Lab #9: Support Vector Machines

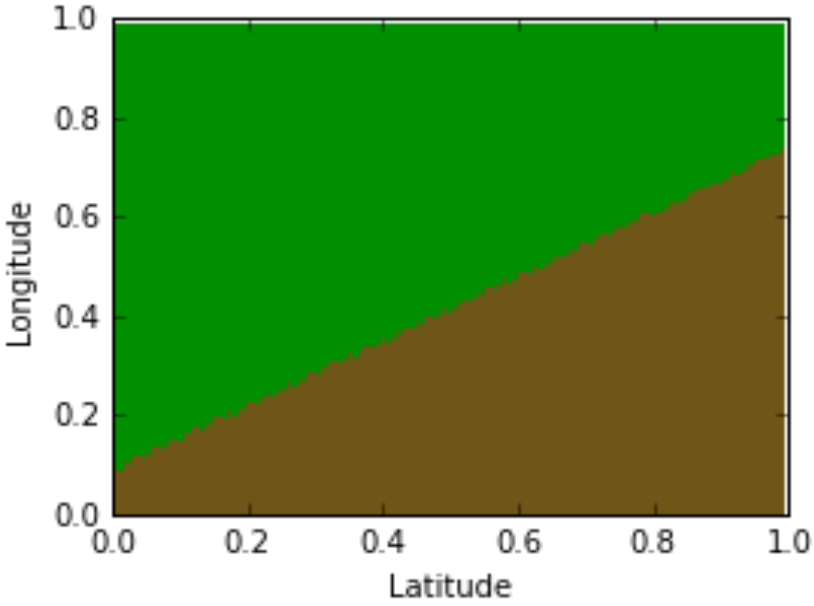
CS 109A, STAT 121A, AC 209A: Data Science

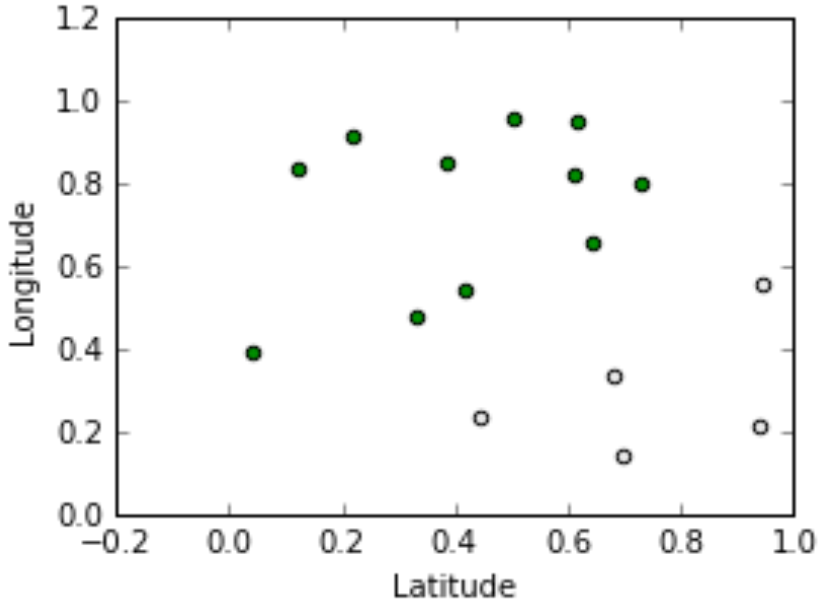
Fall 2016

Harvard

Today's lab: Problem 1

Problem 1: Detecting vegetation cover





Use a support vector machine classifier to separate land and vegetation

So far…

- Linear models
  - Logistic Regression
  - Linear Discriminant Analysis

- Non-linear models
  - Quadratic Discriminant Analysis
  - Decision trees

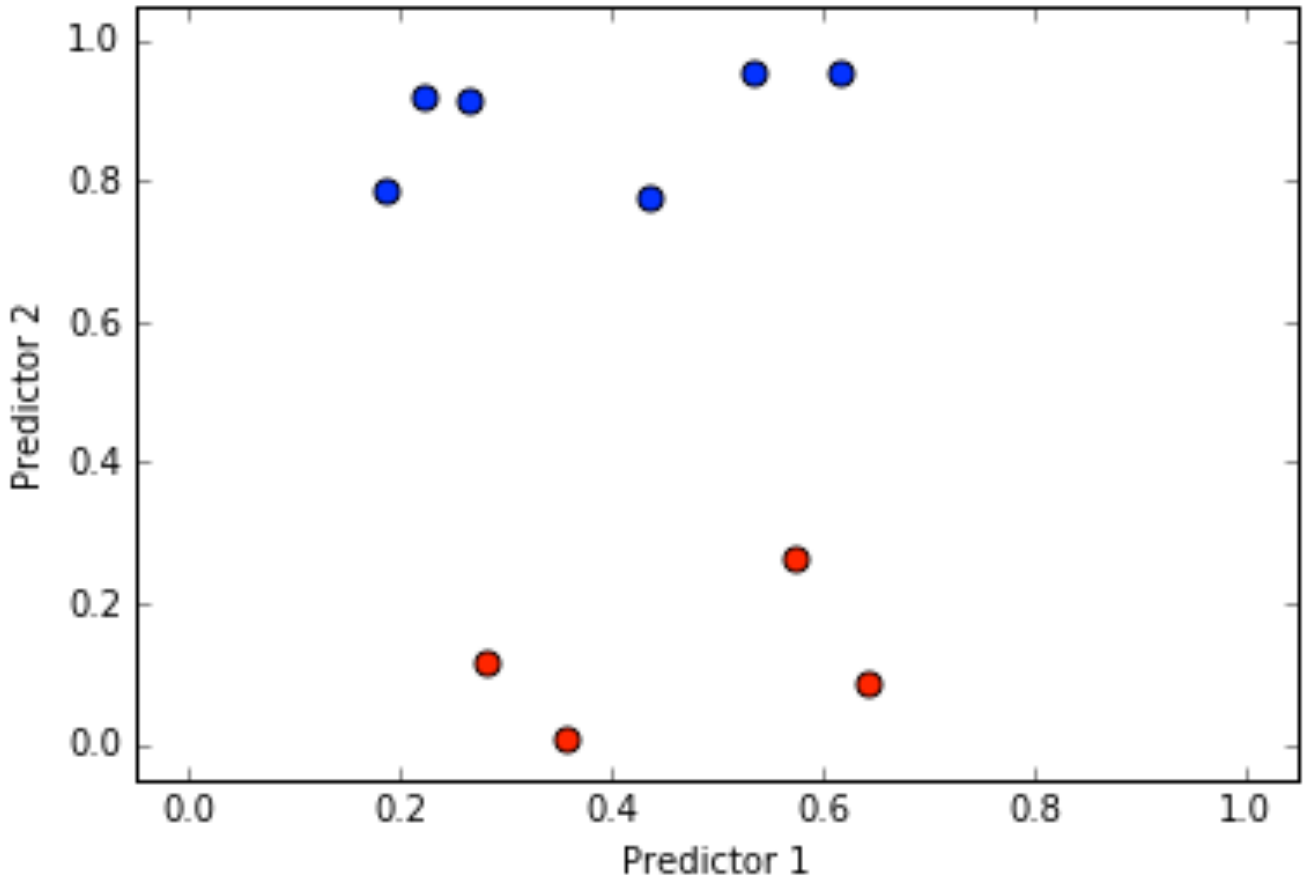
So far…

- Linear models
  - Logistic Regression
  - Linear Discriminant Analysis

This Lab:
Support Vector Machines (SVM)

