A Gentle Introduction To Machine Learning

Kyle Kastner
Southwest Research Institute (SwRI)
University of Texas - San Antonio (UTSA)

Outline

- Why Use Machine Learning?
- Workflow
- Resources
- Final Comments

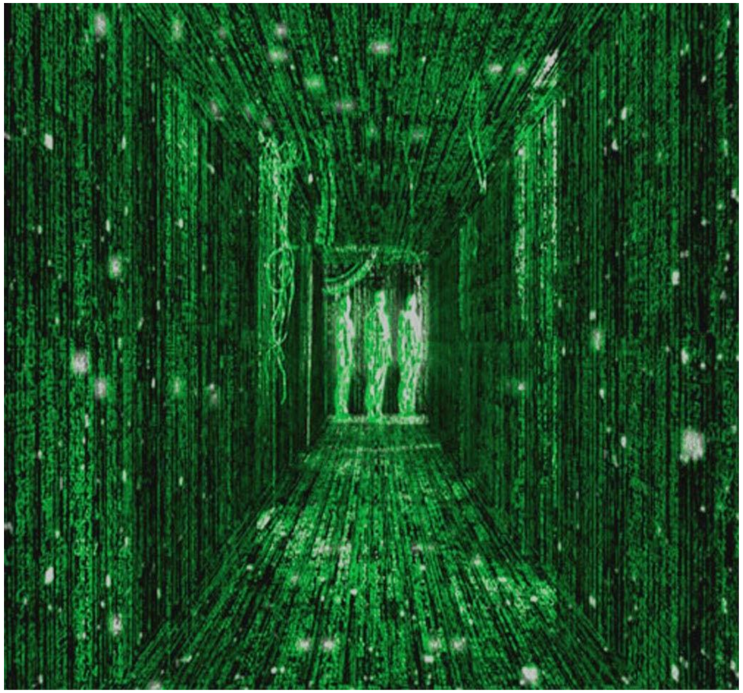
Why Use Machine Learning?

- Drowning in data
- Computers are cheap, humans are expensive
- Psychic superpowers (sometimes)

http://blog.thepertgroup.com

Types of Problems

- Regression (Supervised)
  - Predict housing prices
- Classification (Supervised)
  - Handwritten digit recognition
- Clustering (Unsupervised)
  - Document tagging



http://2.bp.blogspot.com

Where to Start?

- Know your data
  - If labeled, supervised learning
  - Unlabeled, try unsupervised

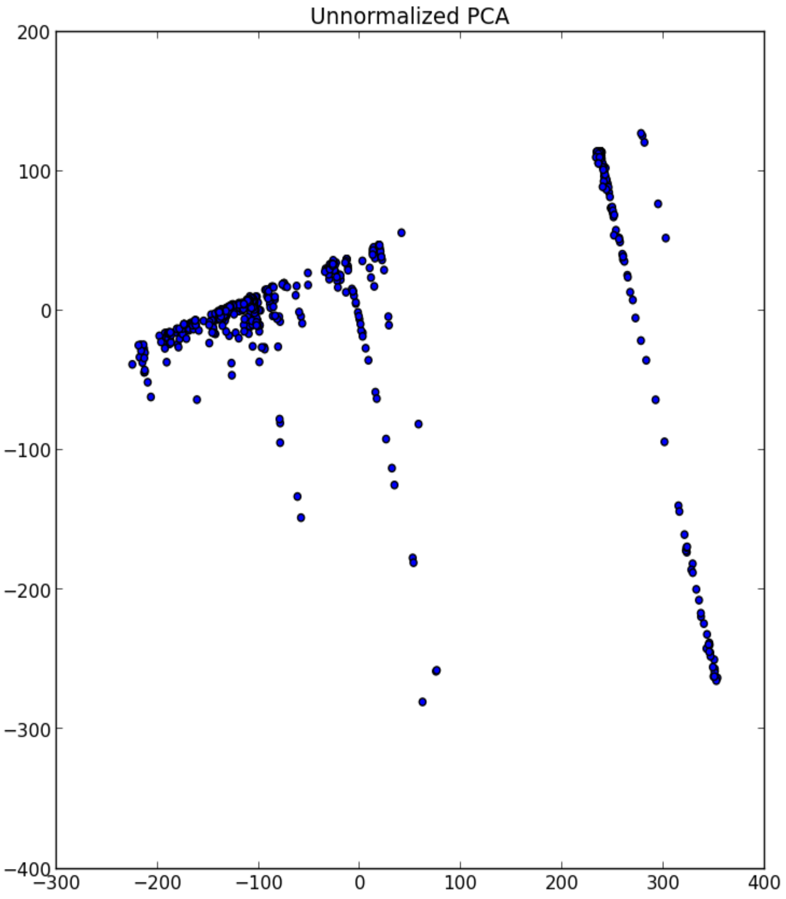
- Clean it up
  - Normalize by removing mean and dividing by variance
  - Visualize in 2D

