Week 1 – Getting ready:

Preparing data for machine learning

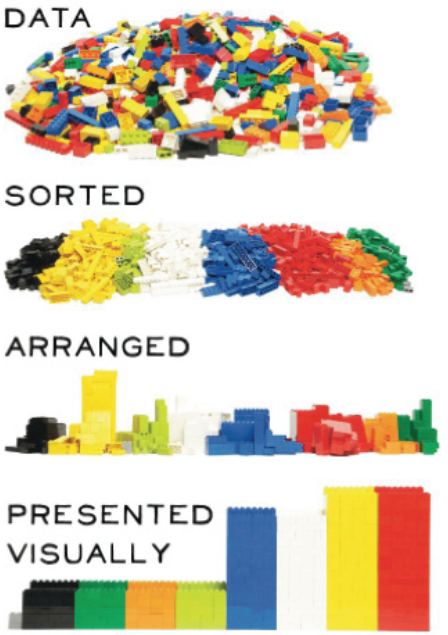
MSDS680 Machine Learning
SW1 2018
Nathan George

First, how was everyone’s break / last 8 weeks?

- Anything exciting?
- Do anything with data science, or make any plans to?

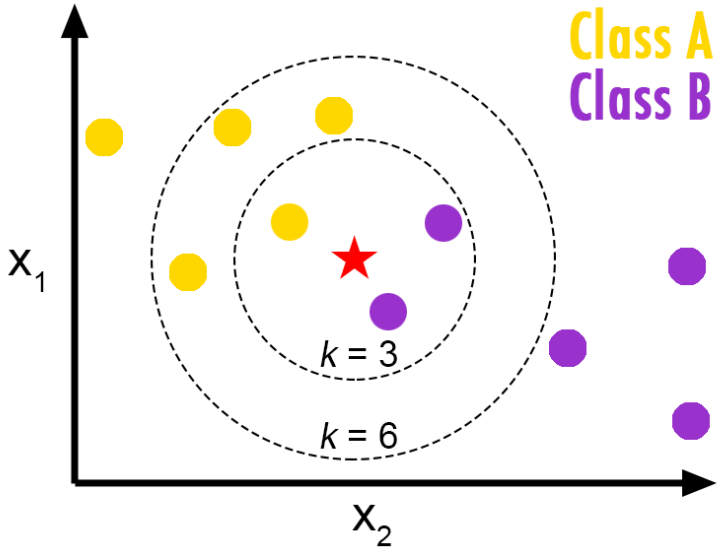
Course overview: week list

1. data preparation
2. K-nearest neighbors
3. Bayesian methods
4. Tree methods
5. Neural nets and SVMs
6. K-means and Hierarchical clustering
7. Performance Metrics
8. Reinforcement Learning



Course overview: week list

1. data preparation
2. K-nearest neighbors
3. Bayesian methods
4. Tree methods
5. Neural nets and SVMs
6. K-means and Hierarchical clustering
7. Performance Metrics
8. Reinforcement Learning



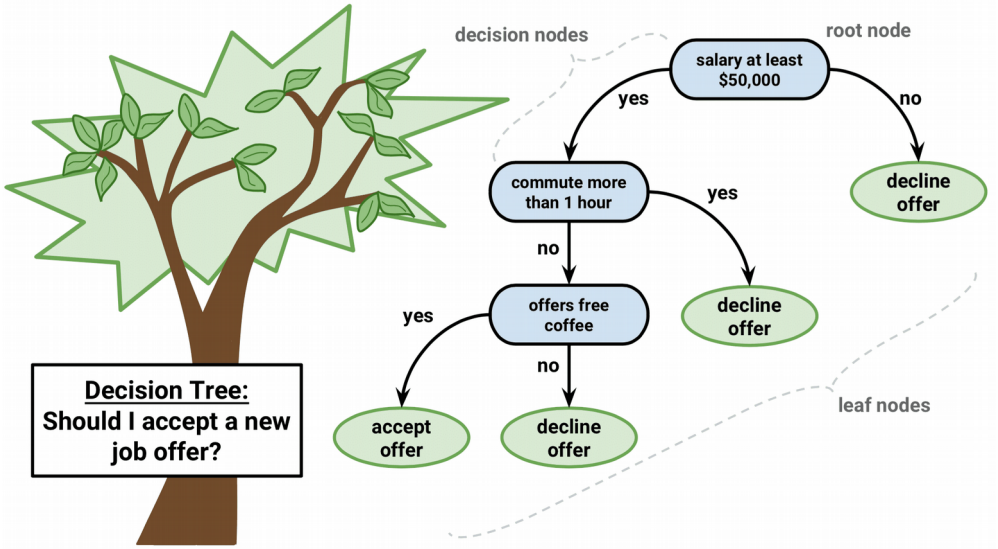
Course overview: week list

1. data preparation
2. K-nearest neighbors
3. Bayesian methods
4. Tree methods
5. Neural nets and SVMs
6. K-means and Hierarchical clustering
7. Performance Metrics
8. Reinforcement Learning