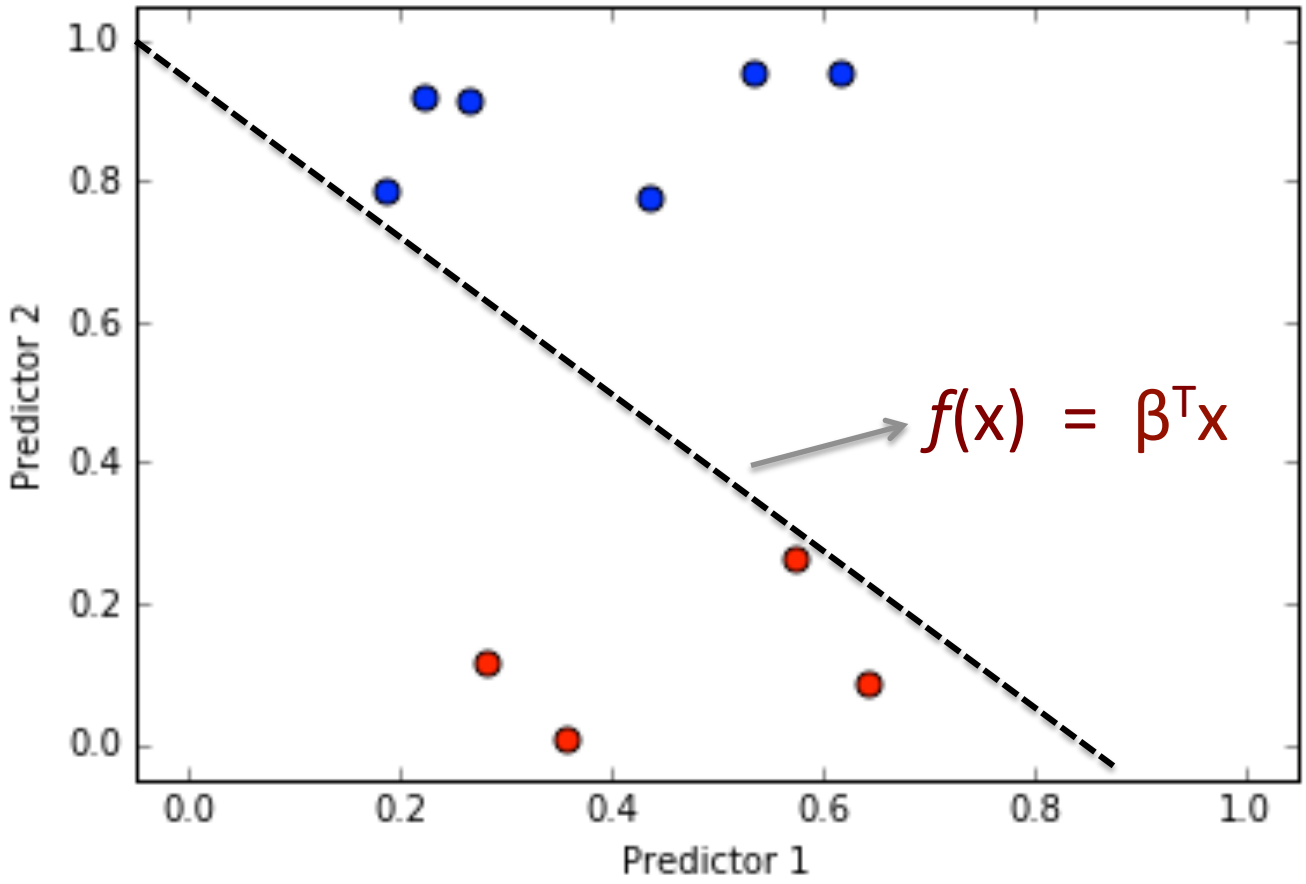
- Non-linear models
  - Quadratic Discriminant Analysis
  - Decision trees

