Know Thy Neighbor: An Introduction to Scikit-Learn and K-NN

Portia Burton
PLB Analytics
www.github.com/pkafei

About Me:

- Organizer of the Portland Data Science group
- Volunteer of HackOregon
- Founder of PLB Analytics

What We will Cover Today

1. Brief Intro to Machine Learning
2. Go Over Scikit-learn
3. Explain the k-Nearest Neighbor algorithm
4. Demo of Scikit-learn and kNN

Machine Learning

Machine Learning

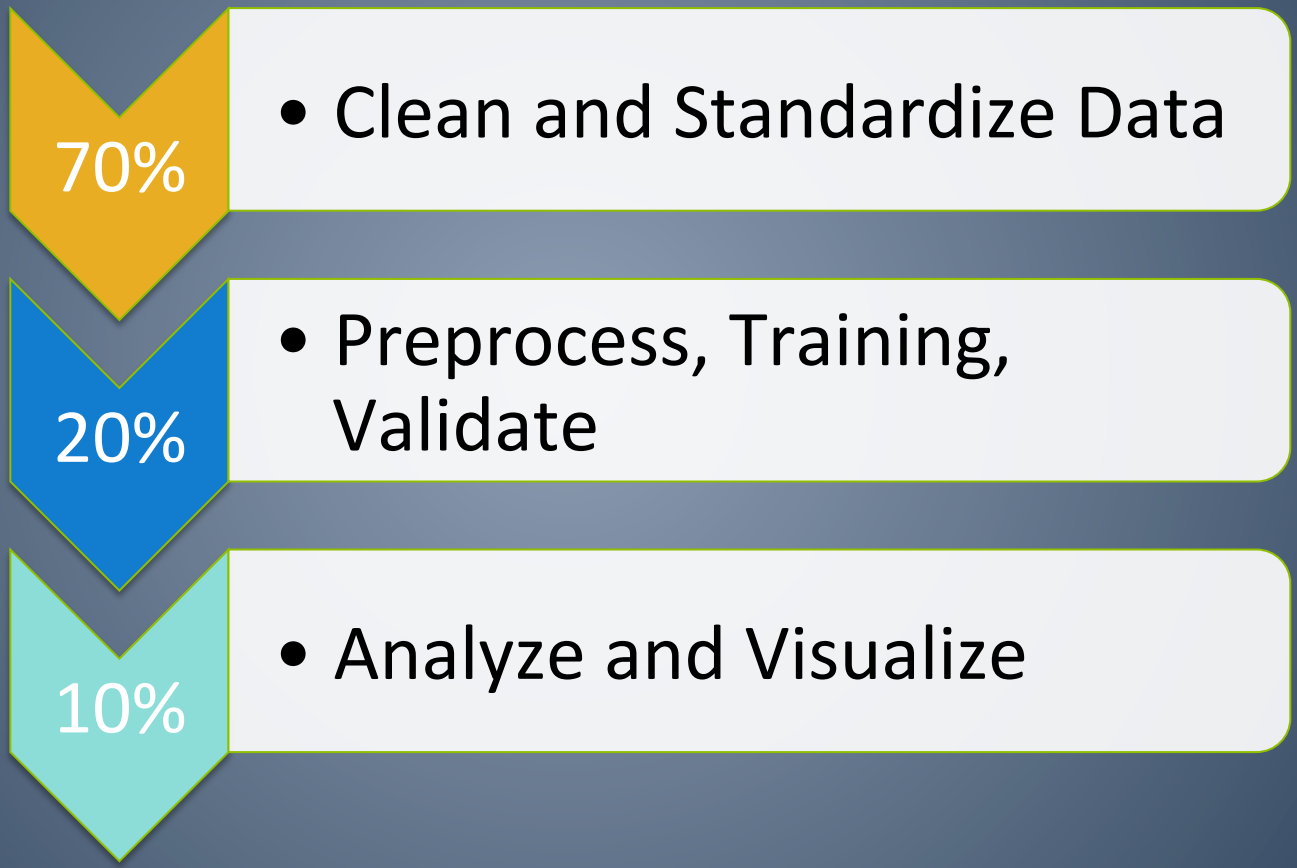
- The algorithm learns from the data

What is Machine Learning

Algorithms use data to….

