Lecture 2

New Contact Information

- email: cmangionetran@ucsd.edu
  - Feel free to send me questions, as you will see I integrate them into the class
- website for lecture notes
  - http://cseweb.ucsd.edu/classes/sp16/cse151-a/
  - I will post the lecture notes from lectures 1 and 2 immediately after class along with a syllabus
- We have added a piazza group for discussion
- TAs
  - Songbai Yang: yansongbai@eng.ucsd.edu
  - Yixing Lao: y1lao@ucsd.edu

Office Hours

- Tuesday, 9-10 AM
  - Will add more if necessary
  - Will post the TAs office hours
- CSE building, room 2272

The Book

- "Machine Learning. The Art and Science of Algorithms that Make Sense of Data", Peter Flach
- We will be skipping around
- It will be augmented by materials from many other sources
- Nice because it 'scales'
  - We will be skipping around quite a bit

Reading For The Next Week

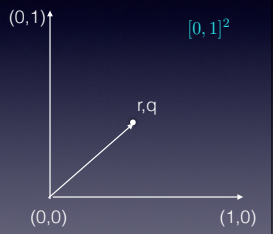
- Prologue
- Chapter 1, through geometric models
- I am working on probabilistic models as his marginal likelihoods seems dodgy
  - Our first implementation will be a geometric model

Mistake in Last Slides

- Pointed out to me after lecture
  - Bad copy and paste job
  - Great find, thanks
    - I have had professors make mistakes on slides and believed the strangest things
- However, never skip a teaching moment
  - Intrinsically it is crucial to understand

Visually

- If Y has no effect on X, then
  - P(X|Y) = P(X)
- Basic definition of 'conditional'



Sensitivity and Specificity

- These concepts are important, but not immediately obvious
- No test or model is perfect… there will be wrong predictions
- We need to consider two concepts
  - False Positives
  - False Negatives

False Positive

- A test result is positive
  - Or an observation is predicted to be positive
- What is the likelihood that the test result was actually negative
  - Hence false positive

$$FP_{rate} = \frac{FP}{FP + TP}$$

- FP_{rate} - Rate of false positives
- FP - Number of false positives
- TP - Number of true positives

False Negative

- A test result is negative
  - Or an observation is predicted to be negative
- What is the likelihood that the test result was actually positive
  - Hence false negative

$$FN_{rate} = \frac{FN}{FN + TN}$$

- FN_{rate} - Rate of false negatives
- FN - Number of false negatives
- TN - Number of true negatives

Sensitivity and Specificity

$$Sensitivity = 1 - FP$$

$$Specitivity = 1 - FN$$

Machine Learning Terms and Concepts

- A few terms will allow us to begin our conversation about models
  - Observations
    - Measurements taken from the real world
  - Features (or variables)
    - Components of the the observations
  - Tasks
    - Problems we wish to solve
  - Models
    - Output of a machine learning algorithm

Observations

- Observations define the machine learning space
- Examples
  - Stock prices over time
  - Gene sequences
  - Words in a tweet
  - Likes / dislikes on a dating site
- Note that these all have very different data in them
- How are these widely varying data sources feed ML algorithms

Hypothesis Space

- Given a set of observations the 'hypothesis space also called the model space
  - All possible conclusions that can be reached given the observations
- A model transforms of an observation to the hypothesis space

Features

- Features (or variables) are the individual components of the an observation
  - Health diagnostic observation may have temperature, body weight, heart rate, etc
- The feature space may be 'discrete' or 'continuous'
- Features must always be mapped into coherent (discrete or continuous) spaces
  - One of the goals for variable calculation

Types of modeling problems

- Output type determines the algorithms and type of model
  - Discrete
  - Continous

Discrete

- Output space is non continuous
  - Binary
    - Spam / not spam, Happy / sad tweets (sentiment analysis)
  - Multi class (may be very many classes)
    - Company names in a cleansing program, Fish species in a image classification program
  - Set of algorithms are called *classifiers*

Continuous

- Continuous - output space is real numbers
  - Satisfaction indexes (1-10), Mollusk age (your gonna see this one again)
- Called *regressions*

Types Of Learning

- Supervised
  - Each observation (exemplar) is labeled (tagged) with a category or value
- Unsupervised
  - There is no labeling, but we can still glean a lot of information
- Modern modeling techniques use a combination of both

Supervised Learning

- Types of problems
  - Whether you will like some content based on previous 'likes'
  - Virus detection from gene sequences
  - Image recognition (is it a fruit, vegetable, cake)
  - Fraud detection
  - Credit worthiness

Unsupervised Learning

- Types of problems
  - Association: 'Users also bought'
  - Market basket analysis
    - Complement
      - What items are likely to be in the same basket
      - Tomatoes and basil
    - Replacement
      - Likely hood of 'replacing' one item with another in the same basket
      - Lemons and limes