Update anything and share with people in 474
Student Project Presentations:

Group Number and Project	Algorithm used	Notes
9 - Price prediction of used cars	Linear Regression	Kijiji -DM-> their interface -> $ price of car SQL db Clean data - remove noise - add 0's
15 - Github issue label prediction	Text classification Clustering	5 main types of issues
4 - Violence in the News	Random forest	Categorize appropriate for kids Vectorizors Tried: bayes, trees, regression
1 - NBA Expected shots GitHub repo: https://github.com/erikreppel/seng474-nba-shots Presentation slides: https://goo.gl/WVwciJ	- Multilayer perceptron (final decision) - Linear regression, naive bayes (bernoulli and gaussian), AdaBoost all tested	- Classification > regression - Calculating expected value of shot, i.e. 75% chance to hit a 3 pointer, 0.75 * 3 = 2.25 expected value.
20 - NHL Points Projection	Clustering Euclidean distance Neural network	Age regression - in prime? Accuracy - 91.6% - 43.4% Good for average - bad for outliers Neural net was having major problems giving stars a 0 prediction
12 - Predictive Policing (Crime Prediction)	Lots used: Gaussian NB, decision tree, multinomial NB, bernoulli np, perceptron, logistic regression, SVM	Government data site: from Victoria - 2006-2016 categorizing
13 - MLB MVP	Multiple Regression models: Linear regression, ridge, lasso Proposal started with ID3, had to change due to working with continuous values not discreet.	Scraped data from 1961-2016 Used player stats for position players and pitchers and evaluated both separately using multiple regression models to create a weighted voting scheme, and apply votes to players for stats they achieved. Tried to mimic BBWAA voting mathematically.

14 - Music Popularity (“hotttness” prediction)	Classification = good Logistic regression, linear regression, support vector machines(support vector regression, RBD vs polynomial kernel, coefficient r^2=0.34), gaussian naive bayes	Million song db 12 statistics 4 algs
6 - Edible Mushrooms (we focused on precision for TP having low FP)	Linear Svc (linear svm) Svm Gaussian Naive bayes	-Very accurate 95% or more -Over fit/100% -Did well 85%-70%
10 - Oscars		Used 2 to predict top 10 then took intersection Social Media
7 - Bike share load balancing	Decision tree	Predicted when bike station will be empty, full or none at any hour of day Used for days in advance Allows for Bay Area Bike Share data for San Francisco
11 - Trump Tweets	Neural Network	Fooled people 17% of time Had website for guessing real trump tweet or theirs
2 - Academy Awards Prediction		
3 - Soccer Premier League Prediction	Regression Models	Didn’t handle relegation data well, chopped all the new teams data.
5 - Sentiment Analysis	Vectorizer gsvc	Yelp texts review -> /5 stars Maps were bad
17 - Eve Online Market Prediction	Linear Regression, Support Vector Regression, AdaBoost Regression	Calculated momentum attribute from simple market data; found a high return of 1.5x (50%) on initial over 3.5 months of trading.

Guest Lectures:

Daniel German		Demonstration using emacs and R on the Iris dataset
Brian Ziebart	a supervised machine learning framework that adversarially approximates the training data and uses the exact performance measure	Supervised machine learning as an adversarial game. Provides flexibility for addressing sample selection bias and for inductively Optimizing multivar performance measures like F-measure discounted Cumulative gain from information retrieval & ranking tasks.
George Tzanetakis	Markov Logic Networks	Automatically classifying the kind of music by sample. Tagging a song > bags of words >
David Johnson	Support Vector Machine	Kinect image processing edge detection in piano learners - computer vision

Pros and Cons of various models:

Model	**When to use**	**When not to use**
Decision tree	Discrete data	Continuous data
Regression tree	Discrete values Continuous prediction	Anything else
Naïve bayes	Text classification Few parameters, large data set	Lots of parameters
HMM	Sequence or temporal models	Anything else
Logistic regression	Lots of parameters Continuous values	Few parameters Discrete
K-means	Clustering Glob forms (spheres) Know how many clusters	Differing sizes, densities, shapes
Bisecting K-means	Clustering Don't know how many clusters spherical clusters	Assignment 3 Q3
Linear Regression	Continuous data and prediction	Discrete, lots of parameters
Neural Networks	Anything graphical Human brain replication	Very complicated Blackbox Slow to make and run
SVM	Linearly separable data	Non-binary set of data
Min/Max/Avg	Min: non-globular shapes Max: Good with noise	Min: Sensitive to noise and outliers Max: Tends to break large clusters Avg: Slow – big calculations
Euclidean distance	Finding similarities	Different scales
Pearson correlation	Finding similarities with different scales	Doesn't handle outliers well