- Separate training data
  - Try 80/20% train/test split, randomly chosen

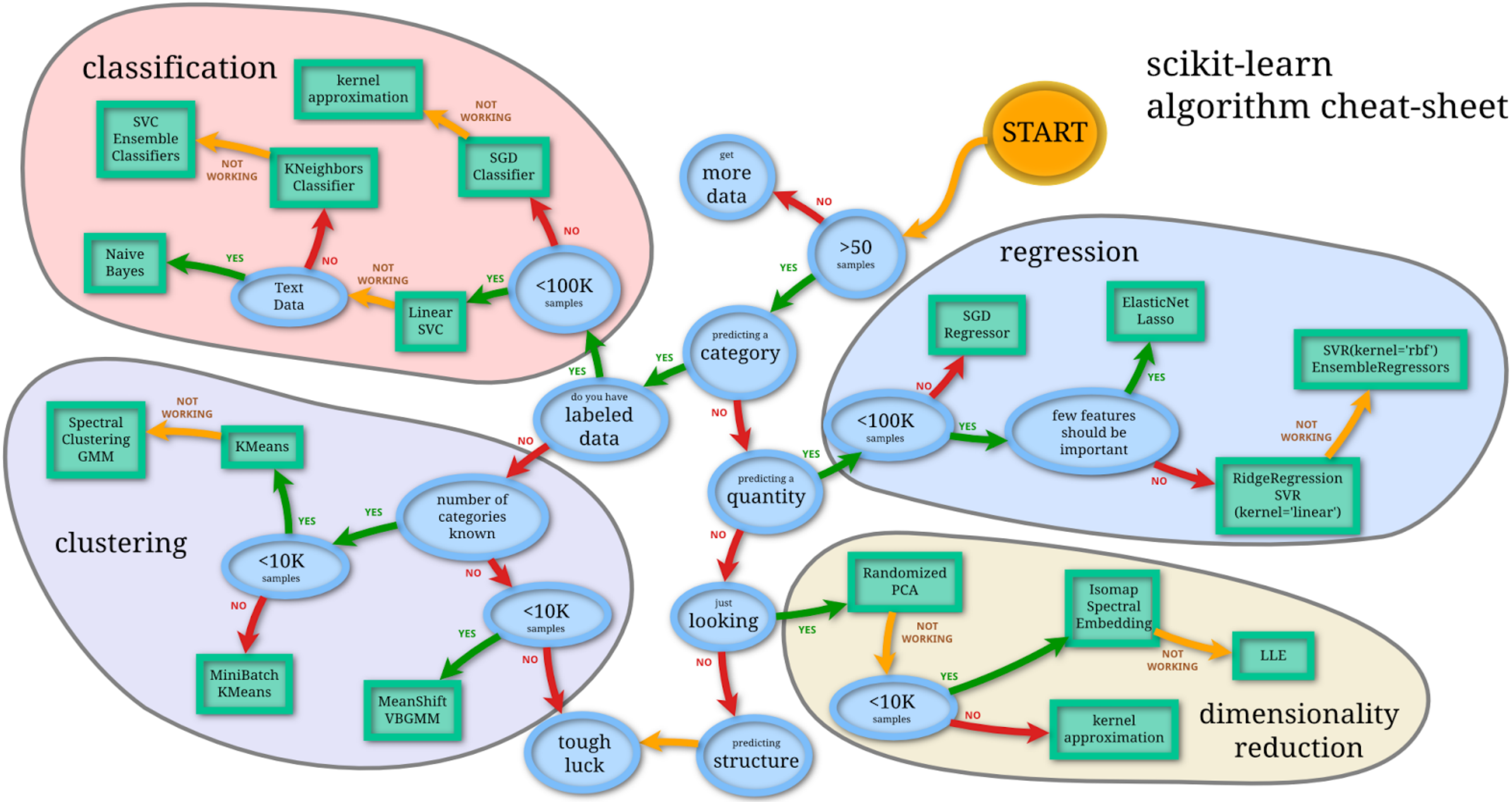
www.glassdoor.com

Preprocessing

- Typically normalize by subtracting mean and dividing by variance
- Use Principle Component Analysis (PCA) to keep structure while reducing dimensions
- PCA to plot N-dimensional data in 2D or 3D



