1. Describe the importance of training and test data. Why do we separate data into these subsets?
2. What is k-fold cross validation and what do we use it for?
3. How is k-fold cross validation different from stratified k-fold cross validation?
4. Name the 4 types of supervised learning models that we have learned thus far that are used to predict *categorical* dependent variables like whether an email is labeled "spam" or "not spam."
5. Name the 3 types of supervised learning models that we have learned thus far that are used to predict *continuous* dependent variables like test scores.

Part B:

1. Import the spam dataset and print the first six rows.
2. Read through the documentation of the original dataset here: <http://archive.ics.uci.edu/ml/machine-learning-databases/spambase/spambase.names>  (<http://archive.ics.uci.edu/ml/machine-learning-databases/spambase/spambase.names>). The dependent variable is "spam" where one indicates that an email is spam and zero otherwise. Which three variables in the dataset do you think will be important predictors in a model of spam? Why?
3. Visualize the univariate distribution of each of the variables in the previous question.
4. Choose one model from Part A Question 4. Split the data into training and test subsets. Build a model with the three variables in the dataset that you think will be good predictors of "spam". Run the model and evaluate prediction error using k-fold cross-validation. Describe why you chose any particular parameters for your model (e.g., if you used KNN how did you decide to choose a specific value for k).
5. Repeat the previous question but with a *different* model from Part A Question 4.
6. Repeat the previous question but with a *different* model from Part A Question 4.
7. Repeat the previous question but with a *different* model from Part A Question 4.
8. Now rerun all 4 models with 3 additional variables that you think will help the prediction accuracy. Did this cause the performance to improve over your previous models?
9. What is a variable that *isn't* available in this dataset but you think *could* increase your final model's predictive power if you had it? Why do you think it would improve your model?

HW3 Rubric		
Criteria	Ratings	Pts
Part A		7 pts
Part B		15 pts
		Total Points: 22