- Create predictive models
- Classify unknown entities
- Discover patterns

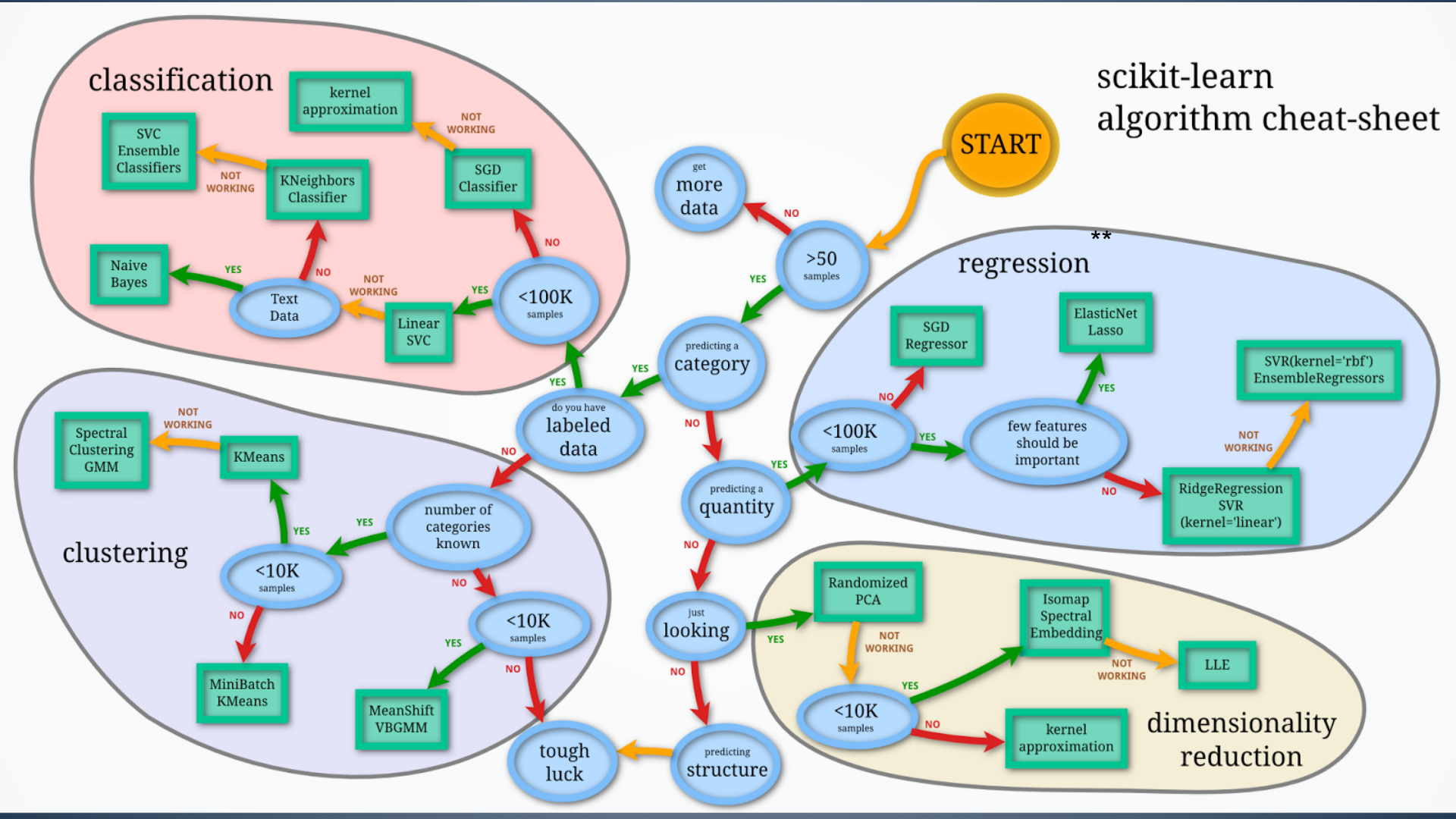
Basic Workflow of Machine Learning


70%
• Clean and Standardize Data
20%
• Preprocess, Training, Validate
10%
• Analyze and Visualize

Scikit-Learn

What is scikit-learn?