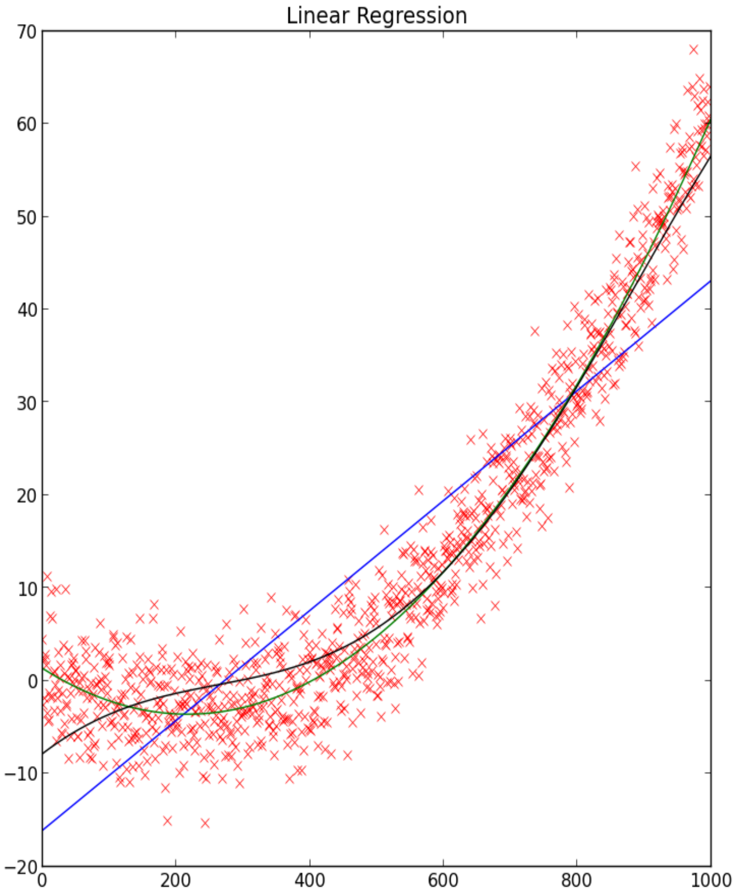
Selecting an Algorithm



http://peekaboo-vision.blogspot.com

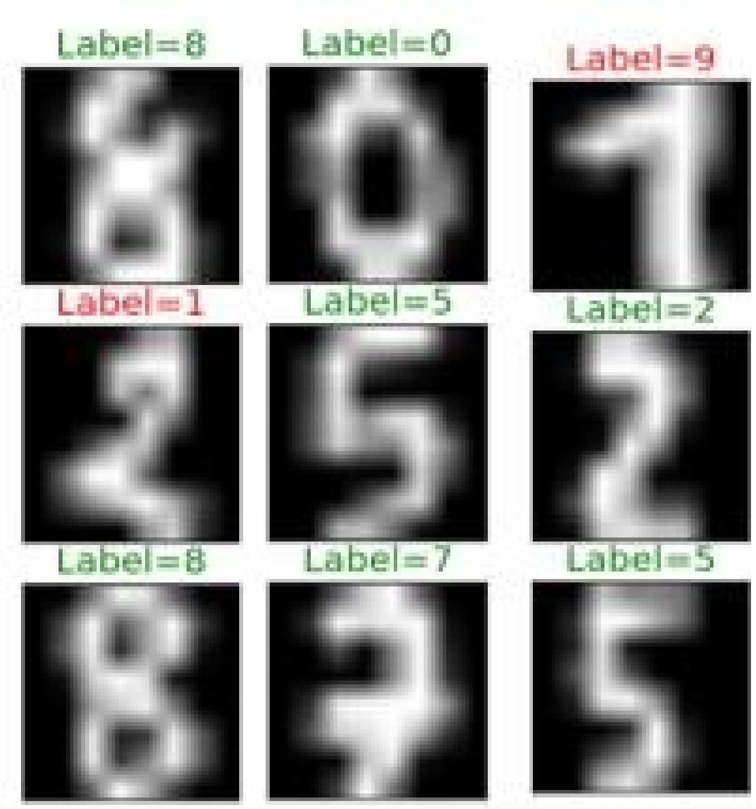
Linear Regression

- Find the "best fit" line
- Outliers will greatly affect results
- Perform regression into different basis
- Basis can be Fourier, polynomial, wavelet, etc.



Logistic Regression

- Optimize parameters for each class label
- Choose class with highest probability
- Can be very powerful, especially after PCA

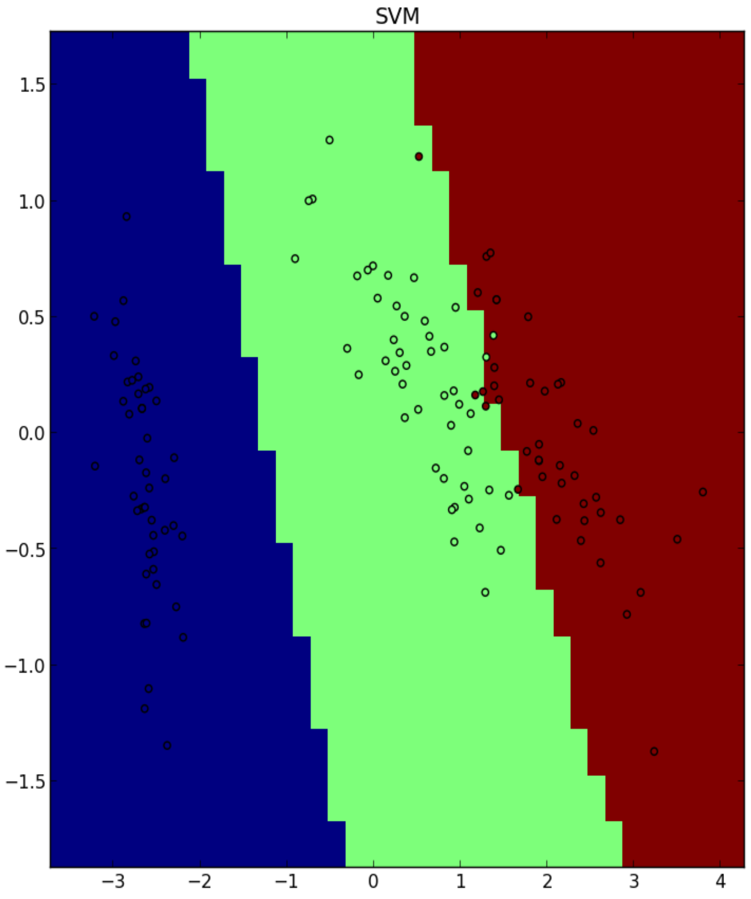


Support Vector Machine (SVM)

- Margin parameter is a configurable "allowed error" to account for class overlap

- Boundaries use a semi-arbitrary "kernel" function

- Linear, polynomial, wavelet, sigmoid



Data

- **from sklearn import datasets**
- Iris, Digits are excellent for classification
- Boston for regression
- Any classification dataset (sans labels) for clustering
- Very good for generating data

http://en.memory-alpha.org

Resources

- Scikit-learn documentation and examples
  - The infamous cheat sheet
- Coursera courses
  - Andrew Ng's Machine Learning
- Pattern Recognition and Machine Learning
  - Christopher M. Bishop

Final Comments

- Machine learning is a spectrum
- Data preprocessing is vital
- Prefer simple models to complex ones
- Use **sklearn**

Questions?

Code on GitHub:

https://github.com/kastnerkyle/SciPy2013

Bonus: Trends in Machine Learning

- Deep networks
- Generative models
- Unsupervised data from Youtube
- Text-to-speech
- Image object recognition
- Google+ untagged image search