H2O.ai Algorithms

Patrick Hall
22 Jan 2017

H2O.ai

Algorithms on H2O

Supervised Learning

Statistical Analysis

- **Penalized Linear Models**: Binomial, Gaussian, Gamma, Poisson and Tweedie – HIGHLY INTERPRETABLE
- **Naïve Bayes**

Ensembles

- **Distributed Random Forest**: Classification or regression models
- **Gradient Boosting Machine**: Produces an ensemble of decision trees with increasing refined approximations

Neural Networks

Multilayer Perceptron

Deep Learning

- **Deep neural networks**: Multi-layer feed forward neural networks for standard supervised learning tasks
- **Convolutional neural networks:** Sophisticated architectures for pattern recognition in images, sound, and text

Unsupervised Learning

Clustering

- **K-means**: Partitions observations into k clusters/groups of the same spatial size. Automatically detect optimal k

Dimensionality Reduction

- **Principal Component Analysis**: Linearly transforms correlated variables to independent components
- **Generalized Low Rank Models:** extend the idea of PCA to handle arbitrary data consisting of numerical, Boolean, categorical, and missing data

Anomaly Detection

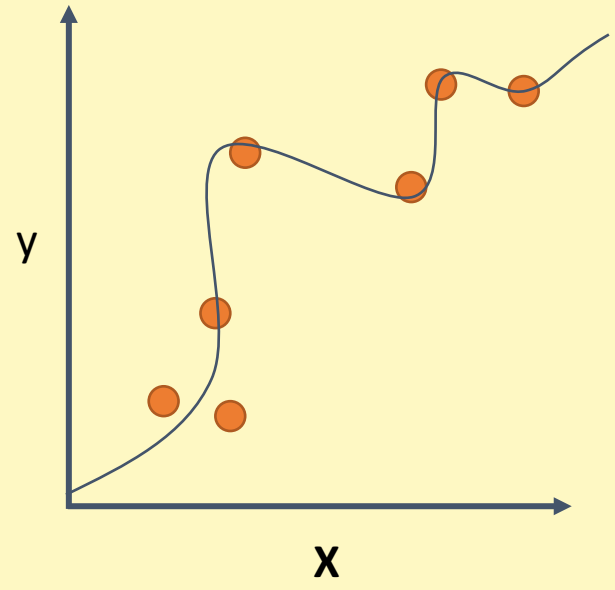
- **Autoencoders**: Find outliers using a nonlinear dimensionality reduction based on neural networks

Term Embeddings

- **Word2vec:** Generate context-sensitive numerical representations of a large text corpus

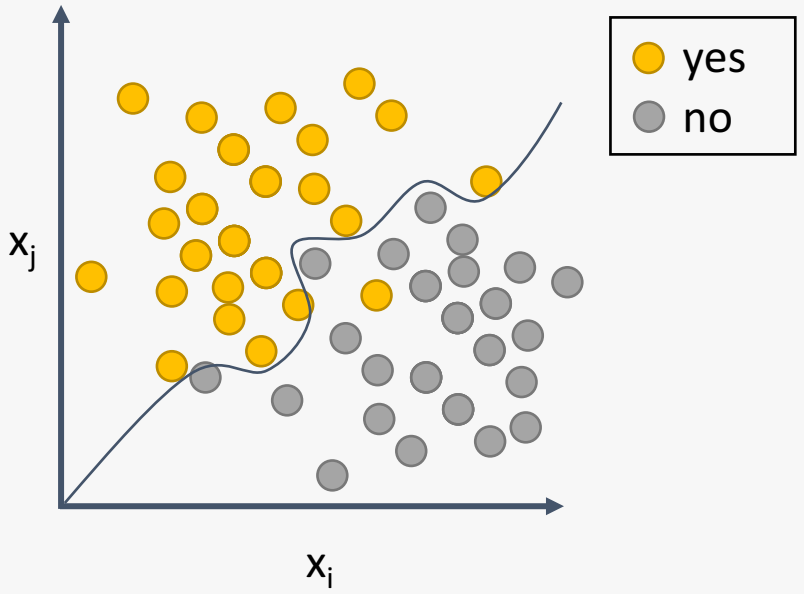
Supervised Learning

Regression:
How much will a customers spend?



