Giving Computers the Ability to Learn from Data

Machine Learning

- Self-learning for spotting patterns in data and make prediction model
- turn data into knowledge
- age of abboundant data, providing powerful open source libraries

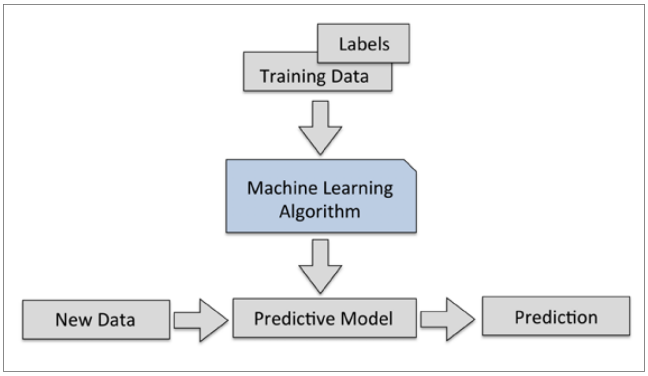
Building intelligent machines to transform data into knowlege

- as a subfield of artificial intelligence with self-learning
- capturing the knowlege in data to gradually improve performance of predictive models
- make data-driven decisions

The three different types of machine learning

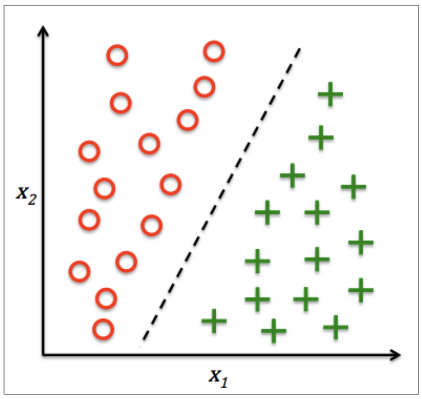
Supervised learning

- learn a model from labeled training data to make predictions about unseen or future data



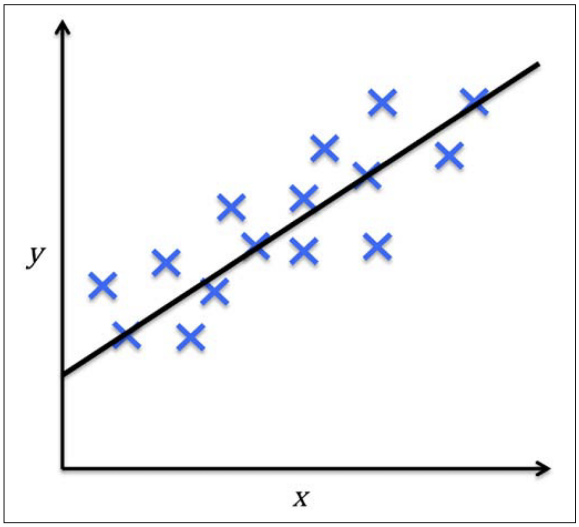
Binary classification

- spam classification
- learn a model (the dicision boundary)



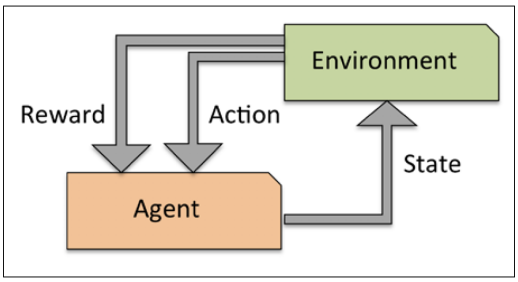
Regression

- find relationship between explanatory variables and continuous response variable
- ex) predict Math SAT score based on time spent studying



Reinforcement learning

- develope a system(agent) that improves performance based on interactions with enviornment
- reward signal but not superviesd signal
- reward is not corrective ground truth label of value
- learn a series of actions that maximize this reward via exploratory trial-and-error approach or deliberative planning

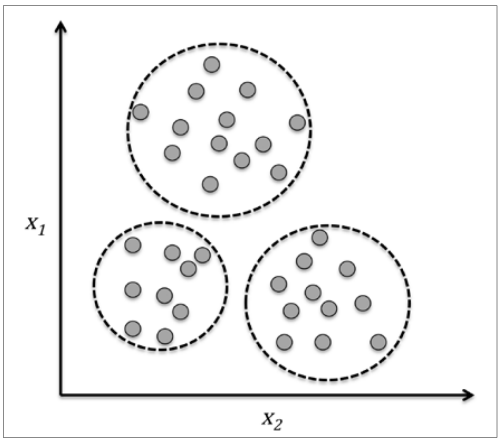


Unsupervised learning

- no right answer or no reward at all
- discover/explore unknown structure of data to extract meaningful information

