CS 660: Mathematical Foundations for Analytics

Dr. Francis Parisi

Pace University

Spring 2017

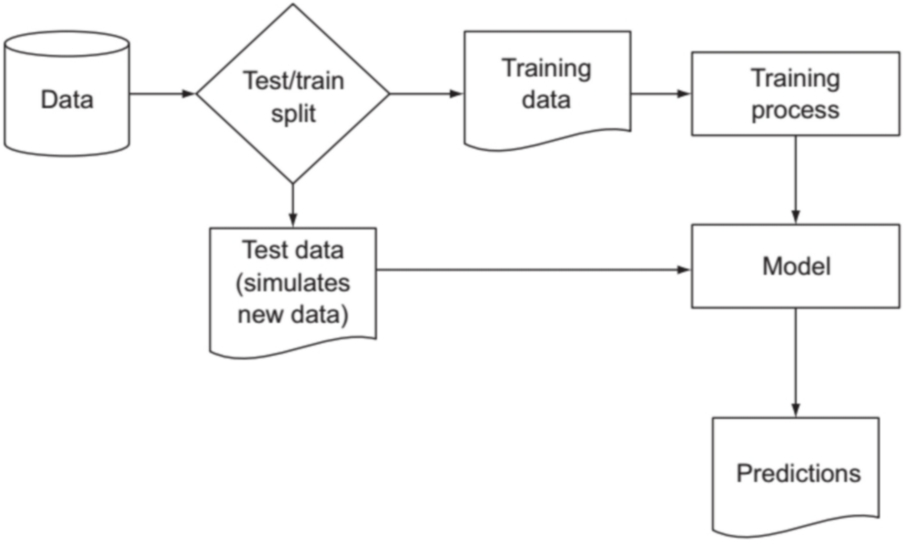
Course Overview

1. Part I: Data Science Fundamentals
    - Data Science Concepts and Process
    - The R Language
    - Exploratory Data Analysis
    - Cleaning & Manipulating Data
    - Presenting Results
2. Part II: Graphs & Statistical Methods
    - Basic Graphics
    - Advanced Graphics
    - Probability & Statistical Methods
3. Part III: Modeling Methods
    - Model Selection and Evaluation
    - Linear and Logistic Regression
    - Unsupervised Methods
    - Advanced Modeling Methods

Modeling Methods

- Match the problem to an analytical method
- Evaluate the model's quality –
  quantifying the performance of a model using a number of different measures
- Validate the model's soundness –
  checking that the model will work in the real world as well as it did during training

Modeling Methods



Modeling Methods

- Most problems in data science fall into one of two categories: *supervised* or *unsupervised*
- Supervised learning methods include:
  - linear regression
  - logistic regression
  - generalized additive models (GAM)
  - support vector machines (SVM)
- Unsupervised learning methods include:
  - K-means clustering
  - A priori algorithm for association rules
  - Nearest neighbor

Modeling Methods

- Use supervised methods when there is a known outcome
  - Classification – observations fall into two or distinct groups
  - Scoring – predicting the value of some measure based on other variables
- Use unsupervised methods when there are no known outcomes but you're looking for patterns and relationships in your data
  - Segmentation into an unknown number of groups (not pre-determined)
  - Make associations based on similar qualities or activities

Modeling Methods

- As a data scientist you must be able to map your problem to the most appropriate method
- Your intended uses of the model influence the methods you should use
- If you want to understand the relationship between input variables and outcome then a regression method would be a good choice
- If you want to know what single variable influences a categorization, then try a decision tree
- Classification is an example of *supervised* learning: in order to learn how to classify objects, you need a dataset of objects that have already been classified
- Cluster analysis and Association rules are examples of *unsupervised* learning: look for patterns in the data, not predict an outcome

Modeling Methods

- *Classification Using Nearest Neighbors* predict something about a data point p (like a customer's future purchases) based on data that are most similar to p
- *Classification Using Naive Bayes* use data about prior events to estimate the probability of future events (weather forecasting)
- *Classification Using Decision Trees and Rules* divide data into smaller and smaller portions to identify patterns that can be used for prediction
- *Regression Methods* used for forecasting numeric data and quantifying the size and strength of a relationship between the dependent and independent variables

    logistic regression can be used to model a binary dependent variable, and *Poisson regression* for count data

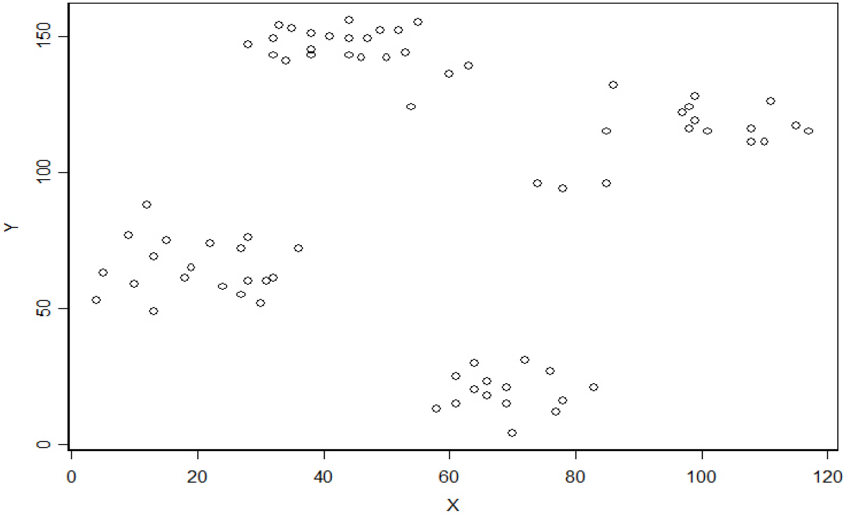
Modeling Methods

- *Neural Networks* model the relationship between input signals and an output signal using a model derived from our understanding of how a brain responds to stimuli from sensory inputs
- *Support Vector Machines* find a surface that makes a boundary between various points of data in multidimensional space according to their feature values
- *Association Rules* specify patterns of relationships among items such as

