Bradley Reardon

DATS 6202

HW 1

5/17/21

1. Machine learning is the process of teaching a computer through inputting data. The goal is to train/teach the computer well enough for it to solve problems involving prediction, classification, clustering, and pattern recognition similarly to how a human might solve those problems.
2. The main difference between supervised and unsupervised learning is that supervised learning requires a training dataset that contains a label feature used to train a model to accurately predict the labels for observations lacking a label. Unsupervised learning does not require a label feature and thus learns to group observations based on similarities in features with other observations.
   1. Supervised learning: linear regression, random forest classification, SVM
   2. Unsupervised learning: K-means clustering, dimension reduction techniques (i.e PCA)
3. 1. Fitting (linear regression)
   2. Pattern recognition (classification)
   3. Clustering
   4. Prediction (time-series)
4. Reinforcement learning can be used to solve problems pertaining to game theory, control theory, and multi agent systems. Some industry applications include autonomous automobiles, industrial robots used to operate more efficiently and safely on issues that might cause harm or use excess energy when human-operated, stock/market/sales prediction through time-series modeling, and so on.
5. Classification example:

Class 1

Binary Classifier

Input Data

Class 2

1. 1. **Regression:** linear, ridge, lasso, polynomial, Bayesian, stepwise
   2. **Classification:** Naïve Bayes, stochastic gradient descent, knn, decision tree, random forest, SVM, logistic regression
2. 1. **Clustering:** centroid-based, density-based, distribution-based, hierarchal
   2. **Reinforcement:** brute force, value function, direct policy search, monte carlo methods, temporal difference methods
3. The curse of dimensionality refers to phenomena that arise when analyzing data in high-dimensional spaces that do not occur with low-dimensional spaces. When dimensionality increases, the space available also increases but at a much larger rate. The scarcity of data that arises from such happening can be an issue for methods that require statistical significance.
4. 1. PCA
   2. Feature selection (backwards & forwards)
   3. Missing value ratio

Machine Learning Process

Gather raw data (input data)

Preprocess/clean data

Create train/test data if using a supervised model

Input data into model (train if using aa supervised model)

Compare with other models

Evaluate model

Choose the best model

Interpret and report results.

Deploy Model