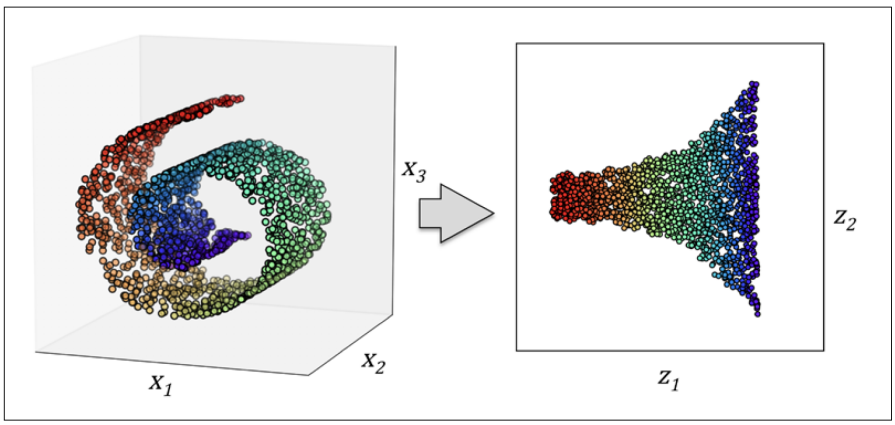
Finding subgroups with clustering

- find meaningful subgropus without any prior
- unsuervised classification



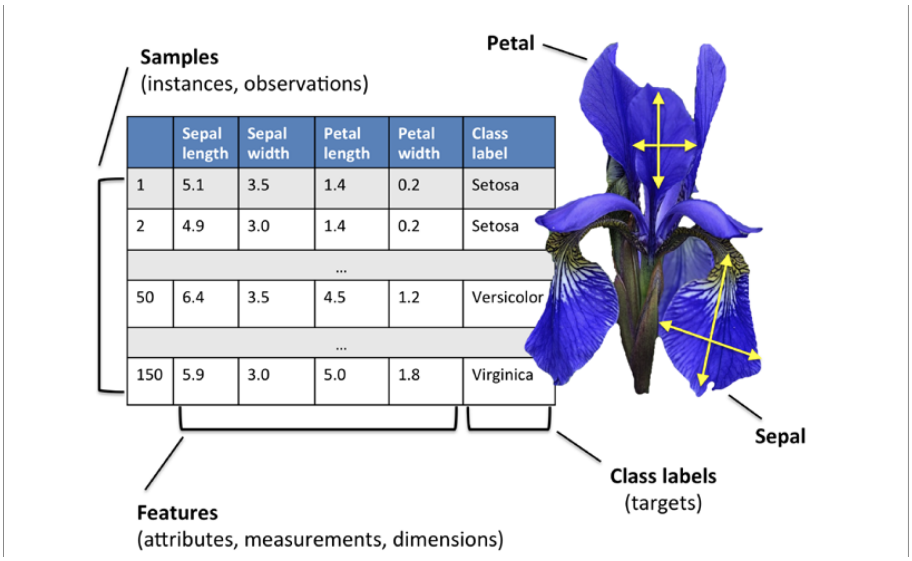
Dimensionality reduction for data compression

- when data of high dimensionality
  - challenge for limited sotrage and computation time
- unsupervised dimensionality reduction
- feature processing to remove noise from data



Basic terminology and notations

- matrix and vector notations
- data instance
- feature

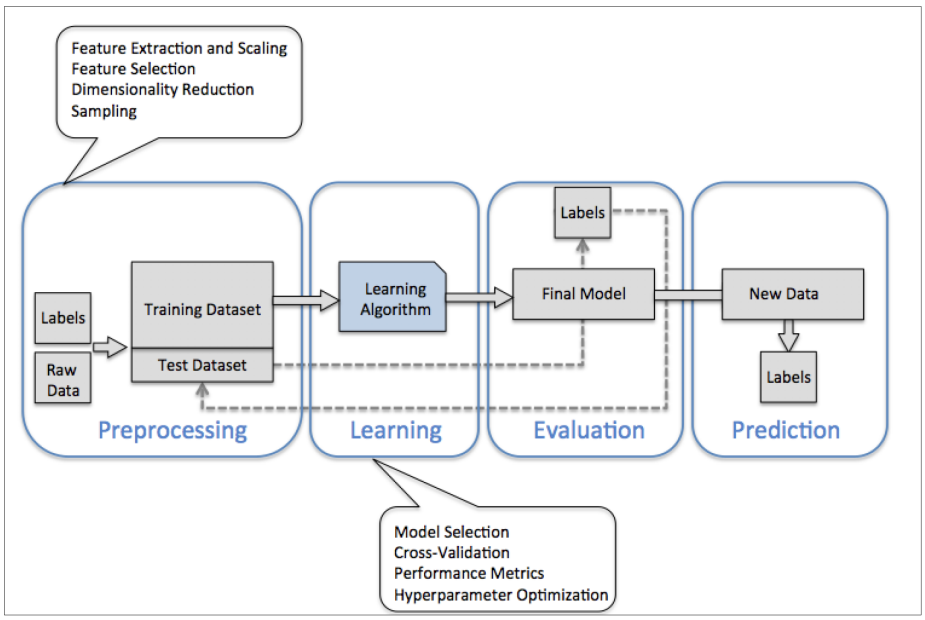


	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

The Iris dataset, consisting of 150 samples and 4 features, can then be written as a 150×4 matrix $\boldsymbol{X} \in \mathbb{R}^{150 \times 4}$:

$$\begin{bmatrix} x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & x_4^{(1)} \\ x_1^{(2)} & x_2^{(2)} & x_3^{(2)} & x_4^{(2)} \\ \vdots & \vdots & \vdots & \vdots \\ x_1^{(150)} & x_2^{(150)} & x_3^{(150)} & x_4^{(150)} \end{bmatrix}$$

A roadmap for building model



Preprocessing – getting data into shape

- find meaningful features from raw data
- feature scaling like normalization
- dimensionality reduction
- randomly divide dataset to seperate train/test set

Training and selecting a predictive model

- various algorithm, various model form
- use of cross-valiation : validation set
- use of metrics to compare
- use of hyperparameter opt for find-tuning

Evaluating models and predicting unseen data instances

- use test set to estimate generalization error
- optimisitc

Using python for machine learning

- numpy, scipy

    - opt for fast and vectorized operations on multi-array
- pandas
    - opt for tabular data
- matplotlib
    - for viz
- conda