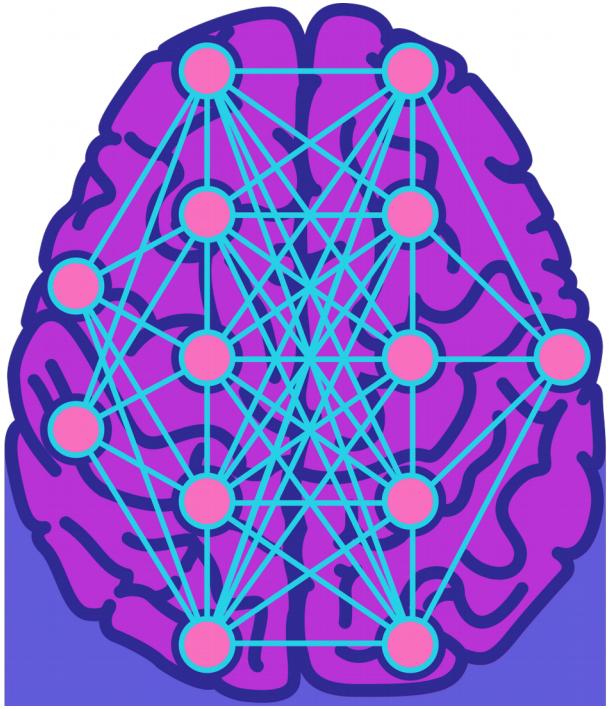
$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Course overview: week list

1. data preparation
2. K-nearest neighbors
3. Bayesian methods
4. Tree methods
5. Neural nets and SVMs
6. K-means and Hierarchical clustering
7. Performance Metrics
8. Reinforcement Learning



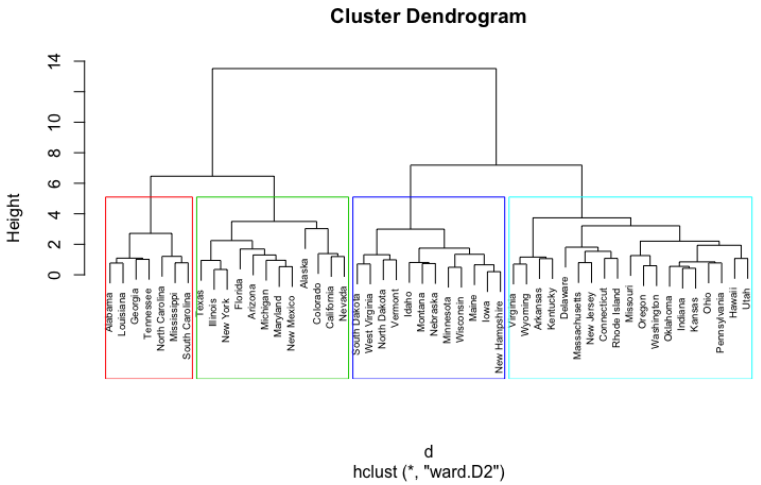
Course overview: week list



1. data preparation
2. K-nearest neighbors
3. Bayesian methods
4. Tree methods
5. Neural nets and SVMs
6. K-means and Hierarchical clustering
7. Performance Metrics
8. Reinforcement Learning

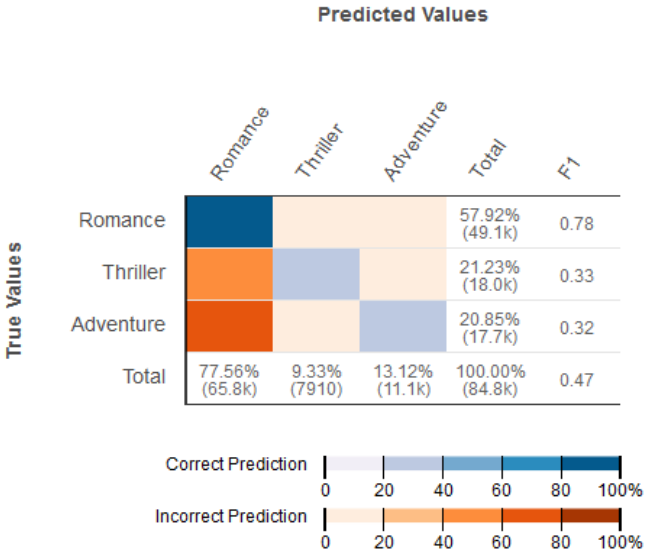
Course overview: week list

1. data preparation
2. K-nearest neighbors
3. Bayesian methods
4. Tree methods
5. Neural nets and SVMs
6. K-means and Hierarchical clustering
7. Performance Metrics
8. Reinforcement Learning



