Analytics and Visualization of Big Data

Fadel M. Megahed

Lecture 25: Machine Learning (Cont.)

AUBURN UNIVERSITY
SAMUEL GINN
COLLEGE OF ENGINEERING

Department of Industrial and Systems Engineering

Spring 13

Refresher: The Machine Learning Model – The Training Set

- The data to which ML algorithm is applied is called a **training set**.

- A training set consists of a set of pairs (x, y), where
  - x is a vector of values, often called a feature vector.
  - y is the label, the classification value for x.

- **The objective of the ML process is to discover a function $y = f(x)$ that best predicts the value of y associated with unseen values of x.**

Perceptrons

- A perceptron is a linear binary classifier.

- Each perceptron has a threshold θ. The output of the perceptron is: +1 if w.x > θ, and is −1 if w.x < θ.
  - **The special case where w.x = θ will be regarded as "wrong".**

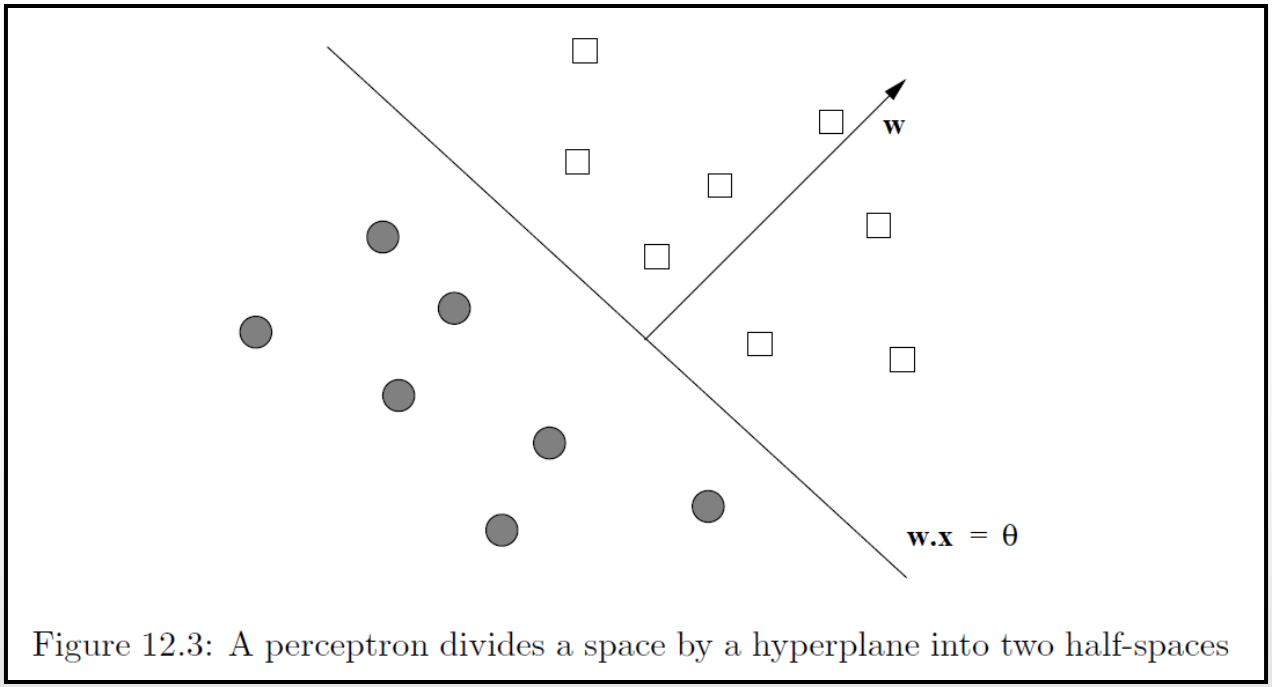
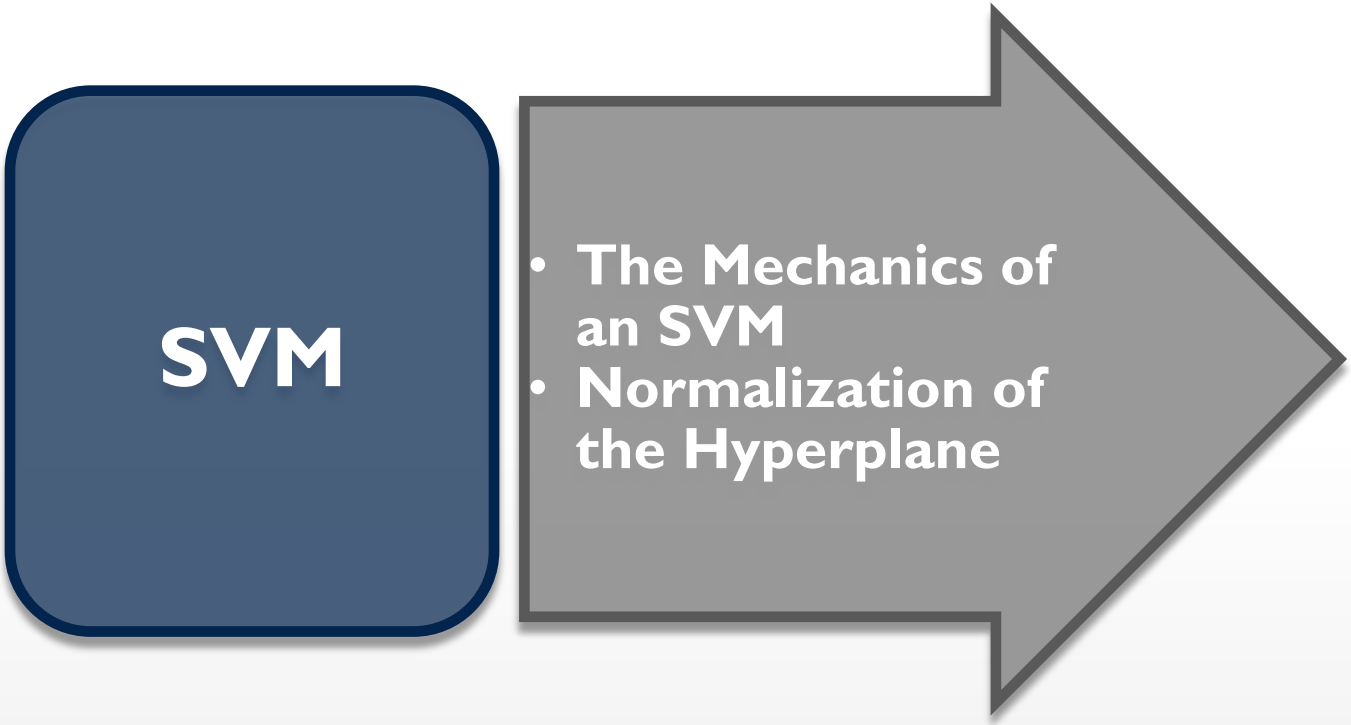


