Machine Learning Techniques

DATASCI 420

Lesson 01-1: Intro to Machine Learning

Some Machine Learning References

- General
  - Jiawei Han, Data Mining: Concepts and Techniques, (The Morgan Kaufmann Series in Data Management Systems)
  - Tom Mitchell, *Machine Learning*, McGraw Hill, 1997
  - Christopher Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1995
- Adaboost
  - Friedman, Hastie, and Tibshirani, “Additive logistic regression: a statistical view of boosting”, Annals of Statistics, 2000
- SVMs
  - http://www.support-vector.net/icml-tutorial.pdf

Supervised vs Unsupervised Machine Learning

Supervised vs. Unsupervised Learning

- Supervised learning: classification is seen as supervised learning from examples.
  - Supervision: The data (observations, measurements, etc.) are labeled with pre-defined classes. It is like that a “teacher” gives the classes (supervision).
  - Test data are classified into these classes too.

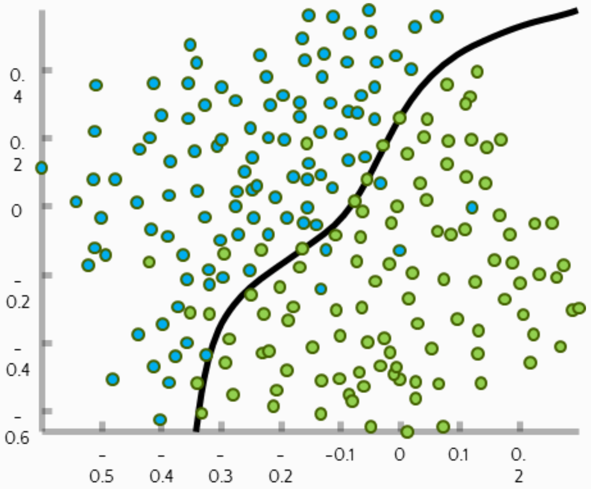
- Unsupervised learning (clustering)
  - Class labels of the data are unknown
  - Given a set of data, the task is to establish the existence of classes or clusters in the data

Supervised Machine Learning

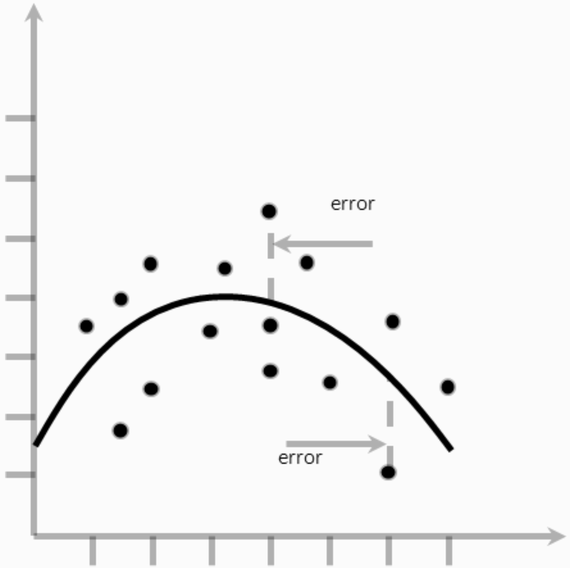
- Machine learning tasks where training data has labels
  - Examples:
    - Fraud transaction: we know which transactions in the training data were fraud (1), which were not (0)
    - Readmission: we know which patients were readmitted to hospital within a certain time window after discharge
    - Recommendation: we know which items were presented to customers, and which items were clicked, added to cart, or purchased.

Three typical supervised machine learning tasks: Classification, Regression and Recommendation

