Linear Machines and SVM – Part 1: Linear Machines Basics

Ira A. Fulton Schools of Engineering
Arizona State University

Objective

Objective

Define general linear classifiers

Revisiting Logistic Regression

In Logistic Regression: given a training set of n labelled samples $<\mathbf{x}^{(i)}, y^{(i)}>$, we learn $P(y|\mathbf{x})$ by assuming a logistic sigmoid function.

➔ We end up with a *linear classifier*.

➔ $g(\mathbf{x}) = \mathbf{w}^t\mathbf{x}$ is called the *discriminant function*.

Linear Discriminant Functions

- In general, taking a discriminative approach, we can *assume* some form for the discriminant function that defines the classifier.

  ➔ The learning task is to use the training samples to estimate the parameters of the classifier.

Linear Decision Boundaries

- Linear discriminant functions give arise to liner decision boundaries

➔ *linear classifiers* or *linear machines*

- We will use both notations:

$$g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} \qquad \textit{or} \qquad g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + w_0$$

Linear Machine for C>2 Classes

- We can define C linear discriminant functions:

$$g_i(\mathbf{x}) = \mathbf{w}_i^t\mathbf{x}, \quad i = 1, 2, \ldots, C$$

- What is the decision rule for the classifier?

The Learning Task

- Finding $\mathbf{w}_i$, $i = 1, 2, \ldots, C$

- Let's use the 2-class case as an example
  - For n samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$, of 2 classes ω_1 and ω_2 , **if** there exists a vector **w** such that $g(\mathbf{x}) = \mathbf{w}^t\mathbf{x}$ classifies them all correctly ➔ Finding **w**

  i.e., finding **w** such that

$$\mathbf{w}^t\mathbf{x}_i \geq 0 \text{ for samples of } \omega_1 \quad \text{and}$$
$$\mathbf{w}^t\mathbf{x}_i < 0 \text{ for samples of } \omega_2,$$

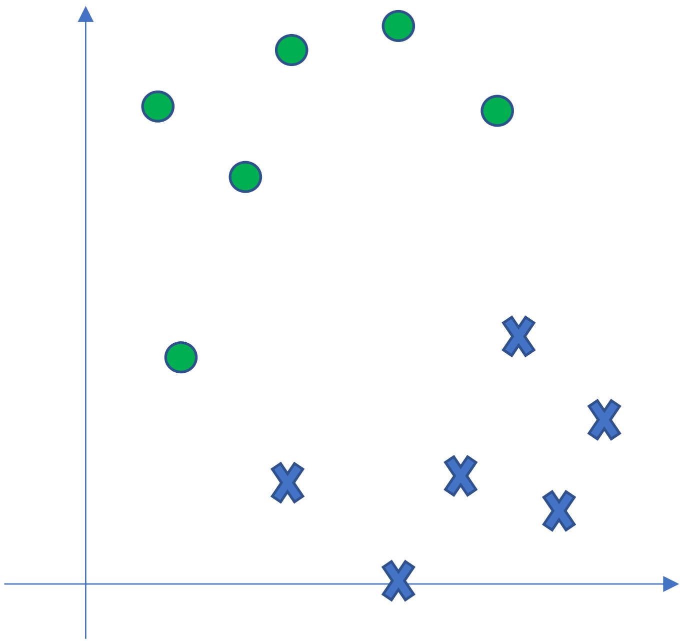
Linear Separability

- If we can find at least one vector $\mathbf{w}$ such that $g(\mathbf{x}) = \mathbf{w}^t\mathbf{x}$ classifies all samples
  - ➔ We say the samples are linearly separable.

- An example of not linearly separable in 2-D:

The Solution Region

- There may be many different weight vectors that can all be valid solutions for a given training set
  - ➔The solution regions



- If the solution vector is not unique, *Which one is the best?*

Solving for the Weight Vector

- Consider the following approach: finding a solution vector which optimizes some objective function.

➔ We may introduce additional constraints for a “good” solution”

➔ **Solving a constrained optimization problem.**

- Theoretical: Lagrange or Karush-Kuhn-Tucker.

- In practice: e.g., gradient-descent-based search

Gradient Descent Procedure

- Basic idea:
  - Define a cost function $J(\mathbf{w})$
  - Starting from an initial weight vector $\mathbf{w}(0)$
  - Update $\mathbf{w}$ by

$$\mathbf{w}(k+1) = \mathbf{w}(k) - \eta(k)\nabla J\big(\mathbf{w}(k)\big),$$

- η>0 is the *learning rate.*