- Python machine learning package
- Great documentation
- Has built in datasets(i.e. Boston housing market)


scikit-learn
algorithm cheat-sheet
START
get more data
NO
>50 samples
YES
predicting a category
YES
do you have labeled data
YES
<100K samples
NO
SVC Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
NOT WORKING
kernel approximation
NOT WORKING
SGD Classifier
NO
Naive Bayes
YES
Text Data
NO
NOT WORKING
Linear SVC
YES
classification
NO
predicting a quantity
YES
<100K samples
YES
few features should be important
NO
SGD Regressor
NO
ElasticNet Lasso
YES
SVR(kernel='rbf') EnsembleRegressors
NOT WORKING
RidgeRegression SVR (kernel='linear')
**
regression
NO
just looking
YES
Randomized PCA
NOT WORKING
<10K samples
YES
Isomap Spectral Embedding
NOT WORKING
LLE
NO
kernel approximation
dimensionality reduction
NO
predicting structure
tough luck
NO
<10K samples
YES
number of categories known
YES
<10K samples
NO
Spectral Clustering GMM
NOT WORKING
KMeans
YES
MiniBatch KMeans
NO
MeanShift VBGMM
YES
clustering

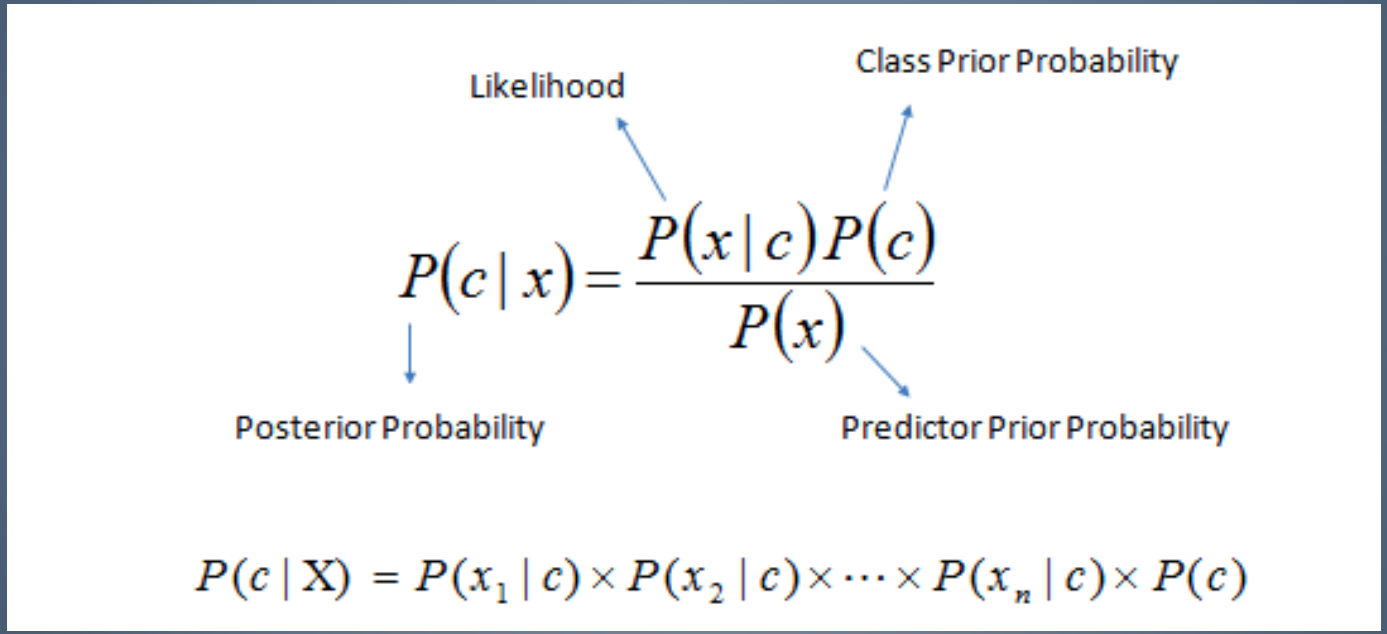
Many companies use Scikit-Learn


EVERNOTE®

Are You a Recipe? Yum.