Classification



Regression



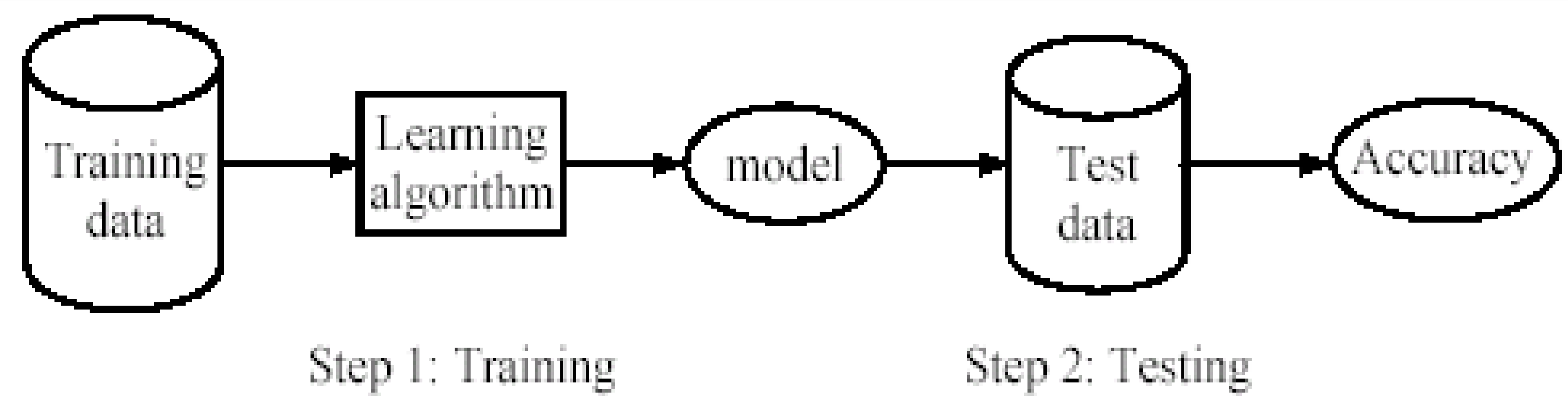
Recommenders

Supervised learning process: two steps

Learning (training): Learn a model using the training data

Testing: Test the model using unseen test data to assess the model accuracy

$$Accuracy = \frac{\text{Number of correct classifications}}{\text{Total number of test cases}},$$



What do we mean by learning?

- Given
  - a data set *D*,
  - a task *T*, and
  - a performance measure *M*,

  a computer system is said to **learn** from *D* to perform the task *T* if after learning the system's performance on *T* improves as measured by *M*.
- In other words, the learned model helps the system to perform *T* better as compared to no learning.

Fundamental assumption of learning

Assumption: The distribution of training examples is identical to the distribution of test examples (including future unseen examples).

- In practice, this assumption is often violated to certain degree.
- Strong violations will clearly result in poor classification accuracy.
- To achieve good accuracy on the test data, training examples must be sufficiently representative of the test data.

The machine learning framework

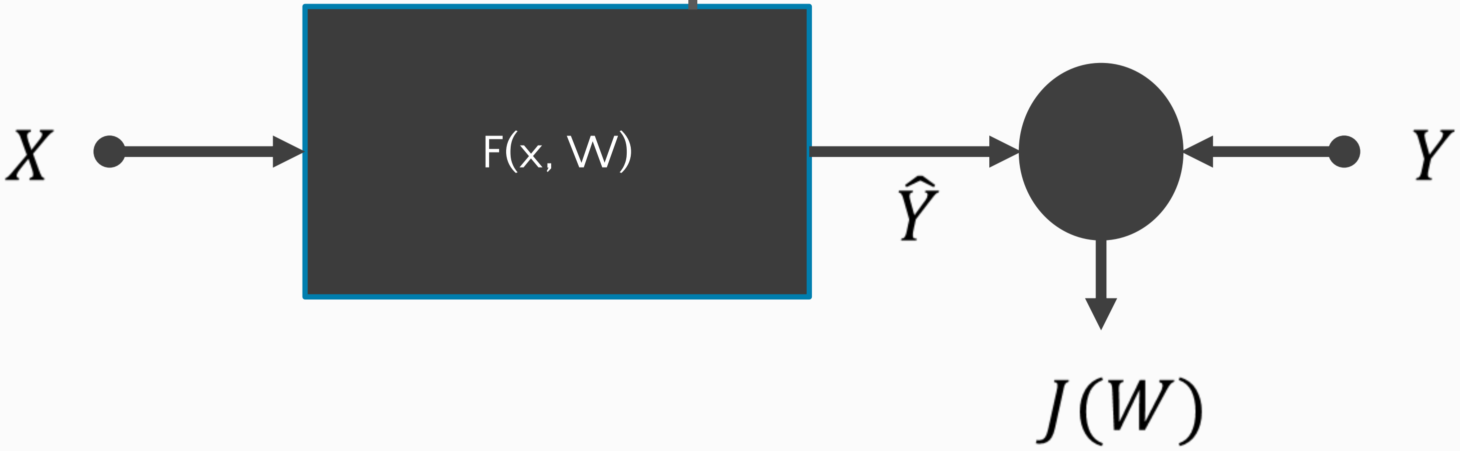
- $$\underset{\substack{\uparrow\\ \text{Observed}\\ \text{dependent}\\ \text{variable}}}{y} = \underset{\substack{\uparrow\\ \text{prediction}\\ \text{function}}}{f}_{\theta}(\underset{\substack{\nwarrow\\ \text{Independent}\\ \text{variables}}}{x}) + \underset{\substack{\nwarrow\\ \text{Random}\\ \text{noise}}}{\varepsilon}$$