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

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



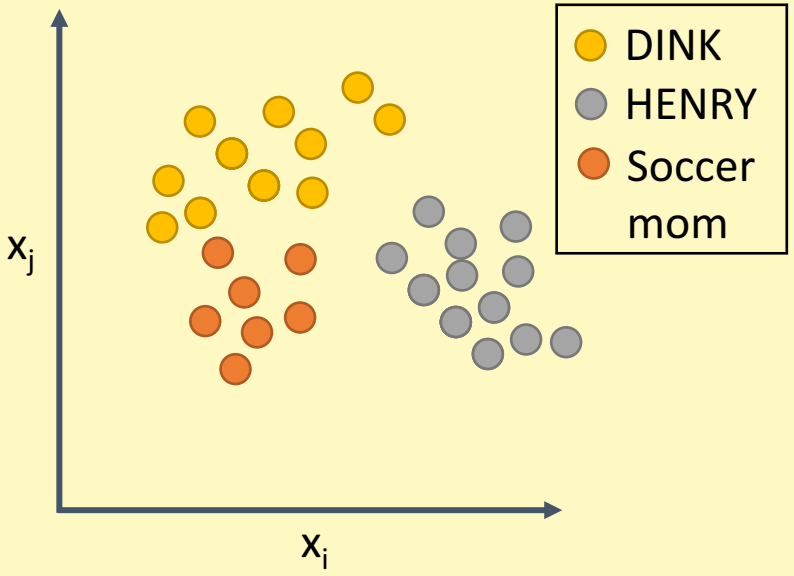
H_2O algos:
Penalized Linear Models
Naïve Bayes
Random Forest
Gradient Boosting
Neural Networks

H2O.ai

Unsupervised Learning

Clustering:

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

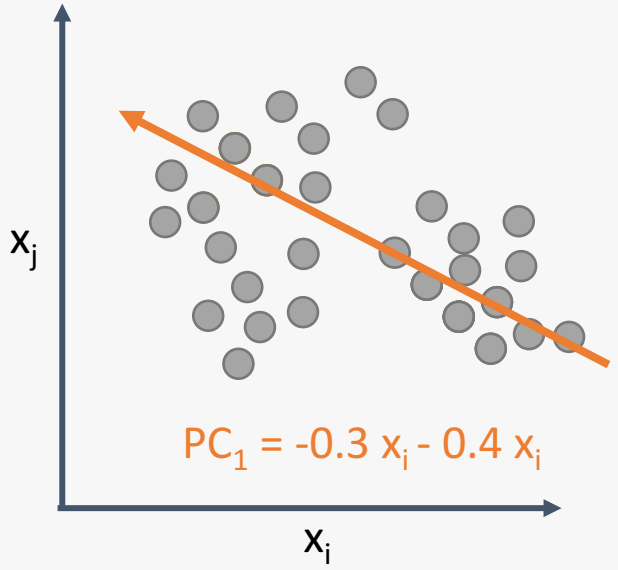


H_2O algos:

k – means

Feature extraction:

Grouping columns – Create a small number of new representative dimensions

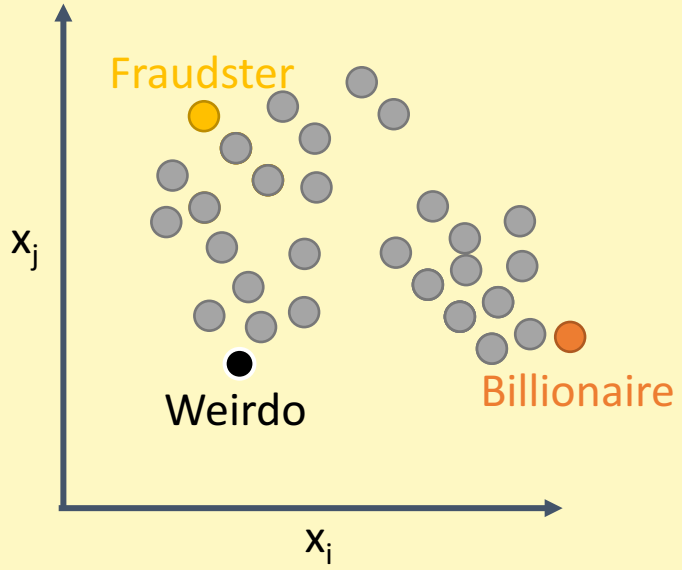


H_2O algos:

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

H_2O.ai

H2O.ai

	Usage	Recommendations	Problems
Penalized Linear Models	• Regression • Classification	• Best for linear or linearly separable relationships • Nonlinear and interaction terms should be specified by users • Great for wide data • Creates interpretable models	• NAs • Outliers/influential points • Strongly correlated inputs • Rare categorical levels in new data
Naïve Bayes	• Classification	• Best for linearly separable relationships in big data • Great for huge data sets where other methods fail	• Linear independence assumption • Rare categorical levels in new data
Random Forest	• Regression • Classification	• Single trees great for non-smooth relationships • Great for nonlinear and nonlinearly separable relationships in dirty data • Great for modeling interactions • Great for NAs and outliers	• Overtraining • Many hyperparameters
Gradient Boosting Machines	• Regression • Classification	• Single trees great for non-smooth relationships • Great for nonlinear and nonlinearly separable relationships in dirty data • Great for modeling interactions • Great for NAs and outliers	• Overtraining • Many hyperparameters • Often more accurate than random forest but also potentially more susceptible to instability with noisy data
Neural Networks (Deep learning & MLP)	• Regression • Classification	• Great for nonlinear and nonlinearly separable relationships • Great for modeling interactions in fully connected topologies • Deep water is well-suited for pattern recognition in images, videos, and sound	• NAs • Overtraining • Outliers/influential points • Long training times • Many hyperparameters • Strongly correlated inputs • Rare categorical levels in new data

H2O.ai

	Usage	Recommendations	Problems
k **- means**	• Clustering	• Great for creating Gaussian, non-overlapping, roughly equally sized clusters • The number of clusters can be unknown	• NAs • Outliers/influential points • Strongly correlated inputs • Cluster labels sensitive to initialization • Curse of dimensionality
Principal Components Analysis	• Feature extraction • Dimension reduction • Anomaly detection	• Great for extracting a number <= ***N*** of linear, orthogonal features from i.i.d. numeric data • Great for plotting extracted features in a reduced-dimensional space to analyze data structure, e.g. clusters, hierarchy, sparsity, outliers	• NAs • Outliers/influential points • Categorical inputs
Generalized Low Rank Models	• Feature extraction • Dimension reduction • Anomaly detection • Matrix completion	• Great for extracting linear features from mixed data • Great for plotting extracted features in a reduced-dimensional space to analyze data structure, e.g. clusters, hierarchy, sparsity, outliers • Great for imputing NAs	• Outliers/influential points
Autoencoders (Neural Networks)	• Feature extraction • Dimension reduction • Anomaly detection	• Great for extracting a number of nonlinear features from mixed data • Great for plotting extracted features in a reduced dimensional space to analyze structure, e.g. clusters, hierarchy, sparsity, outliers	• NAs • Overtraining • Outliers/influential points • Long training times • Many hyperparameters • Strongly correlated inputs • Rare categorical levels in new data
Word2Vec	• Highly representative feature extraction from text	• Great for extracting highly representative, context sensitive term embeddings (e.g. numerical vectors) from text • Great for text preprocessing prior to further supervised or unsupervised analysis	• Many Hyperparameters • Overtraining • Specifying term weightings prior to training • Long training times