  $$\{\text{peanut butter, jelly}\} \Rightarrow \{\text{bread}\}$$

  if peanut butter and jelly are purchased, then bread is also likely to be purchased
- k*-means Clustering* is an unsupervised method that automatically divides the data into clusters, or group of similar items

Modeling Methods



Model Selection and Evaluation

Modeling Problem	**Typical Modeling Methods**
Classification: assigning known labels to objects	Decision trees Naive Bayes Logistic regression (with threshold) Support vector machines
Regression: predicting or forecasting numerical values	Linear regression Logistic regresssion
Association rules: finding objects that tend to appear in the data together	Apriori
Clustering: finding groups of objects that are more similar to each other than to objects in other groups	k-means
Nearest Neighbor: predicting a property of a datum based on the data that are most similar to it	Nearest neighbor

Model Selection and Evaluation

- After building a model we need to if the model even works on the data it was trained from
- We use a set of quantitative measures to assess model performance
- The measures we use vary by class of model
- To help us decide if our scores are “good enough” we compare our model to several ideal models

Model Selection and Evaluation

- A *null* model gives us the low end of performance
- A *Bayes rate* model gives us high end of performance
- A *single-variable* model tells us what a simple model can achieve

Model Selection and Evaluation

- Null Model
  - A null model is the best simple model you're trying to outperform
  - We use null models to lower-bound desired performance
  - We usually compare to a best null model
- Bayes Model
  - A Bayes rate model is a best possible model given the data
  - The Bayes rate model is the perfect model
  - The Bayes rate model is an upper bound on a model evaluation score
- Single-variable Model
  - Compare your model against the best single-variable model
  - A complicated model that doesn't outperform the best single-variable model can't be justified

Model Selection and Evaluation

Evaluating classification models

- The most common measure of classifier quality is *accuracy*
- The confusion matrix is a powerful tool for measuring classifier performance
- The confusion matrix is a table counting the number of known outcomes versus each prediction type

	Predicted	
Actual	Default	Non-default
Default	264	14
Non-default	22	158

Model Selection and Evaluation

Evaluating classification models

	Predicted	
Actual	TRUE	FALSE
TRUE	True Positive (TP)	False Negative (FN)
FALSE	False Positive (FP)	True Negative (TN)

Evaluating classification models	
Accuracy	number of items categorized correctly divided by the total number of items $$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Evaluating classification models

Precision	fraction of the items the classifier flags as being in the class actually are in the class

$$\text{Precision} = \frac{TP}{TP + FP}$$

Evaluating classification models

Recall	fraction of the things that are in the class are detected by the classifier; what fraction of class members are identified as positive $$\text{Recall} = \frac{TP}{TP + FN}$$
Sensitivity	Same as Recall; the true positive rate
Specificity	True negative rate – what fraction of class members are identified as negative $$\text{Specificity} = \frac{TN}{TN + FP}$$

Model Selection and Evaluation

Evaluating scoring models

- Main measure is residuals – the difference between the fitted value and the actual value for Y
- R^2 tells us how much of the variability in the dependent variable is explained by our model
- The significance of each variable is given the the corresponding p-values
- The overall model fit is measured by the F-statistic and p-value

Evaluating scoring models

```
# Make up some data
d <- data.frame(y=(1:10)^2,x=1:10,z=1:10
                  +rpois(10,lambda = 2))

# fit a regression model using the lm() function
model <- lm(y~x+z,data=d)
```

Model Selection and Evaluation

```
summary(model)

Call:
lm(formula = y ~ x + z, data = d)

Residuals:
Min      1Q Median      3Q     Max
-10.37  -4.66   0.51   2.19  12.48

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   -17.23        5.87   -2.94   0.0218 *
x              14.99        2.61    5.75   0.0007 ***
z              -3.76        2.33   -1.61   0.1509
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 7.42 on 7 degrees of freedom
Multiple R-squared:  0.963,Adjusted R-squared:  0.953
F-statistic: 92.1 on 2 and 7 DF,  p-value: 9.41e-06
```

References

James, G., Hastie, T., Witten, D. and Tibshirani, R. (2013).
An Introduction to Statistical Learning with Applications in R.
Springer, second edition.
6th Printing 2015.

Kabacoff, R. I. (2015).
R in Action.
Manning, Shelter Island, NY, second edition.

Lander, J. P. (2014).
R for Everyone.
Addison-Wesley, Upper Saddle River.

Zumel, N. and Mount, J. (2014).
Practical Data Science with R.
Manning, Shelter Island, NY, second edition.