Example 1



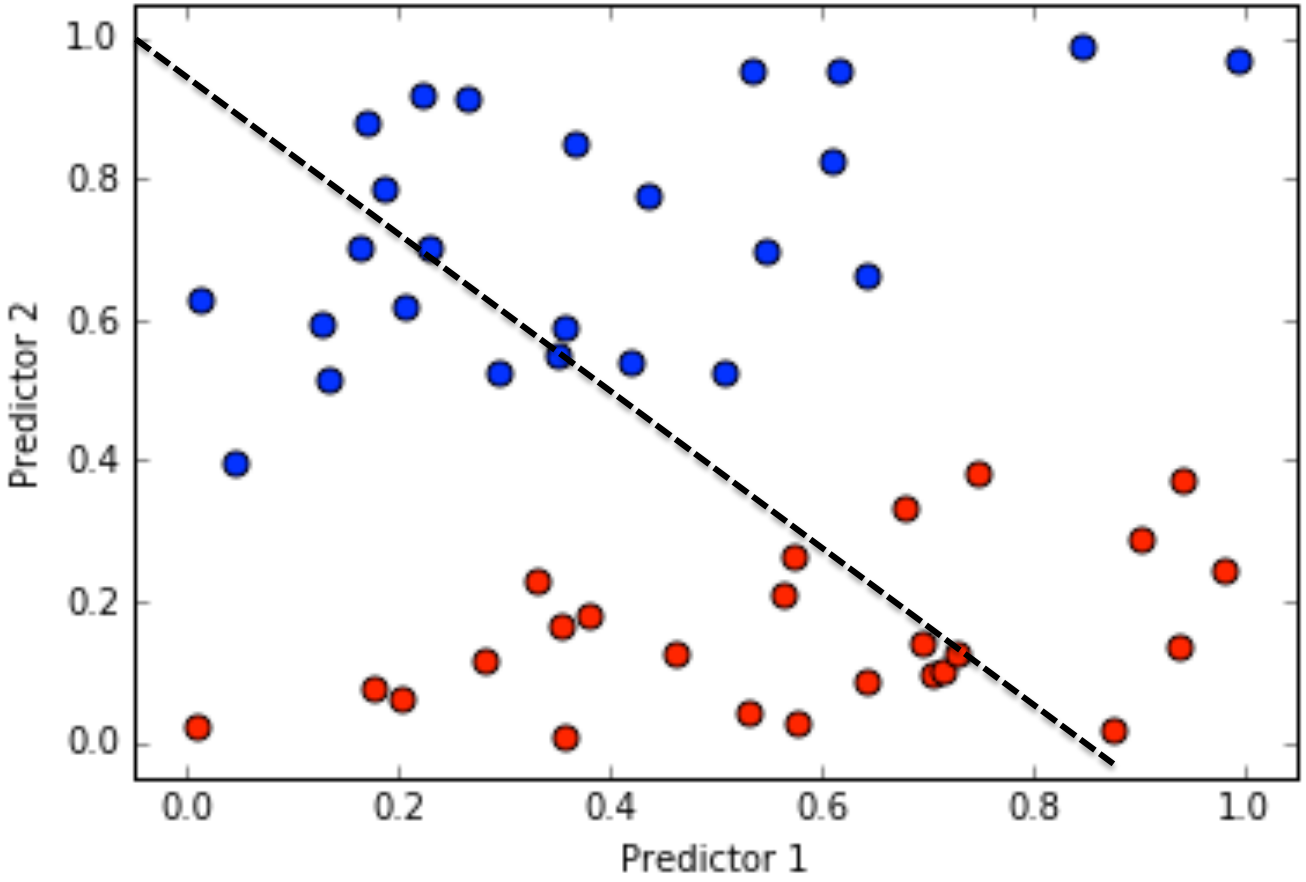
Training set

Example 1: Model A



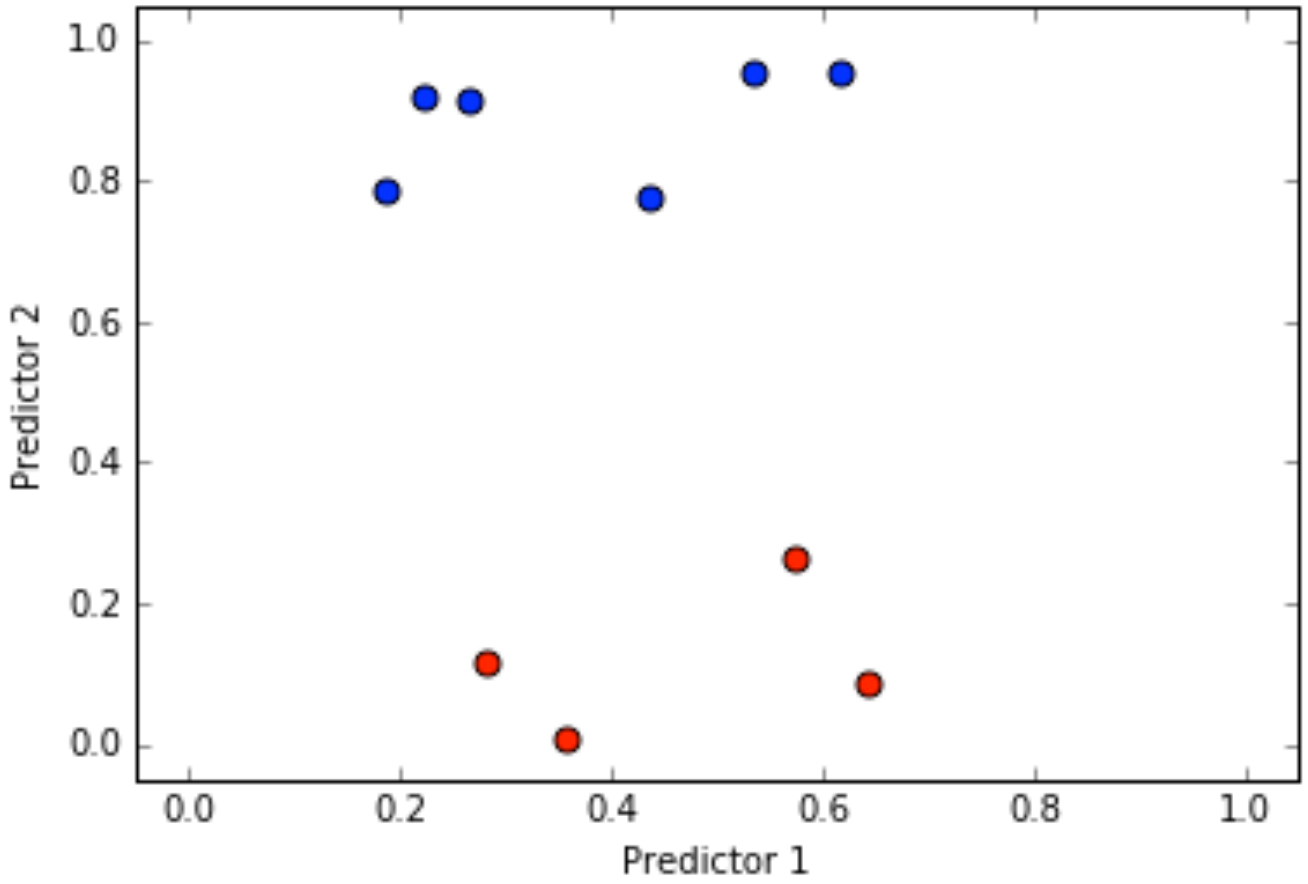
Training set

Example 1: Model A