- Distinguishes ‘recipe’ notes from ‘work’ notes
- Suggesting notebooks is a classification problem
- Implements naïve bayes classification algorithm

Naïve Bayes Classification



"naive" assumption of independence between every pair of features

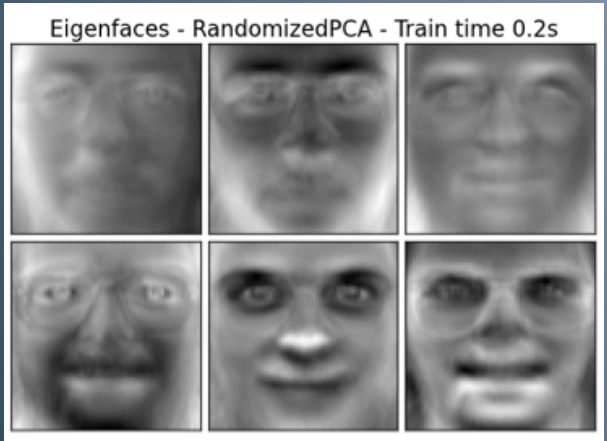
Supervised vs. Unsupervised Learning

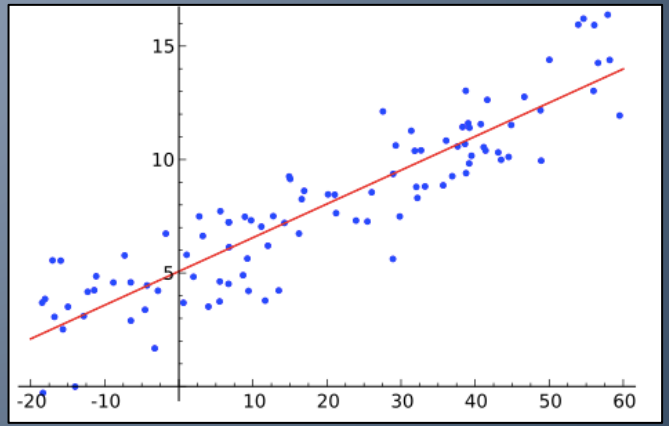
Unsupervised Learning

Data points are not labeled with *outcomes*.
Patterns are found by the algorithm.

