Course Notes - Introduction to Analytics Modeling

Nolan MacDonald

Fall 2024

Contents

Chapter 1

Test

Chapter 2

Module 1 - Introduction

2.1 M1L1 - Course Overview

2.2 M1L2 - Course Structure

2.3 M1L2a - Homework Grading and Q&A

2.4 M1L3 - Modeling

Chapter 3

Module 2 - Classification

3.1 M2L1 - Introduction to Classification

3.2 M2L2 - Choosing a Classifier

3.3 M2L3 - Data Definitions

3.4 M2L4 - Support Vector Machines (SVMs)

3.5 M2L5 - SVM: What the Name Means

3.6 M2L6 - Advanced Support Vector Machines

3.7 M2L7 - Scaling and Standardization

3.8 M2L8 - K-Nearest-Neighbor (KNN) Algorithm

Chapter 4

Module 3 - Validation

4.1 Overview

Module 3 will continue with basic machine learning algorithms. The modules will cover couple of cross-cutting concepts and the important topic of model validation.

4.2 M3L1 - Introduction to Validation

Validation

- How good is the model?

Data has two types of patterns

- **Real Effect** - Real relationship between attributes and response
- **Random Effect** - Random, but looks like a real effect

Fitting matches both real and random effects

- Real effects - Same in all data sets
- Random effects - Different in all data sets

Example: What day of the month were you born?

- Training Data: 3, 21, 24, 24, 25, 26, 27, 30, 30, 31
- Best Predictor: You were born on the 26th
  - Right in the middle of 9/10 data points
- **This is a random effect!**
- This model using 9/10 from 21-31 doesn't have a large error
- If new data showed 2, 9, 11, 12, 14, 21, 24, 24, 29, 31
  - Much larger error due to the uniform spread over the month
- Was this just luck? (3, 21, 24, etc.)

  - No, some random pattern would have shown up
    - Early in month
    - Middle of month
    - Even/odd numbered day
    - Day is multiple of 3
    - Day is close to one of my kids birthdays
    - Etc.

M3L1 - Summary

- The example proves we can't measure the model's effectiveness on data it was trained on
- Model fit captures real and random effects
- Only real effects are duplicated in other data

Don't judge a model based on how well it fits the training data.

Key Points on Validation

- Validation is crucial to determine how good a model is and how accurately it performs on new data
- Measuring a model's performance on the same training data used to create it is not a good approach, as it will be too optimistic
- Any dataset contains both real effects (true relationships) and random effects (patterns that occur by chance)
- When fitting a model to training data, it captures both real and random effects
- However, when using the model on new data, only the real effects will persist, while the random effects will be different
- An example is given of a silly model that predicts people's birth dates based on a random pattern in the training data, which would not generalize well
- The key takeaway is that we cannot rely on training data performance to evaluate a model - we need a separate validation process to get an accurate assessment of its effectiveness

4.3 M3L2 - Validation and Test Data Sets

4.4 M3L3 - Splitting Data

4.5 M3L4 - Cross-Validation

Chapter 5

Module 4 - Clustering

5.1 Overview

Module 4 will continue with basic machine learning algorithms. The focus will be on clustering models. The modules will cover couple of cross-cutting concepts, including distance norms, and k-means clustering.

5.2 M4L1 - Introduction to Clustering

5.3 M4L2 - Distance Norms

5.4 M4L3 - K-Means Clustering

5.5 M4L4 - Practical Details for K-Means

Chapter 6

Appendix A: Glossary

6.1 Basic Machine Learning

Lessons 2.1-2.2, 2.4-2.6, 2.8, 4.1, 4.3-4.6, 6.1-6.3, 16.4

Algorithm: Step-by-step procedure designed to carry out a task.

Change detection: Identifying when a significant change has taken place in a process.

Classification: The separation of data into two or more categories, or (a point's classification) the category a data point is put into.

Classifier: A boundary that separates the data into two or more categories. Also (more generally) an algorithm that performs classification.

Cluster: A group of points identified as near/similar to each other.

Cluster center: In some clustering algorithms (like -means clustering), the central point (often the centroid) of a cluster of data points.

Clustering: Separation of data points into groups ("clusters") based on nearness/similarity to each other. A common form of unsupervised learning.

CUSUM: Change detection method that compares observed distribution mean with a threshold level of change. Short for "cumulative sum".

Deep learning: Neural network-type model with many hidden layers.

Dimension: A feature of the data points (for example, height or credit score). (Note that there is also a mathematical definition for this word.)

EM algorithm: Expectation-maximization algorithm.

Expectation-maximization algorithm (EM algorithm): General description of an algorithm with two steps (often iterated), one that finds the function for the expected likelihood of getting the response given current parameters, and one that finds new parameter values to maximize that probability.

Heuristic: Algorithm that is not guaranteed to find the absolute best (optimal) solution.

k-means algorithm: Clustering algorithm that defines clusters of data points, each corresponding to one of cluster centers selected by the algorithm.

k-Nearest-Neighbor (KNN): Classification algorithm that defines a data point's category as a function of the nearest data points to it.

Kernel: A type of function that computes the similarity between two inputs; thanks to what's (really!) sometimes known as the "kernel trick", nonlinear classifiers can be found almost as easily as linear ones.

Learning: Finding/discovering patterns (or rules) in data, often that can be applied to new data.

Machine: Apparatus that can do something; in "machine learning", it often refers to both an algorithm and the computer it's run on. (Fun fact: before computers were developed, the term "computers" referred to people who did calculations quickly in their heads or on paper!)

Margin: For a single point, the distance between the point and the classification boundary; for a set of points, the minimum distance between a point in the set and the classification boundary. Also called the separation.

Machine learning: Use of computer algorithms to learn and discover patterns or structure in data, without being programmed specifically for them.

Misclassified: Put into the wrong category by a classifier.

Neural network: A machine learning model that itself is modeled after the workings of neurons in the brain.

Supervised learning: Machine learning where the "correct" answer is known for each data point in the training set.

Support vector: In SVM models, the closest point to the classifier, among those in a category. (Note that there is a more-technical mathematical definition too.)

Support vector machine (SVM): Classification algorithm that uses a boundary to separate the data into two or more categories ("classes").

SVM: Support vector machine.

Unsupervised learning: Machine learning where the "correct" answer is not known for the data points in the training set.

Voronoi diagram: Graphical representation of splitting a plane with two or more special points into regions with one special point each, where each region's points are closest to that special point.