Recitation 6

Data, Inference and Applied Machine Learning

Friday 11 November 2022

Assignment Objectives

- Understand and deal with nonlinearity
- Fit classification models
- Choose optimal model parameters
- Perform cross validation on the given dataset
- Evaluate the performance of linear models

Question 1 - Nonlinearity

- Explain the necessity to consider nonlinear relationships between variables. To back up your explanation

- Mathematical equation for a nonlinear model and provide an example of an application where it might be appropriate.

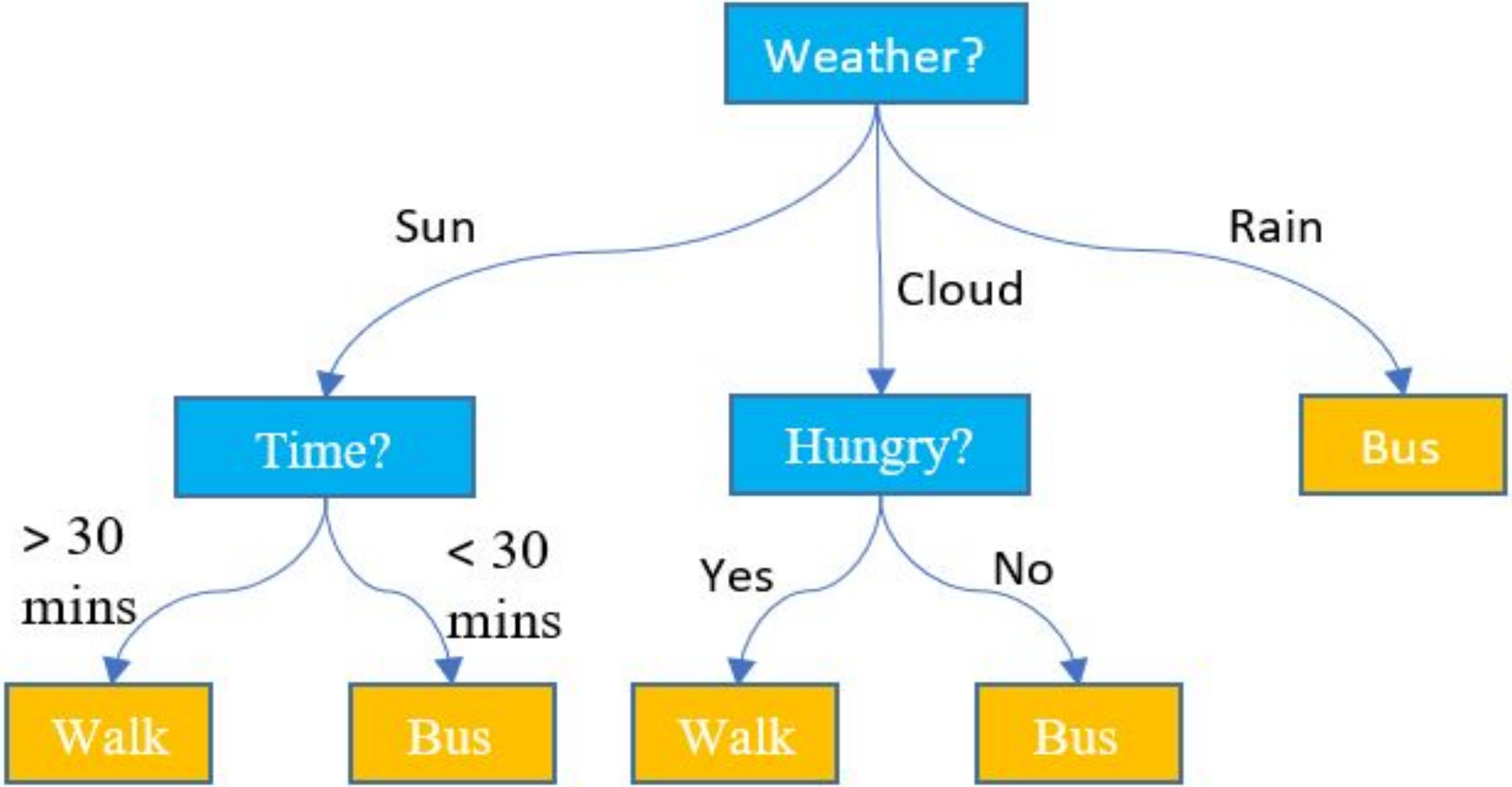
- Can a nonlinear model be more parsimonious than a linear model? Mathematical formulae for both the linear and nonlinear models to support your answer.