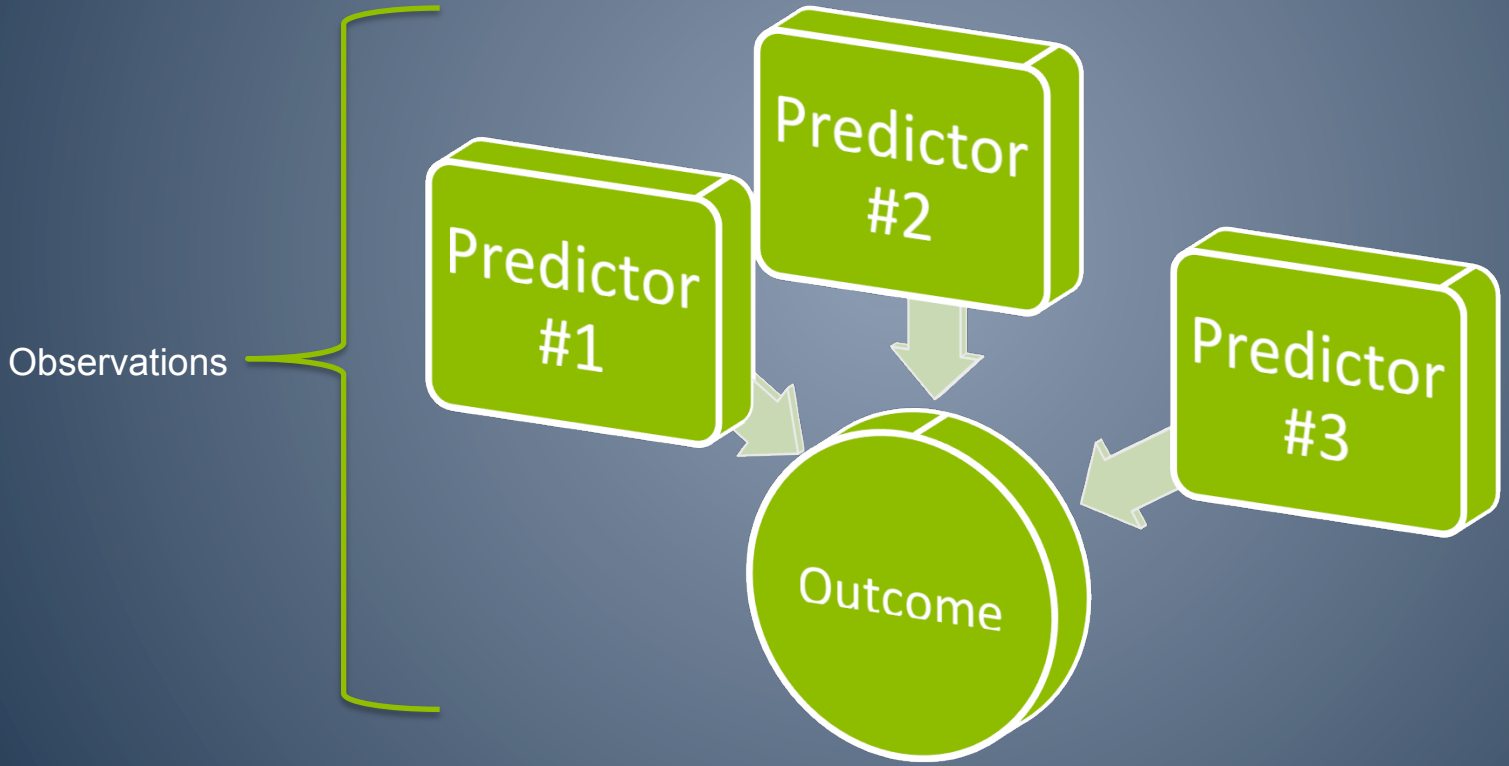
Supervised Learning

When your samples are labeled





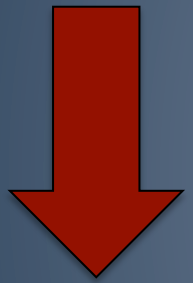
Theoretical Data Model for Supervised Learning