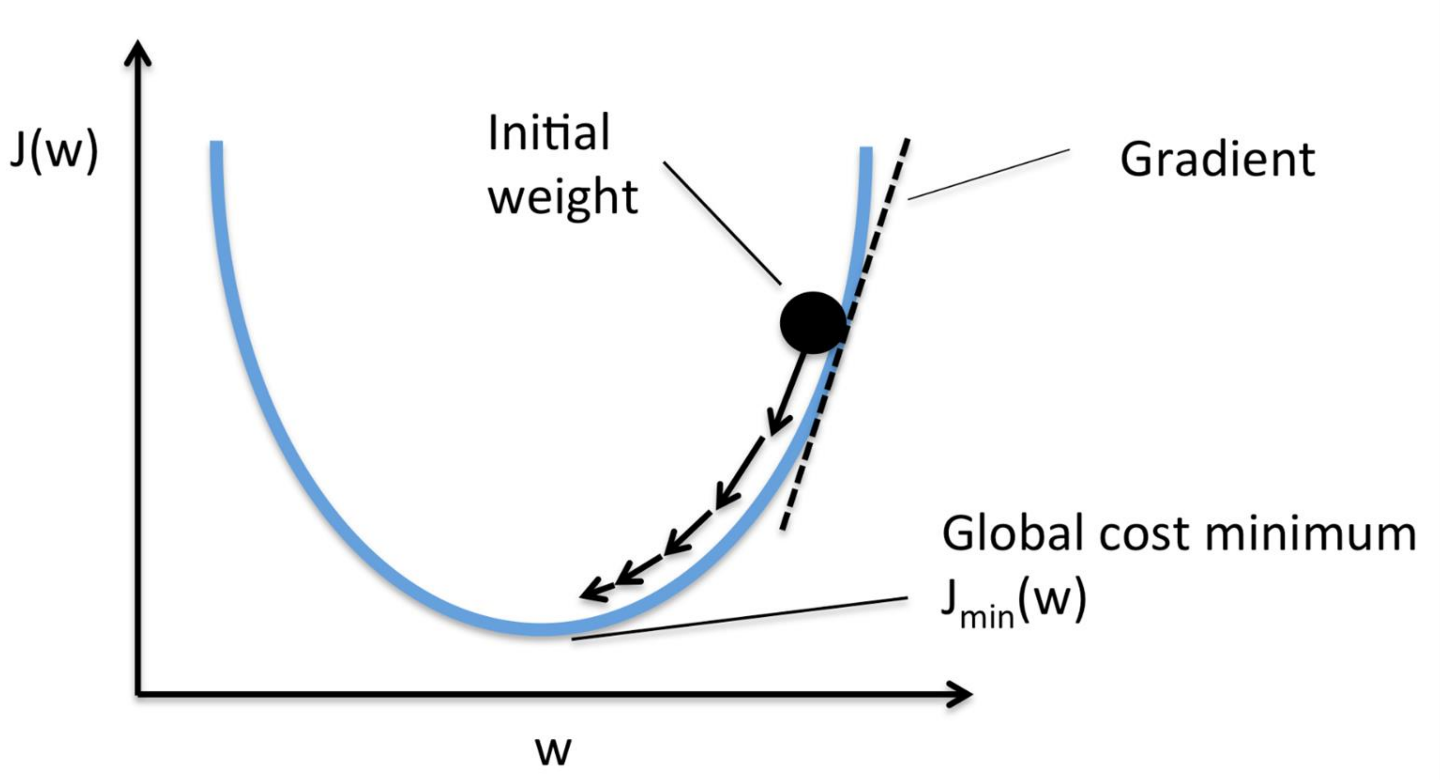
- Training: given a *training set* of labeled examples $\{(x_1,y_1), \ldots, (x_N,y_N)\}$, estimate the prediction function f and parameters θ which minimizes the prediction error on the training set

$$E_{\theta}(Y,X) = \sum_{i=1}^{N}\left(y_i - \hat{f}_{\theta}(x_i)\right)^2$$

- Testing: apply f to a never before seen *test example* x and output the predicted value $y = f(x)$

How to learn model parameters θ?





Many classifiers to choose from

- SVM
- Neural networks
- Naïve Bayes
- Bayesian network
- Logistic regression
- Randomized Forests
- Boosted Decision Trees
- K-nearest neighbor
- Etc.

Which is the best one?