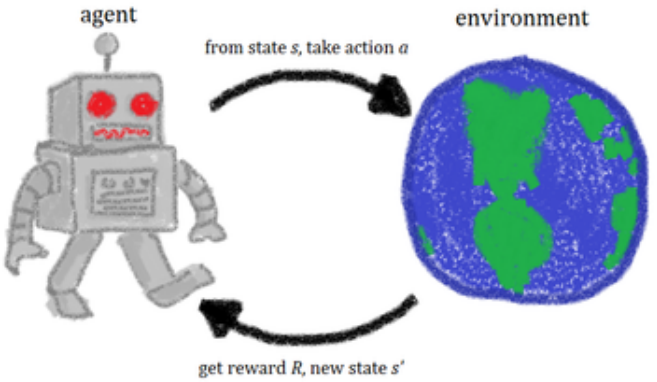
Course overview: week list

1. data preparation
2. K-nearest neighbors
3. Bayesian methods
4. Tree methods
5. Neural nets and SVMs
6. K-means and Hierarchical clustering
7. Performance Metrics
8. Reinforcement Learning



Course overview: week list

1. data preparation
2. K-nearest neighbors
3. Bayesian methods
4. Tree methods
5. Neural nets and SVMs
6. K-means and Hierarchical clustering
7. Performance Metrics
8. Reinforcement Learning



Essential data science skills

-
-
-
-

Essential data science skills

- Comprehending new data
- Programming (R and Python)
- Math and statistics
- Presentation skills!
  - Format for the course will be:
    - Review previous week / Introduce topics (30m presentation)
    - Demo to learn skills (30m)
    - Pair programming project (1h)
    - Post results, a few groups present (30m)
    - Individual programming skills quiz on last week’s topic (30m)

General pipeline for machine learning

- Define a (good) problem
- Collect data
- Clean data

- Transform data

General pipeline for machine learning

- Define a (good) problem
- Collect data
- Clean data
  - Deal with missing values – throwout, impute (median, mean, KNN)
  - Deal with outliers – throwout, clip

General pipeline for machine learning (continued)

- Transform data
  - Dummy variables (R will handle this automatically for factors, mostly)
    - Sometimes we may actually want a factor to be in sequential levels, like with education levels, so we have to order and convert the factor to an integer type
  - Scale data
    - Sometimes helps for SVM, usually necessary for neural networks and KNN
  - (optional) Reduce dimensionality
    - Principle component analysis (PCA)
    - Chi-square (usually for text applications)
    - TNSE (more often used for visualization)
    - Variance/correlation filters
    - Tree methods (e.g. random forest) feature selection

Good vs Bad Problems

- Characteristics of a good problem?

- Bad problem?

Good vs Bad Problems

- Characteristics of a good problem?
  - Objective target (e.g. has heart disease or not)
  - Enough data (at least 100s of points, preferably thousands or more)

- Bad problem?
  - Subjective target (e.g. is a review “overenthusiastic”)
  - Not much data/lots of missing data

Demo problem:

- Car mpg – predict mpg from characteristics of a car

https://archive.ics.uci.edu/ml/datasets/auto+mpg

The problem we will start with: predicting heart disease

- UCI Riverside Heart Disease dataset
- We will use the heart.disease.data file, which has 14 attributes and 1 target variable
- Objectives:
  - Impute missing data (don't just throw it away)
    - -9 means missing
  - Look at the data with EDA
  - Throw away columns if needed
  - Deal with outliers (maybe make a boxplot)

```
remove_outliers <- function(x) clamp(x, lower = boxplot.stats(x)$stats[1], upper = boxplot.stats(x)$stats[5])
```

https://stackoverflow.com/a/27109891/4549682

https://stackoverflow.com/a/4700136/4549682

http://archive.ics.uci.edu/ml/datasets/heart+Disease

Next week

- Next week we will use this data prep to start, then use KNN to predict heart disease from the data.