Recitation 5

Data, Inference and Applied Machine Learning

Friday 28 October 2022

Question 1

- Describe the four or more steps to implementing a rule based approach to decision making,
- Give an example of the rule based approach.
- State whether domain knowledge is required to establish a rule and give an explanation.

- Explain what is **overfitting** and why it is a problem in statistical learning.
- Choose between a simple model with one parameter or a complex model with ten parameters and support your choice.

Question 1 (continued)

- State and describe two commonly used approaches to avoid overfitting.
- Give two examples of metrics used to evaluate the performance of a model and give a formula for each one.
- Give two examples of applications and the appropriate metrics for each.
- Explain why benchmarks are useful in machine learning and give two examples of benchmarks

Question 2

- Explain what is machine learning
- Discuss the evolution of machine learning over time
- Explain why machine learning is popular

- Give three examples of machine learning techniques

- Explain the difference between classification and regression

Question 2 (continued)

- Explain the difference between supervised and unsupervised learning

- Give examples of successful applications of machine learning
- Explain the technique which is appropriate for each application
- State the type of learning that is involved

Question 3

- Load the diabetes dataset into MATLAB/Python
- Produce a correlation matrix of the explanatory (predictor) variables - MATLAB: corrcoef(), Python: corr()
- Make a heatmap of the matrix (using imagesc and colorbar for MATLAB, heatmap() for Python)
- Describe the relationships between the variables

- Explain what is collinearity
- Explain the effect that collinearity amongst predictor variables has on their estimated coefficient value
- Create a multivariate linear model using all ten variables and a constant - MATLAB: fitlm(), Python: fit()
- What are the mean squared error and adjusted R2 for model1 - Python: OLS()

Question 3 (continued)

- Compare significance values of the variables to the level of significance (alpha level) $\alpha = 0.05$ and determine whether all variables are significant
- Explain whether it is a problem of collinearity

- Explain the difference between forward selection and backward selection
- Explain how the stepwise approach works in selecting variables
- Compose a model using **forward selection** - MATLAB: stepwise(), Python: forward_regression ()
- State the variables that were selected
- Explain how the stepwise function works
- What are the mean squared error and R2 for the new model

Question 4

- Explain the difference between logistic regression and linear regression

- Load the titanic dataset into MATLAB/Python
- Calculate the probability of survival for a passenger on the titanic
- Provide a table giving survival probabilities broken down by passenger class, sex and age. Group the ages into classes and the interval width is not restricted.

- Build a logistic regression model with passenger class, sex and age as your explanatory variables and survived as the dependent variable- MATLAB: fitglm(), Python: LogisticRegression()
- Give the parameter estimates and compare them to the the level of significance (alpha level) $\alpha = 0.05$

- Use the confusion matrix to determine the classification accuracy- MATLAB: confusionmat(), Python: confusion_matrix()

Kaggle Titanic challenge

Extra credit: You are encouraged to enter the Kaggle challenge referencing this data set. At the end of this course, extra-credit will be given to students based on their final score on the challenge, coinciding with the deadline for the final assignment. Go to this link https://www.kaggle.com/c/titanic-gettingStarted and follow the instructions to register and enter the challenge.

Submission Files (MATLAB)

- **Single** MATLAB code file (.m) - **andrewID_DIAML_AssignmentNo.m**
- Assignment report(.pdf) - **andrewID_DIAML_AssignmentNo.pdf**
- Data files (as given)

Submission process:

- Put all data files and the source code in a **single folder** named with your **andrewID**
- Zip this folder and submit the zipped **(.zip)** with your report (**.pdf**)to CANVAS

Submission Files (Python)

- **Single** MATLAB code file (.m) - **andrewID_DIAML_AssignmentNo.ipynb**
- Assignment report(.pdf) - **andrewID_DIAML_AssignmentNo.pdf**
- Data files (as given)

Submission process:

- Put all data files and the source code in a **single folder** named with your **andrewID**
- Zip this folder and submit the zipped **(.zip)** with your report (**.pdf**)to CANVAS

Q&A