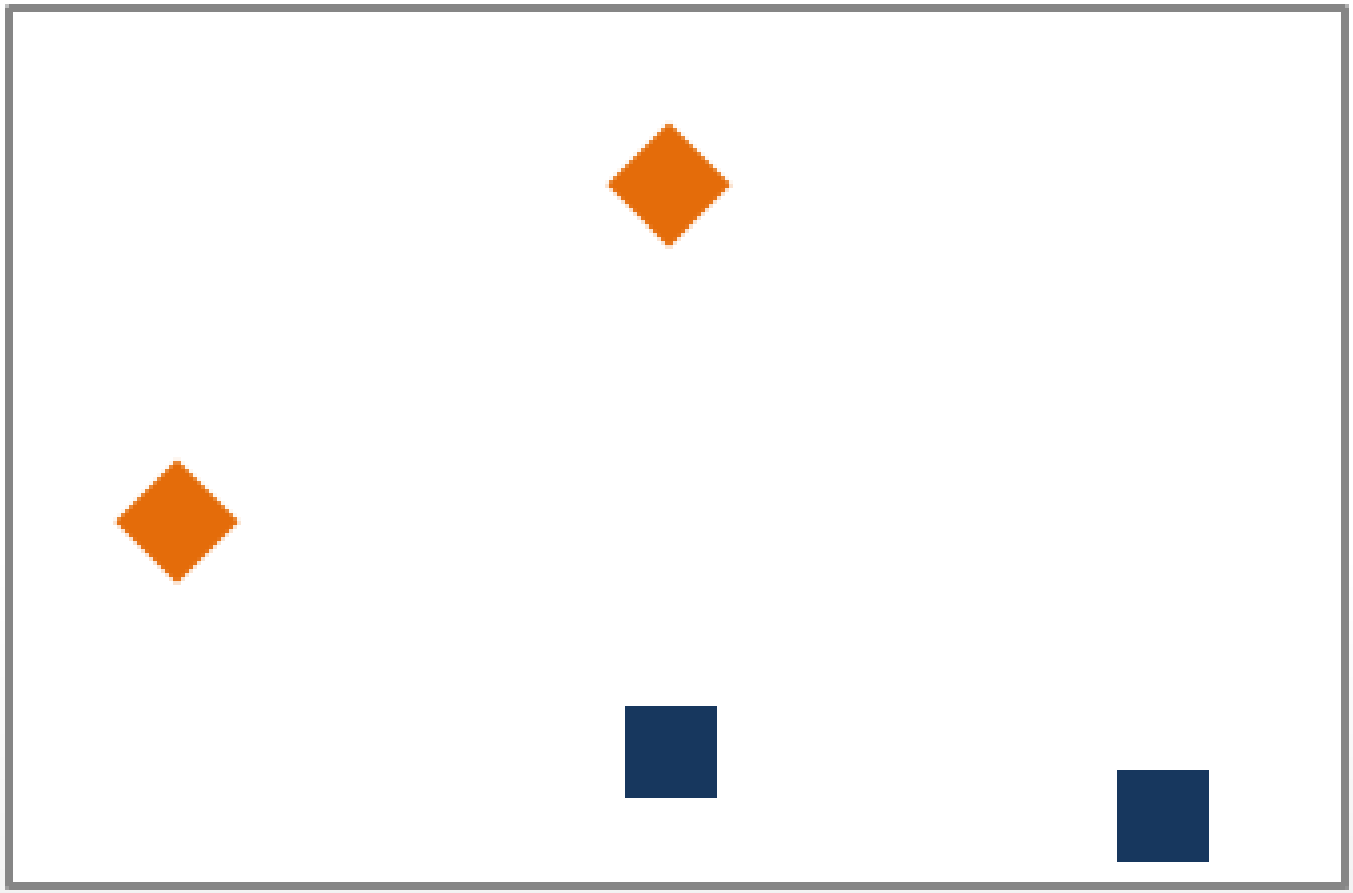
Figure 12.3: A perceptron divides a space by a hyperplane into two half-spaces