Remember to…

Keep your sample size high

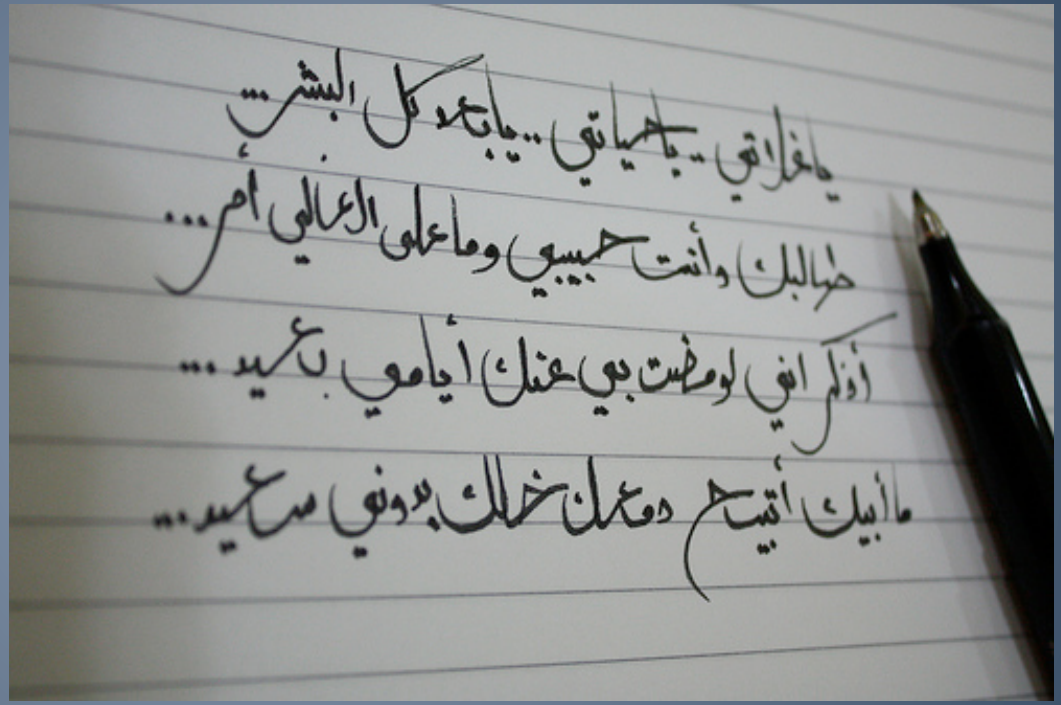
And don’t forget to



Keep your feature set low

Examples of Supervised Learning

Handwriting Analysis

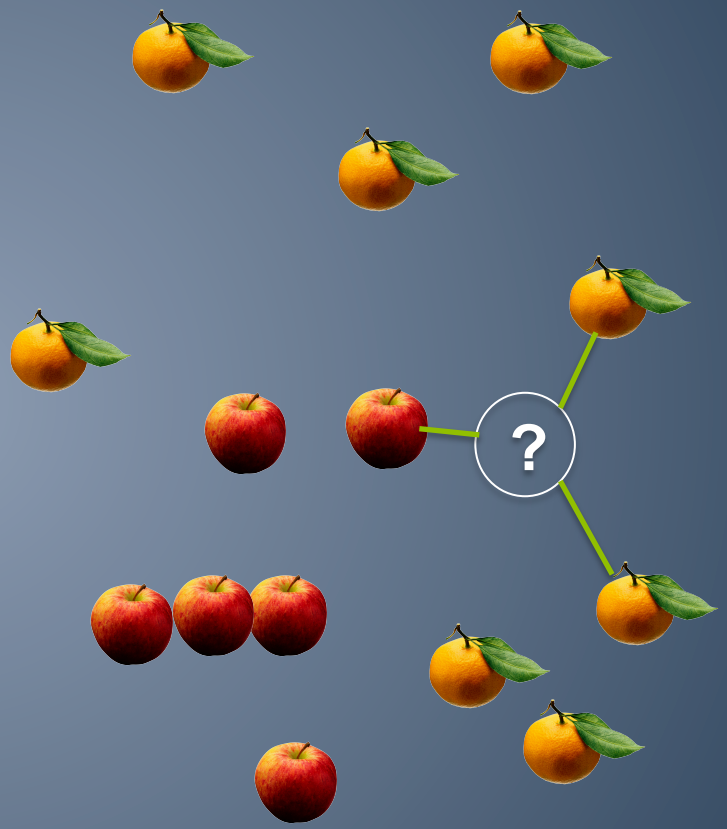


Spam Filters

k-NN

- k Nearest Neighbor algorithm
  - The simplest machine learning algorithm
  - It is a lazy algorithm : doesn’t run computations on the dataset until you give it a new data point you are trying to test
  - Our example uses k-NN for supervised learning