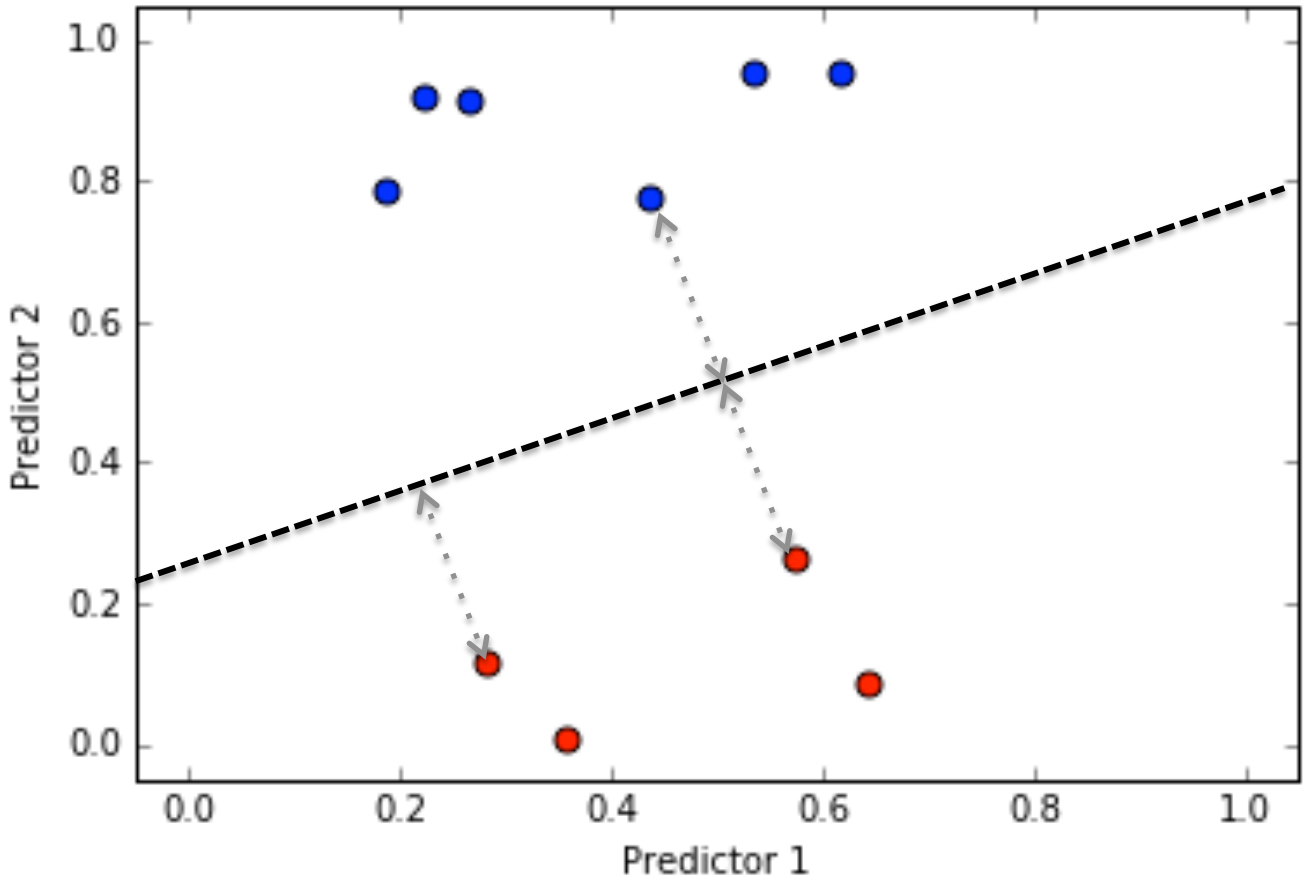
Training + Test set

Example 1



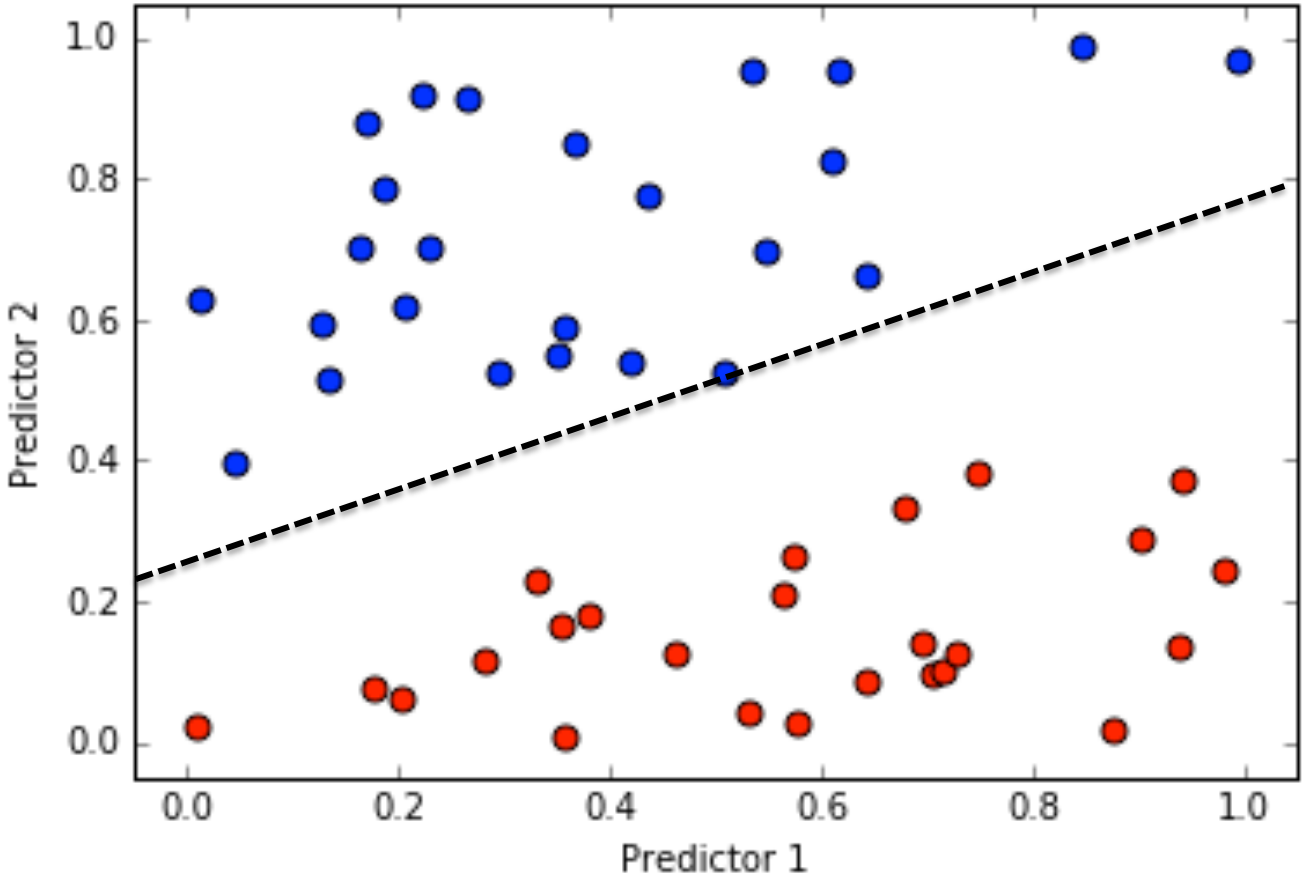
Training set

Example 1: Model B



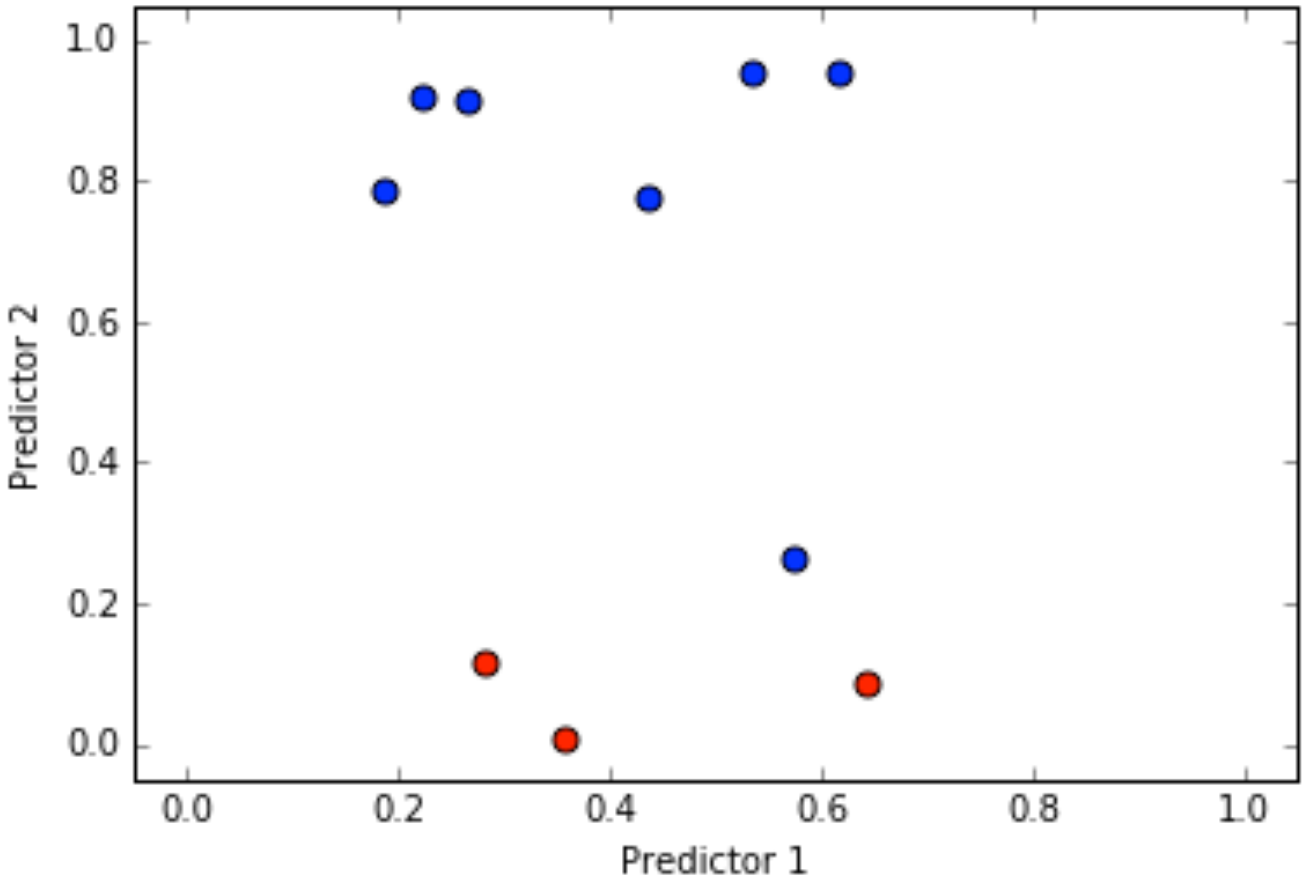