Outline for Today's Topics – Chapter 12



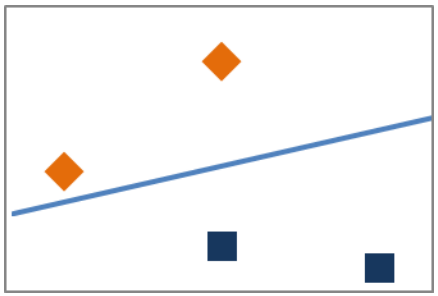
Chapter link: http://i.stanford.edu/~ullman/pub/ch12.pdf

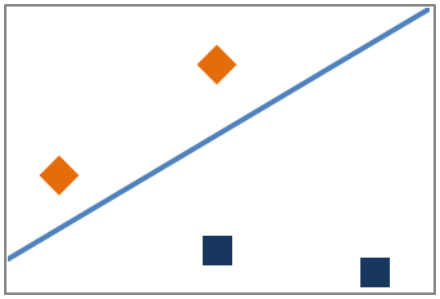
Better Line Separator – Data

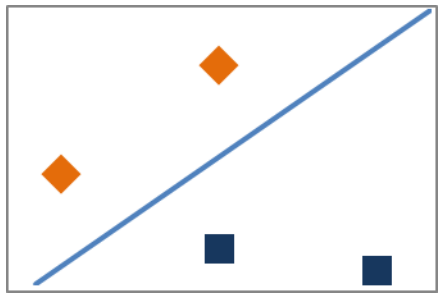


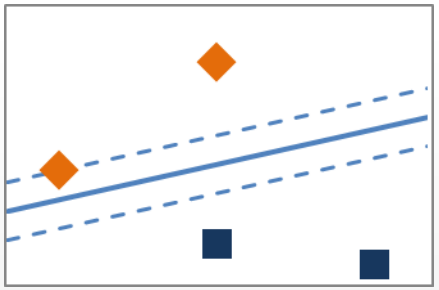
x	y
1	4
4	8
4	1.3
7	0.5

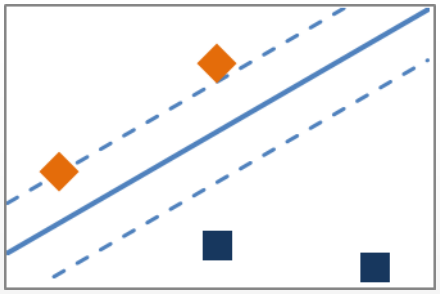
Better Line Separator – Data

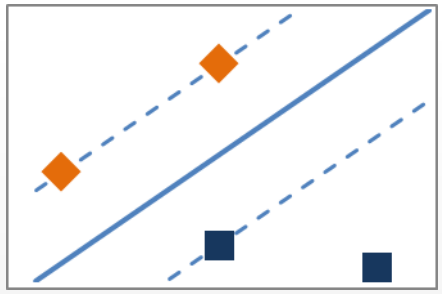












Questions: What is the best line? Why?

Main Questions to be addressed Today

1. **Why is the bigger margin better?**
   a. We addressed this question intuitively
   b. For the purpose of this class, this is sufficient. If you want to blog about the mathematical reasons for that, you are definitely welcome to ☺

2. **Which w maximizes the margin?**

The Mechanics of an SVM