20200218【Data Analyst Nanodegree】P04M01L16

Part 04: Practical Modeling
Learn how to apply statistical statistics and probability to important, real-world scenarios, such as analyzing A/B tests and building supervised learning models.

- 20200218【Data Analyst Nanodegree】P04M01L16
- Module 01: Practical Stats
  - Lesson 16: Logistic Regression
    - 01. Video: Introduction
    - 02. Video: Fitting Logistic Regression
    - 03. Quiz: Logistic Regression Quick Check
    - 04. Video: Fitting Logistic Regression in Python
    - 05. Notebook + Quiz: Fitting Logistic Regression in Python
    - 06. Video: Interpreting Results - Part I
    - 07. Video: Interpreting Results - Part II
    - 08. Notebook + Quiz: Interpret Results
    - 09. Video: Model Diagnostics + Performance Metrics
    - 10. Confusion Matrices
    - 11. Confusion Matrix Practice 1
    - 12. Confusion Matrix Practice 2
    - 13. Filling in a Confusion Matrix
    - 14. Confusion Matrix: False Alarms
    - 15. Confusion Matrix for Eigenfaces
    - 16. How Many Schroeders
    - 17. How Many Schroeder Prediction
    - 18. Classifying Cheesy Correctly 1
    - 19. Classifying Cheesy Correctly 2
    - 20. Precision and Recall
    - 21. Powell Precision and Recall
    - 22. Bush Precision and Recall
    - 23. True Positives in Eigenfaces
    - 24. False Positives in Eigenfaces
    - 25. False Negatives in Eigenfaces
    - 26. Practicing TP, FP, FN with Rumsfeld
    - 27. Equation for Precision
    - 28. Equation for Recall
    - 29. Screencast: Model Diagnostics in Python - Part I
    - 30. Notebook + Quiz: Model Diagnostics
    - 31. Video: Final Thoughts On Shifting to Machine Learning
    - 32. Text: Recap
    - 33. Video: Congratulations
    - 34. Appendix: Glossary

Module 01: Practical Stats

Lesson 16: Logistic Regression

Learn to apply logistic regression models in python. Learn to interpret the results and understand if your model fits well.

01. Video: Introduction

In the previous lesson, you saw how we can use **multiple linear regression** to take multiple explanatory variables both **quantitative** and **categorical**, and use them all at the same time to predict a quantitative response variable. Using linear regression, we were optimizing for best predict a continuous variable.

Logistic regression is a regression approach used to predict only two possible outcomes. This can be predicting whether a particular credit card transaction is fraud or not, whether an individual will vote or not, whether an individual will default on a loan or not, or anything else you might be to predict with only two outcomes.

02. Video: Fitting Logistic Regression

Logistic regression is set up in a way where our linear model will predict the log odds instead of predicting the response itself.

This function here is called the sigmoid function takes the value from your linear regression line and maps it between zero and one as a probability.

Here, P is the probability of a one value occurring. So, this is a ratio of a probability of an event occurring as compared to the probability of the event not occurring. This is called the **odds ratio(胜算比)** and by taking the log and equating this to our linear regression line, we control our predictions to be between zero and one for the probability of a success.

03. Quiz: Logistic Regression Quick Check

Match each term to the description that holds true for it (linear and logistic Regression).

Description	Term
Predicts a probability between 0 and 1	Logistic Regression
Predicts any value between negative and positive infinity	Linear Regression
Explanatory variables can be either categorical or quantitative	Both

04. Video: Fitting Logistic Regression in Python

See: `01. Introduction.ipynb` / `Logistic Regression Starter.ipynb`

05. Notebook + Quiz: Fitting Logistic Regression in Python

See: `01. Notebook + Quiz: Fitting Logistic Regression in Python.ipynb`

06. Video: Interpreting Results - Part I

Interpreting Results:

The most important interpretation each of the coefficients:

For qualitative explanatory variables, any flip in any unit increase in your explanatory variables, we expect a multiplicative change in the odds of being in the one category to be the exponential of your coefficient holding all other variables constant.

07. Video (Screen Cast): Interpret Results – Part II

Interpret Results

- On weekdays, the chance of fraud is 70.5% times more likely than on weekends holding duration constant.
- For each minute less spent on the transaction, the chance of fraud is 0.02 times more likely holding the day of week constant.

08. Notebook + Quiz: Interpret Results

See: `01. Introduction.ipynb` / `Interpret Results.ipynb`

09. Video: Model Diagnostics + Performance Metrics

When determining how well your logistic regression model is doing or predicting the correct labels, the most common measure to use is called, **accuracy(准确率)**.

Accuracy is the proportion of correctly labeled rows divided by the total number of rows in the data set.

There are some cases where accuracy won't work well, particularly when you have large class imbalances in your data set.

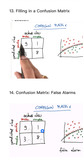
Resources:

Kaggle Machine Learning Courses is a great place to get started with the ideas that connect with these two lessons, as well as demonstrate some more advanced techniques.

The Machine Learning Flashcards is another great step following this program.

10. Confusion Matrices

Confusion Matrices(混淆矩阵):



11. Confusion Matrix Practice 1

Start Quiz:

12. Confusion Matrix Practice 2

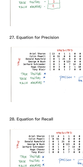
13. Filling in a Confusion Matrix

14. Confusion Matrix: False Alarms

15. Confusion Matrix for Eigenfaces

16. How Many Schroeders

17. How Many Schroeder Prediction



18. Classifying Cheesy Correctly 1

19. Classifying Cheesy Correctly 2

20. Precision and Recall

Precision and Recall

Recall(召回率/查全率): True Positive/(**真正例**) | True Positive + **False Negative(假反例)**: Out of all the items that are truly positive, how many were correctly classified as positive. Or simply, how many positive items were "recalled" from the dataset.

Precision(精确率/查准率): True Positive / (True Positive + **False Positive(假正例)**): Out of all the items labeled as positive, how many truly belong to the positive class.

21. Powell Precision and Recall

22. Bush Precision and Recall

- Note that the labels for the columns is the transpose of the rows (the first being "Actual Sharon", second being "Colin Powell" and so forth).

23. True Positives in Eigenfaces

24. False Positives in Eigenfaces

25. False Negatives in Eigenfaces

26. Practicing TP, FP, FN with Rumsfeld

27. Equation for Precision

28. Equation for Recall

29. Screencast: Model Diagnostics in Python – Part I

See: `01. Introduction.ipynb` / `Interpret Results.ipynb` / `Model Diagnostics.ipynb`

```
[0 1 0 1 0 1 
```