Training set

Example 1: Model B



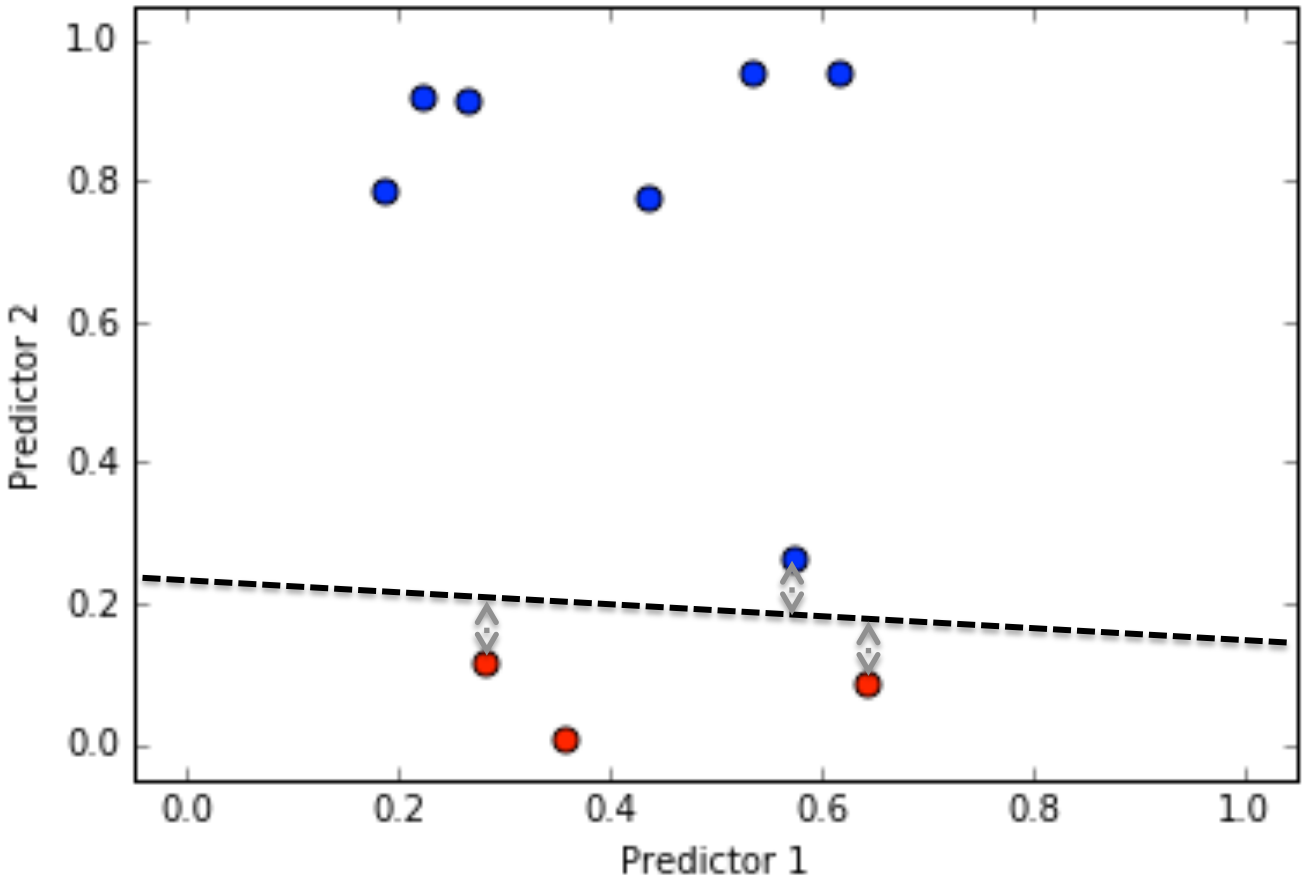
Training + Test set

Example 2



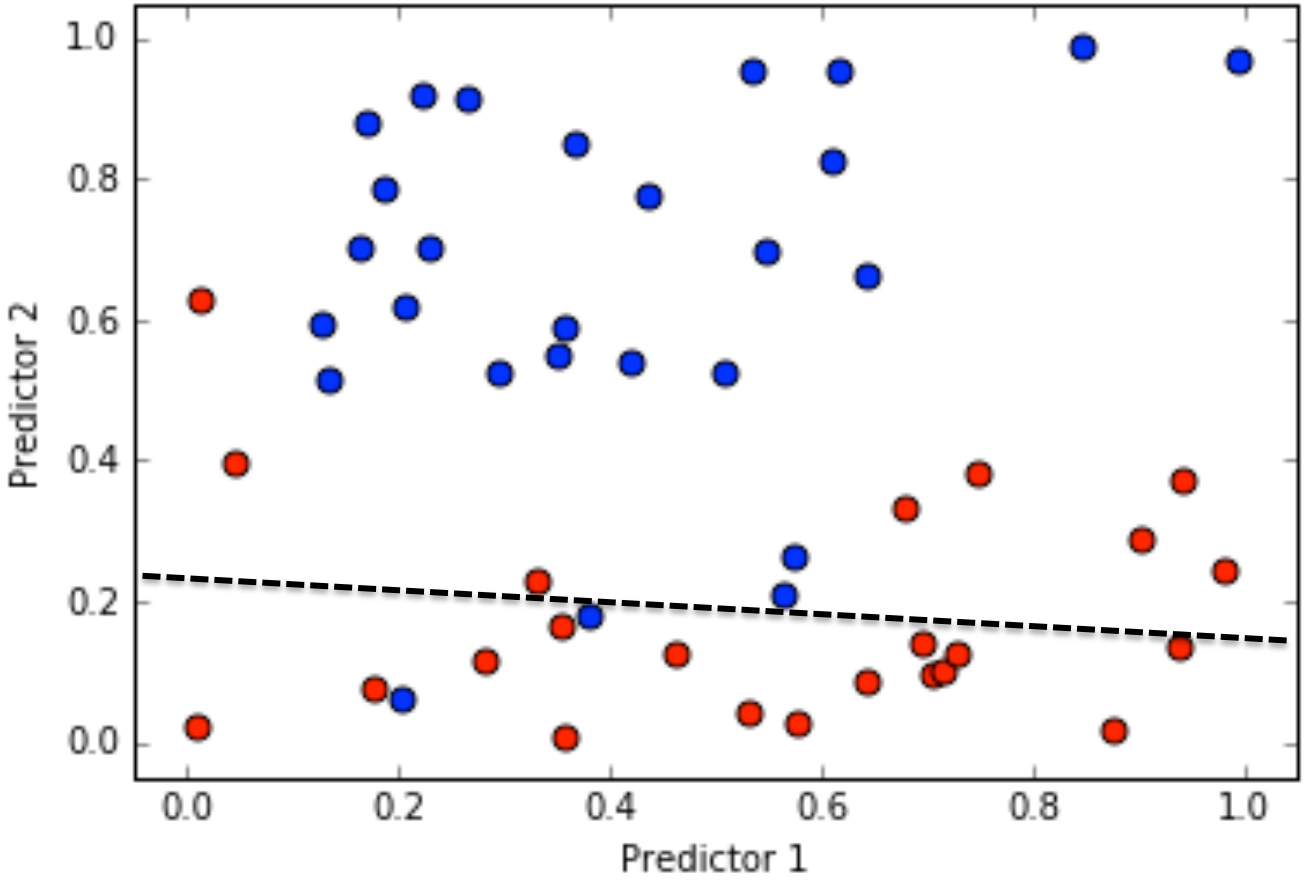
Training set