Mystery Fruit

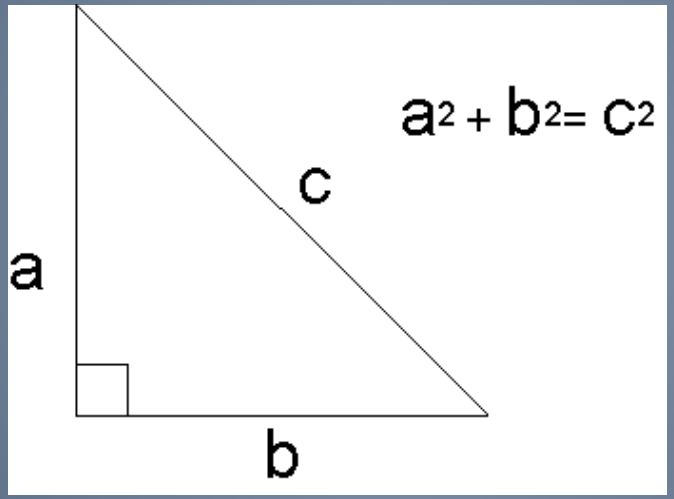


Majority Vote

- Equal weight: Each kNN neighbor has equal weight

- Distance weight: Each kNN neighbor’s vote is based on the distance

How k-NN works



Downsides of kNN

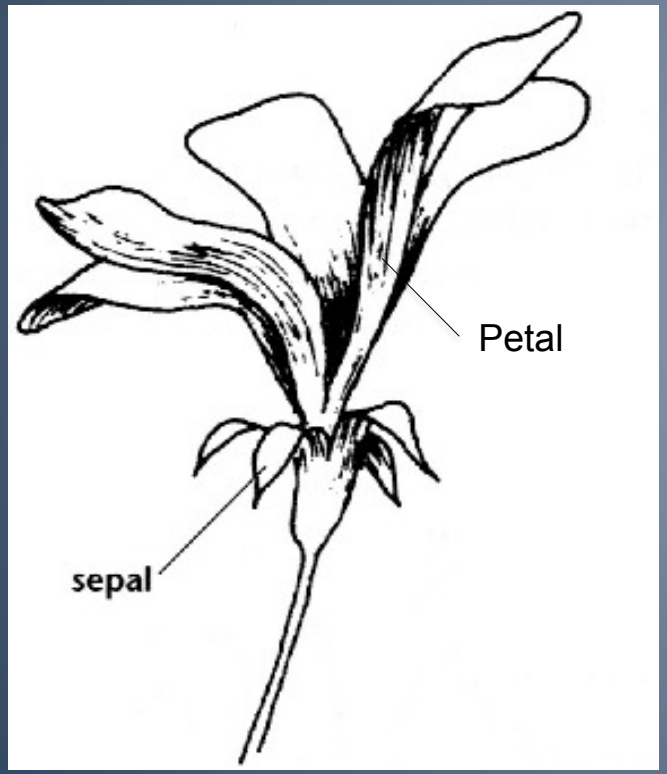
- Since there is minimum training there is a high computational cost in testing new data
- Correlation is falsely high (data points can be given too much weight)

Live demo time!

Our Data Set:

- Typical!
- Multivariate data set was created in 1936
- Analyzed by Sir Ronald Fischer
- Collected by Edgar Anderson

Live coding demo: the data set

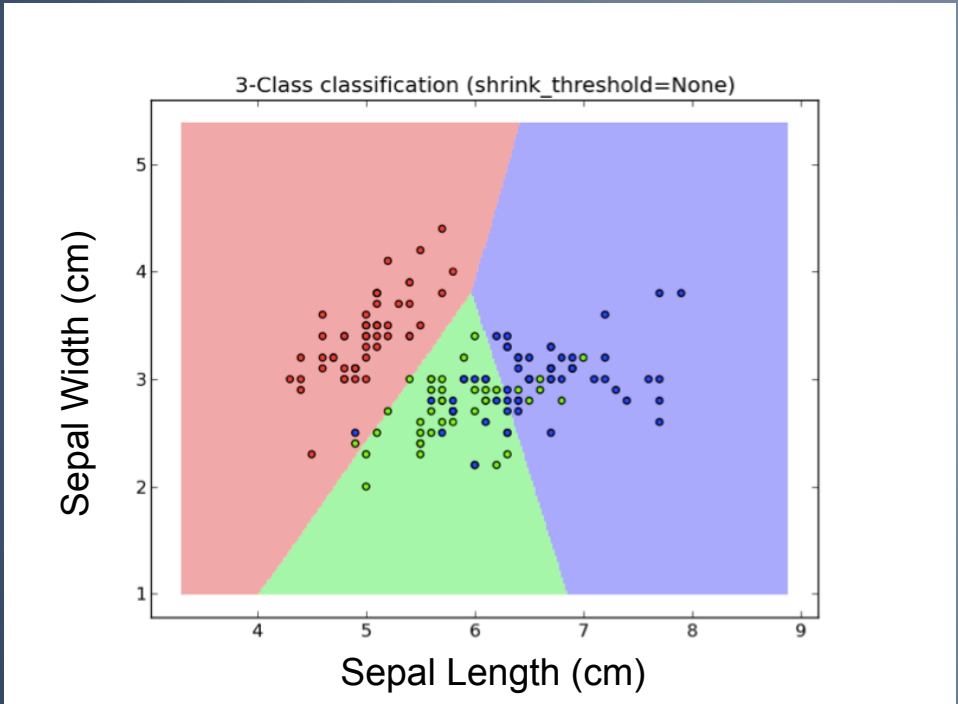


Iris setosa

Iris virginica

Iris versicolor

The plot from the use case



Training Data

Test Data

Example data points for each iris species

Sepal length (x-axis)	Sepal width (y-axis)	Species
5.1	3.5	*I. setosa*
5.5	2.3	*I. versicolor*
6.7	2.5	*I. virginica*

