A Crash Course in H2O Machine Learning Using Python

Online Resources

- H2O Tutorials and Training Material
  - https://github.com/h2oai/h2o-tutorials
  - https://github.com/h2oai/h2o-tutorials/tree/master/training/lending_club_exercise
  - https://github.com/h2oai/app-consumer-loan
- Additional Datasets
  - https://s3.amazonaws.com/h2o-public-test-data/bigdata/laptop

H2O Training

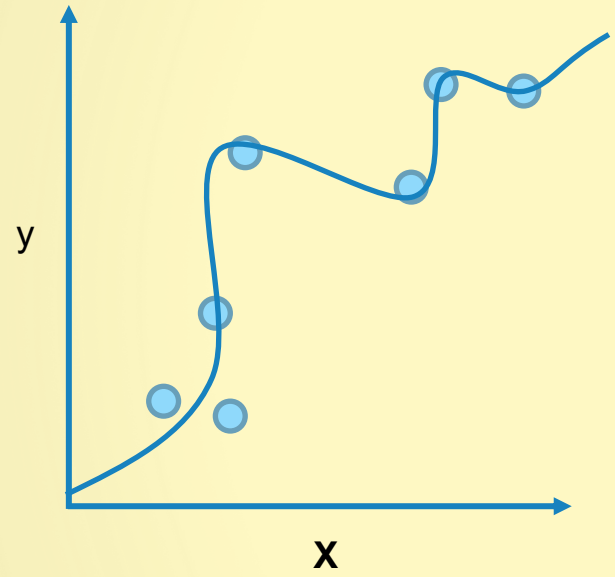
1. Introduction to Machine Learning
2. Introduction to H2O
3. H2O Data Munging
4. Machine Learning Concepts
5. H2O Algorithms
6. Productionalizing Models

What is Machine Learning?

- Set of statistical and optimization tools to model data
  - Supervised Learning
  - Unsupervised Learning
- Focus on “production analytics”
  - Reducing need for
    - Sampling
    - Periodic revision of models
    - Subjective priors
    - Hand coding models in a production language

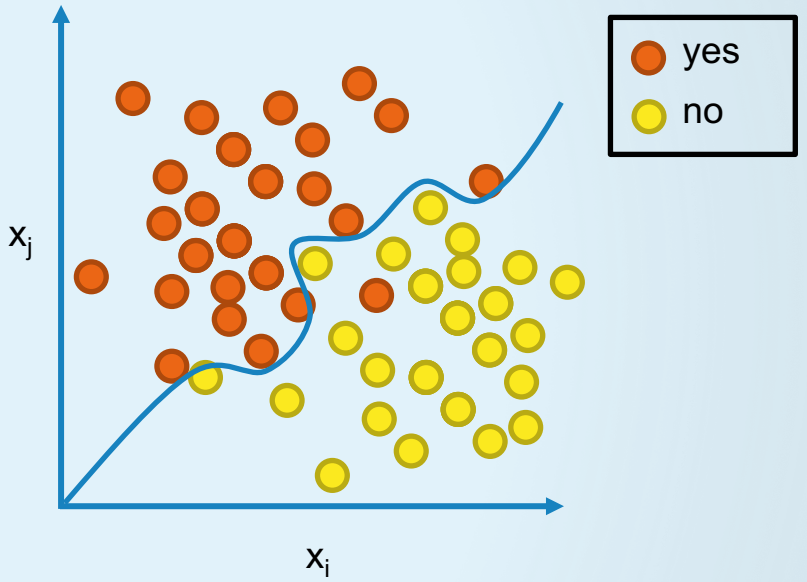
Supervised Learning

Regression:
How much will a customers spend?



H_2O algos:
Penalized Linear Models
Random Forest
Gradient Boosting
Neural Networks
Stacked Ensembles