Example 2: Model A



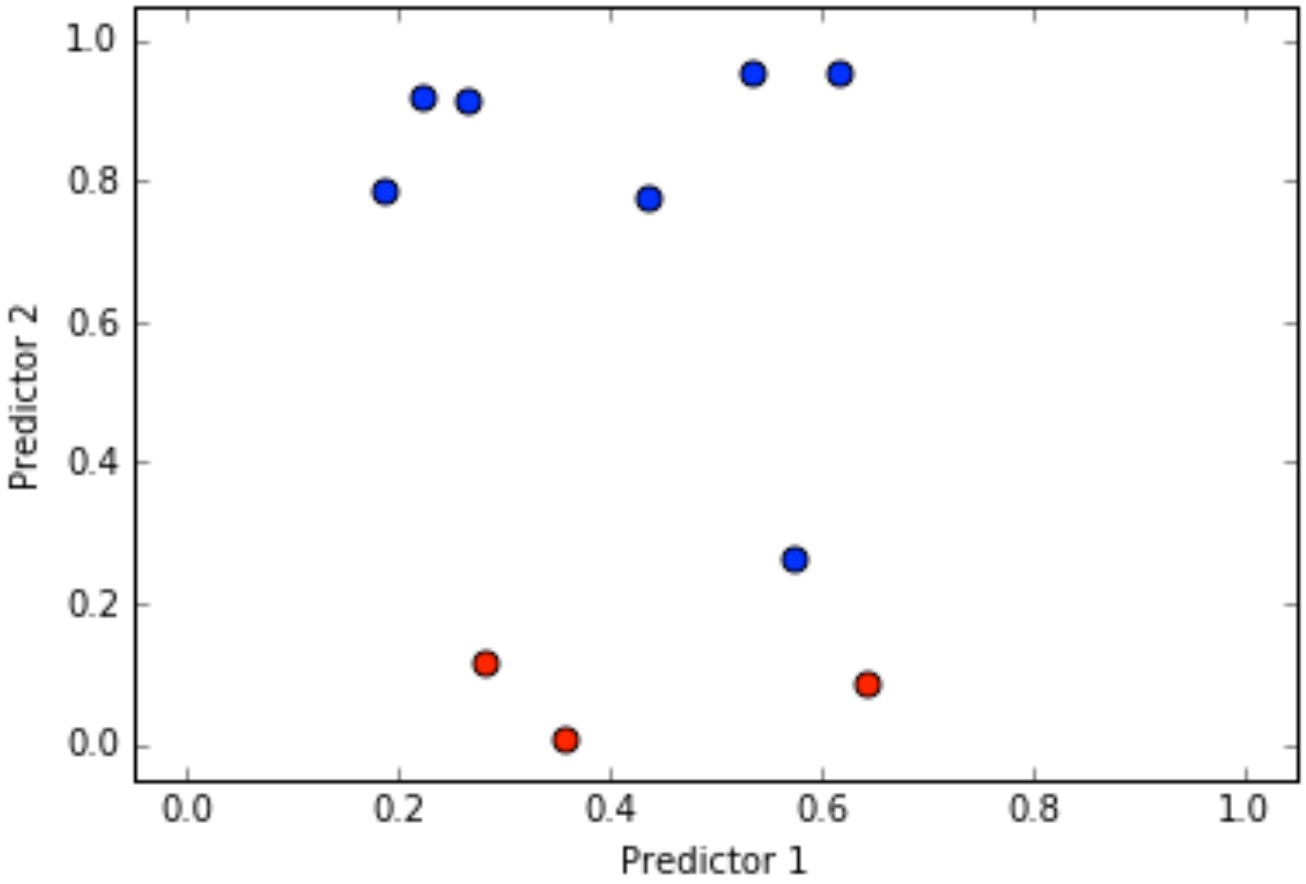
Training set

Example 2: Model A



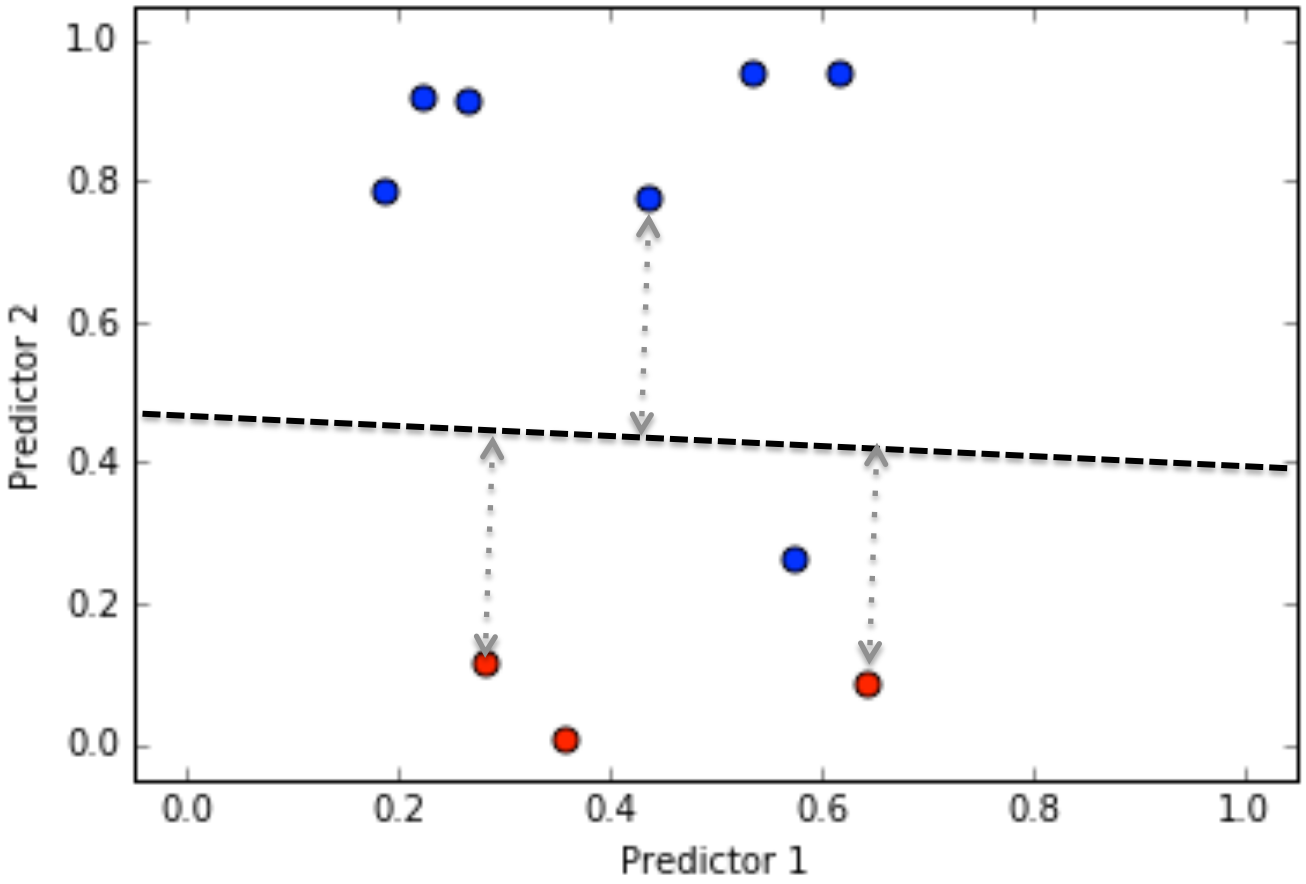
Training + Test set

Example 2