Question 1

- Surrogate data are used for testing for nonlinearity. What characteristics are typically preserved when generating surrogates? **Name** two surrogate techniques and **describe** the approaches for implementing them. —(hint refer to lecture slides)

- Define information, entropy and mutual information using mathematical formula.
  Describe how entropy can be used for constructing a feature for measuring regularity and **give an example of an application**.

  Explanation how mutual information can be used for feature selection and **why it might be better than correlation**.


Weather?
Sun
Cloud
Rain
Time?
Hungry?
Bus
> 30 mins
< 30 mins
Yes
No
Walk
Bus
Walk
Bus

Question 2 - Decision Trees

- Describe components of decision trees.
- Explain the conditions under which a tree might be pruned and why
- Give practical applications of decision trees and their explanations
- Highlight all the important steps involved in constructing a data-driven classifier
- Specify how you would validate the built classifier

Question 2 - Fitting the classifier

Data Preparation

- Load the titanic dataset and extract the necessary columns (age, sex, pclass and survived)
- Impute missing values using mean
- Transform categorical features (especially sex) to numerical

Question 2 - Fitting the classifier

Model fitting

- Fit a classification model (hint -- ***ClassificationTree*** in MATLAB, decisiontreeclassifier , tree from sklearn to plot in PYTHON)
- Plot the tree in graph mode (***view*** in MATLAB and ***tree*** from SKLEARN)
- Find misclassification errors of the tree using cross validation and in-sample techniques (hint -- *resubLoss*, *crossval*, *kfoldLoss in* MATLAB and *cross_validation,kfold in* PYTHON)

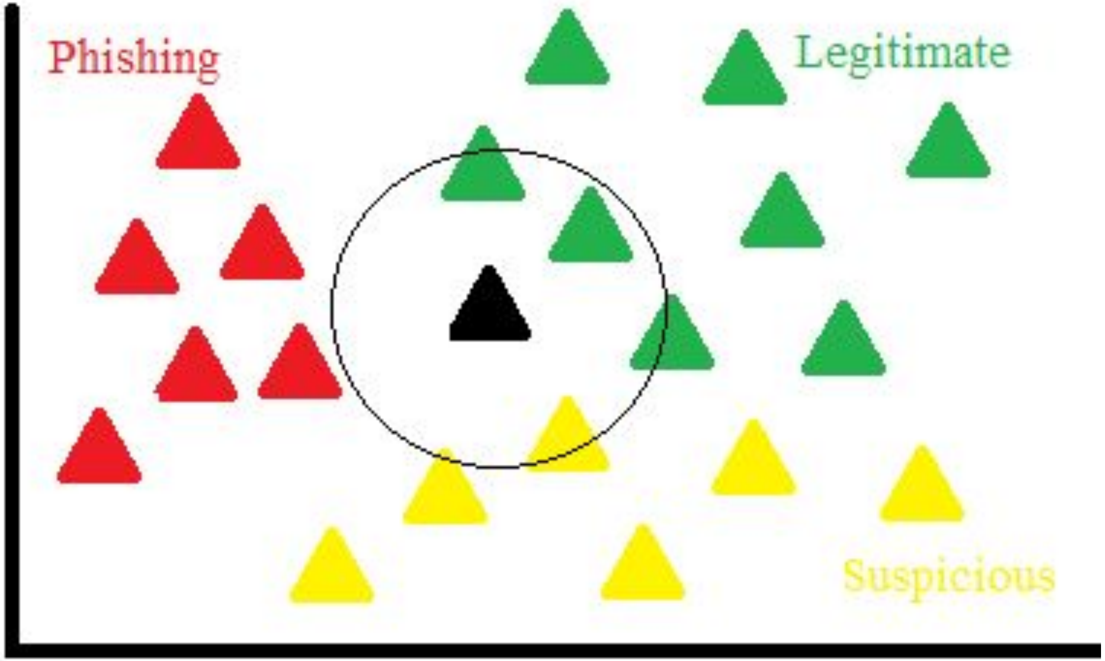
Question 2 - Cont.