References:

http://www.solver.com/xlminer/help/k-nearest-neighbors-prediction-example

http://saravananthirumuruganathan.wordpress.com/2010/05/17/a-detailed-introduction-to-k-nearest-neighbor-knn-algorithm/

http://scikit-learn.org/stable/modules/neighbors.html

http://peekaboo-vision.blogspot.com/2013/01/machine-learning-cheat-sheet-for-scikit.html

http://stackoverflow.com/questions/1832076/what-is-the-difference-between-supervised-learning-and-unsupervised-learning

http://stackoverflow.com/questions/2620343/what-is-machine-learning

References:

http://blog.evernote.com/tech/2013/01/22/stay-classified/

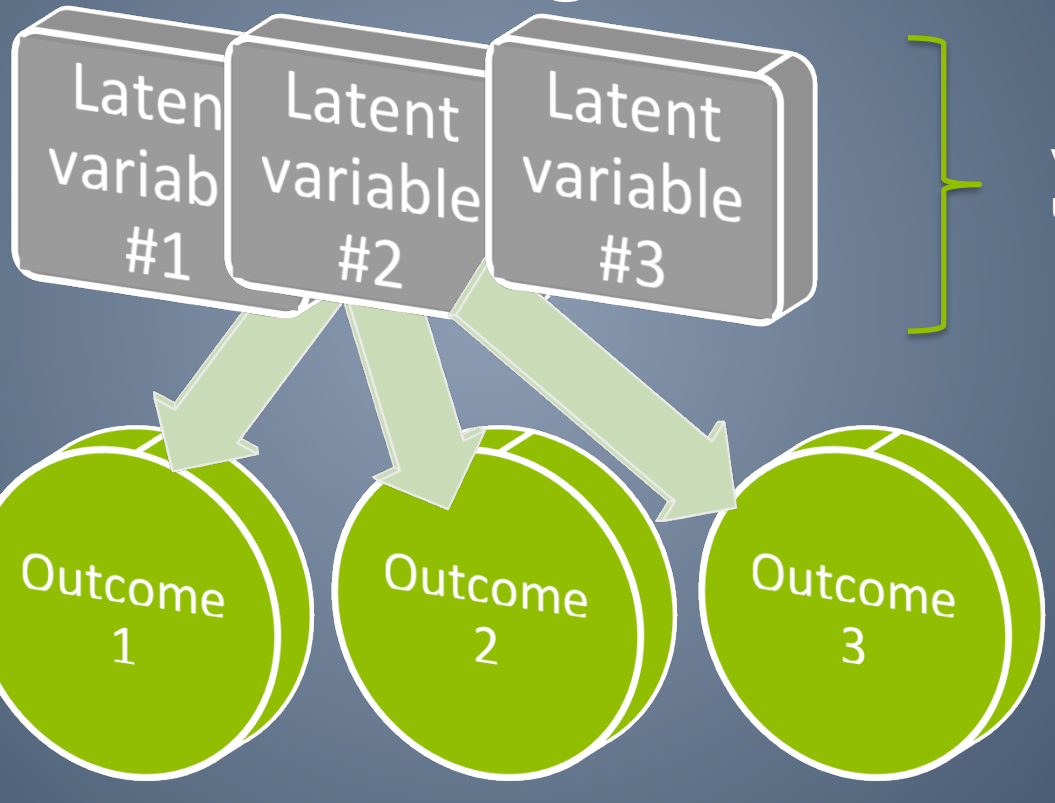
http://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf

http://en.wikipedia.org/wiki/Iris_flower_data_set

http://en.wikipedia.org/wiki/Support_vector_machine

Extra Slides

Theoretical data model for unsupervised learning



Variables that are unknown to us

The "outcomes" are our observations. This is what is given to the algorithm

Output of algorithm: Relationships among the 'outcomes'. Ex: clusters of data points