Classification:
Will a customer make a purchase? Yes or No

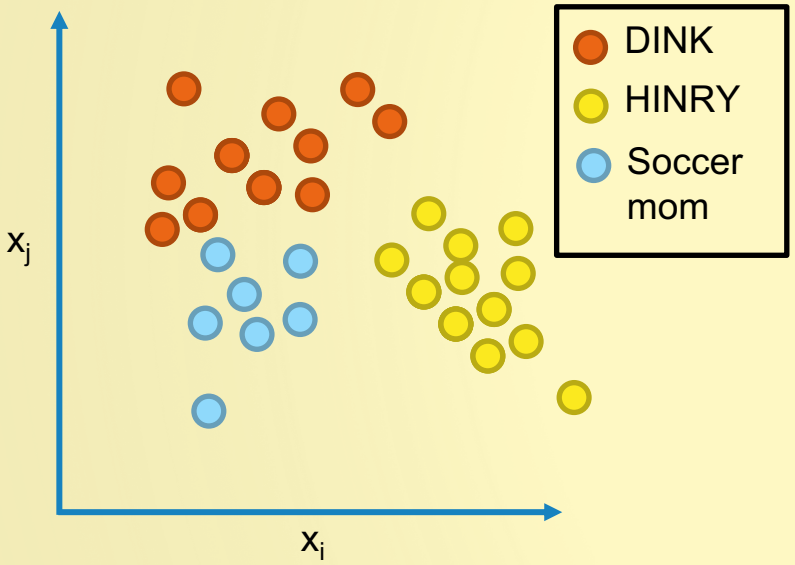


H_2O algos:
Penalized Linear Models
Naïve Bayes
Random Forest
Gradient Boosting
Neural Networks
Stacked Ensembles

Unsupervised Learning

Clustering:

Grouping rows – e.g. creating groups of similar customers

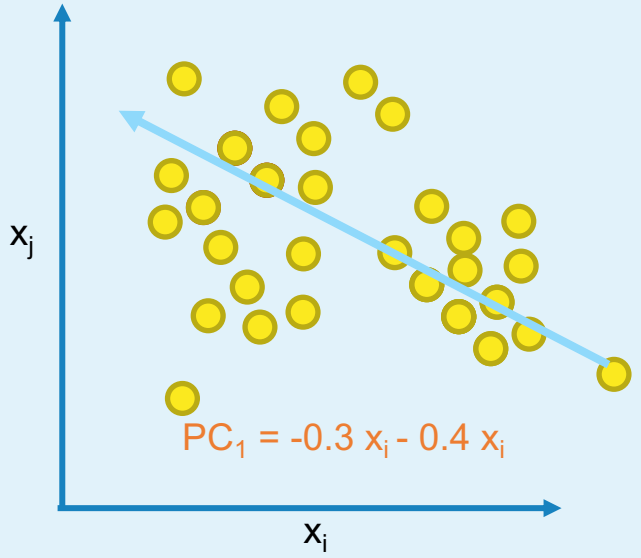


H_2O algos:

k – means

Feature extraction:

Grouping columns – Create a small number of new representative dimensions

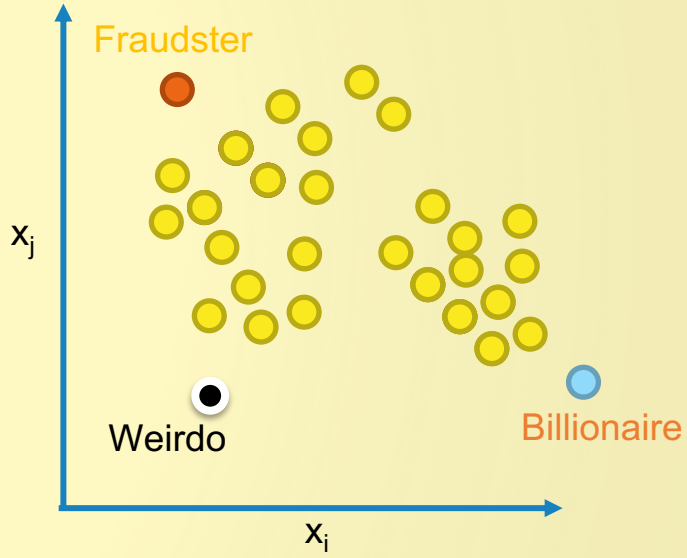


H_2O algos:

Principal components
Generalized low rank models
Autoencoders
Word2Vec

Anomaly detection:

Detecting outlying rows - Finding high-value, fraudulent, or weird customers



H_2O algos:

Principal components
Generalized low rank models
Autoencoders

Vocabulary

Concept	Statistics\Econometrics	Machine Learning
“Computation”	Fit\Estimate	Train
“Left-hand side”	Dependent variable	Target
“Right-hand side”	Regressor\Predictor\Class	Feature\Factor\Enum
“Goal”	Estimation\Explanation	Prediction

Machine Learning Methods

Supervised Learning Methods

- Regression (GLM)
  - Lasso
  - Ridge
  - E
- De
- Ra
- Gradient Boosted Models
- Support Vector Machine
- Neural Network
- Deep Learning

Know Y

Unsupervised Learning Methods

- Clustering
  - Kmeans
  - Hierarchical
- Princ
- Auto
- Non-
- Generalized Low Rank Models

Don’t know Y

When do I use what tool?