Pruning

- Compute the best pruning level and prune the tree with it
- Compute the misclassification errors after pruning
- Compare and comment on your results

Logistics Regression

- Build a logistic regression model using cross validation techniques (i.e. with in-sample and out-of-sample dataset)
- Compute the error of your logistic regression model and **compare** it with the decision tree classifier


Phishing
Legitimate
Suspicious

Question 3 - KNN Classifier

Some theory

3.1 Constructing a parsimonious KNN classifier by focusing on a small neighborhood region. Hint -- read from the lecture slides or read online resources

3.2 Explain how will you transform the available variables in order to construct a KNN classifier

Question 3.3 - Fitting the KNN Classifier

3.3 Fit a KNN classifier with default params, and evaluate the performance using resubloss, cross validation and k-fold loss.

Fine Tuning the classifier - num_neighbours

i) Fit the classifier and test different neighborhood size (1-10, 15, 20)

ii) For each fit, compute the in-sample and cross validated loss. Keep track of these values.

iii) Identify the optimal number of neighbors using cross validation(Hint: k with minimum loss)

iii) Plot the graph of loss against the number of k. Also show the optimal point

Question 3

3.4 (i) Why are some distance metrics sensitive to the kind of features used? . Read from the lecture slides or online resources and provide explanation.

(ii) Evaluate the performance of the KNN classifier using different distance metrics.

[chebyshev, euclidean, mahalanobis, spearman, hamming, etc]

Compare best KNN with Logistic regression

3.5 Calculate performance of both classifier using cross-validation.

- Fit KNN wit optimal k
- Fit Logistic Regression

Provide advantages and disadvantage of both classifiers.

Based on the performance, which model is appropriate to be used for kaggle competition.

Question 4 - Wine Quality Regression

4.1 Calculate the average of each feature for the red and white wines separately using mean() function. Plot **bar graph** to show comparison. Infer on the results.

4.2 Calculate correlation of these features with the dependent variable and identify the most relevant feature based on the correlation values.

4.3 Use Lasso and cross validation to provide a plot of MSE for each wine type. Provide a plot of parameter estimates versus lambda. Hint - lasso() & lassoPlot() functions. Explain how the features were selected by LASSO.

Question 4 - Wine Quality Regression

4.4 Use the features identified by LASSO to construct a KNN regression model for the red wine.

4.5 Choose between linear regression model and KNN model, the model that performs better based on MSE and R2 values. Describe the advantages and disadvantages of both models.

Submission instructions

- Submissions should be made via Canvas.

- Single Python/MATLAB code file(.ipynb or .m) **[Do not Submit checkpoints for .ipynb]**. In addition, each line of code should be documented by text. This demonstrates that the code is unique and owned by the student

- Assignment report(.pdf) with full evidence that the assignment was completed by the student and demonstrate a full understanding of each step in the process including textual descriptions of each result (statistics, table, graph etc) represents and insights that can be gained

- Indicate the libraries you have used in your code at the beginning of the report (After the title page)

- Data files (as given)

Submission instructions

Submission process:

1. Put the source code **file and data files** in a single folder
2. Name of the folder should be the same as your Andrew ID
3. **Zip this folder and attach the zipped file on the assignment submission page (CANVAS)**
4. After attaching the zipped file, click on “Add Another File” from the assignment submission page and **attach your report**
5. Submit your assignment

N.B. This process will allow us to compile your reports in **Turnitin** to check for plagiarism.

Submission Process.......

Specific reasons for a submission being classified as incomplete include:

- Failure to correctly name your folder with your Andrew ID, report, and code file with andrewID_DIAML_AssignmentNo. For example, mcsharry_DIAML_Assignment1, mcsharry_DIAML_Assignment2 and mcsharry_DIAML_Assignment3.
- A missing report describing the steps, results, and insights
- A missing dataset required for running the code
- A missing code file such as .ipynb or .m file
- An error in the file path needed to run the code

Submission Instructions

The student is responsible for checking that their submission is complete. Students will lose 10% as for usual late submission even if the submission is repaired during the 24 hours after the deadline has passed and receive 0 for the assignment if it is not repaired.

The submission deadline is **on Monday 21st, November,2022 16:59 Eastern Time (ET) / Monday 21st, November, 2022 23:59 Rwandan Time (CAT) .**

Q&A