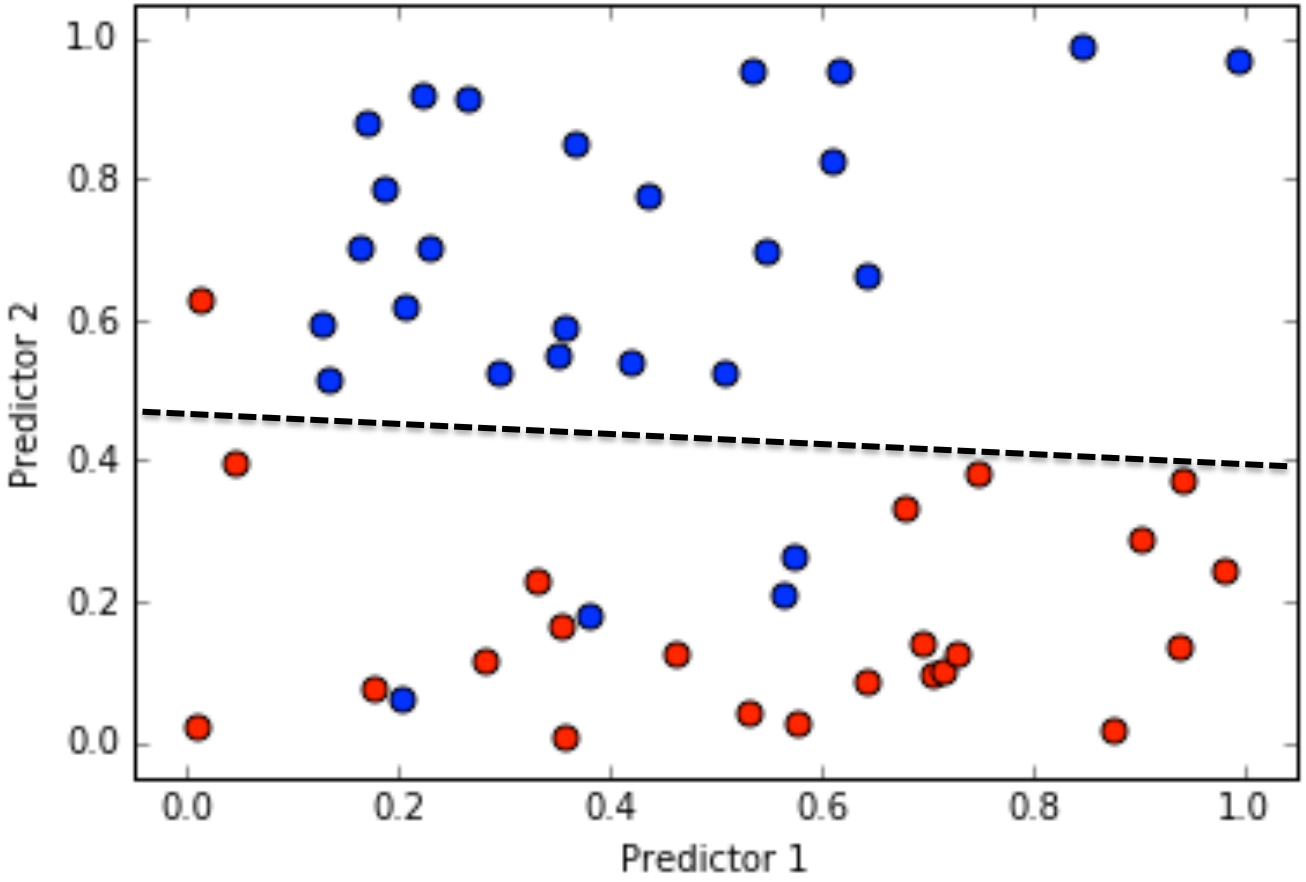
Training set

Example 2: Model B



Training set

Example 2: Model B



Training + Test set

SVM Objective

- Trade-off between accuracy and separation:

$$C \times \text{Accuracy} + \text{Margin of separation}$$

SVM Objective

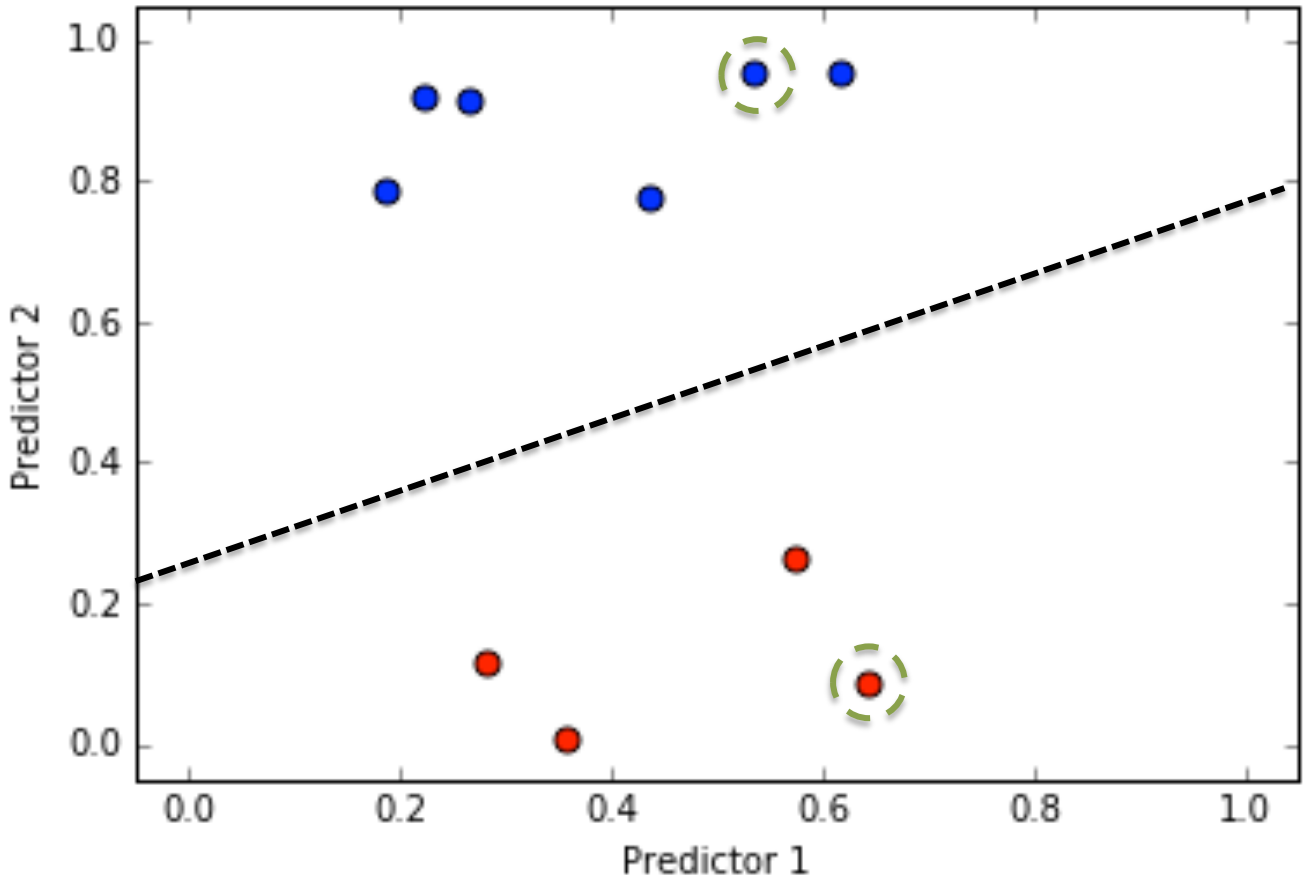
- Trade-off between accuracy and separation:

$$C \times \text{Accuracy} + \text{Margin of separation}$$

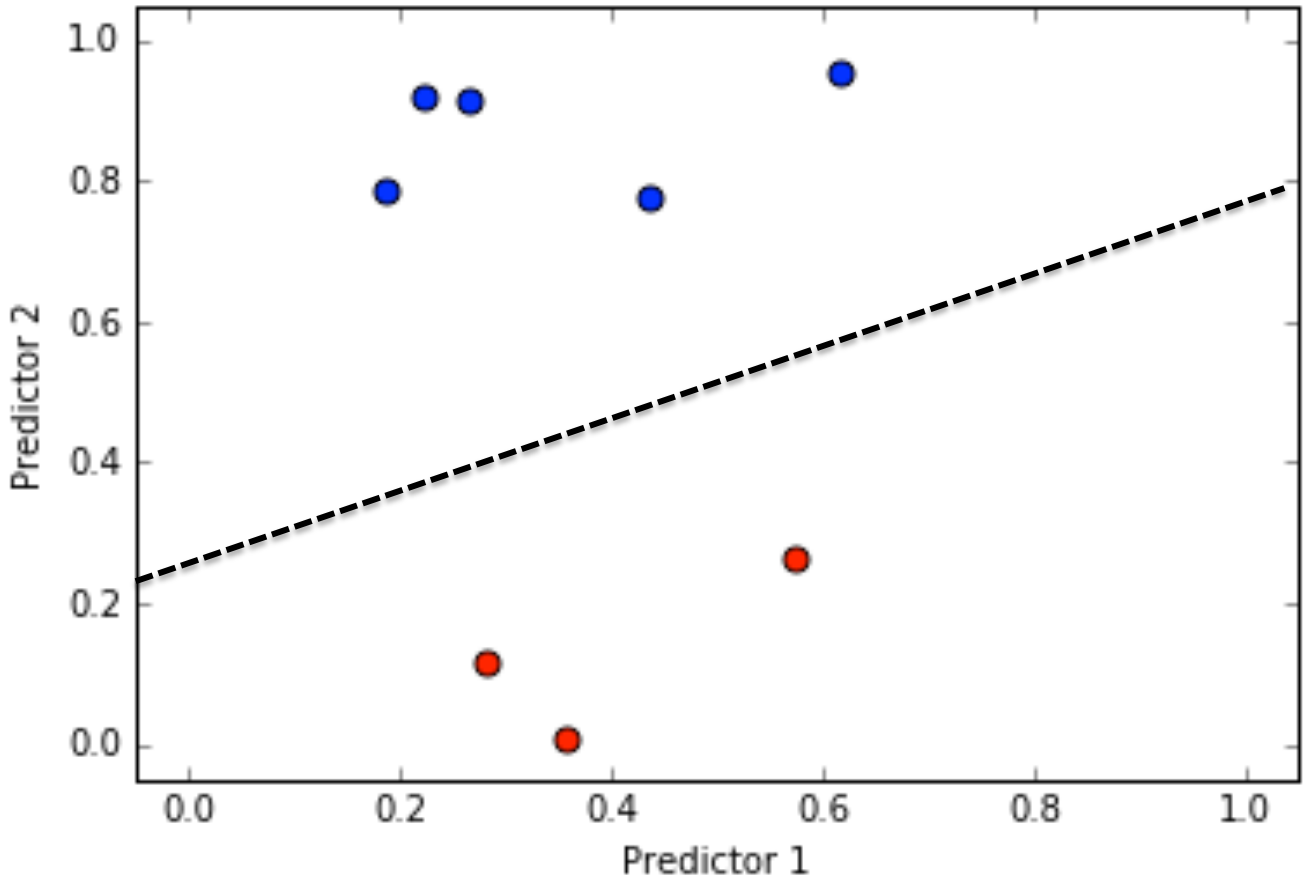
- $C > 0$ controls the trade-off:
  - larger C means larger emphasis on accuracy

- *How can one optimize this objective?*

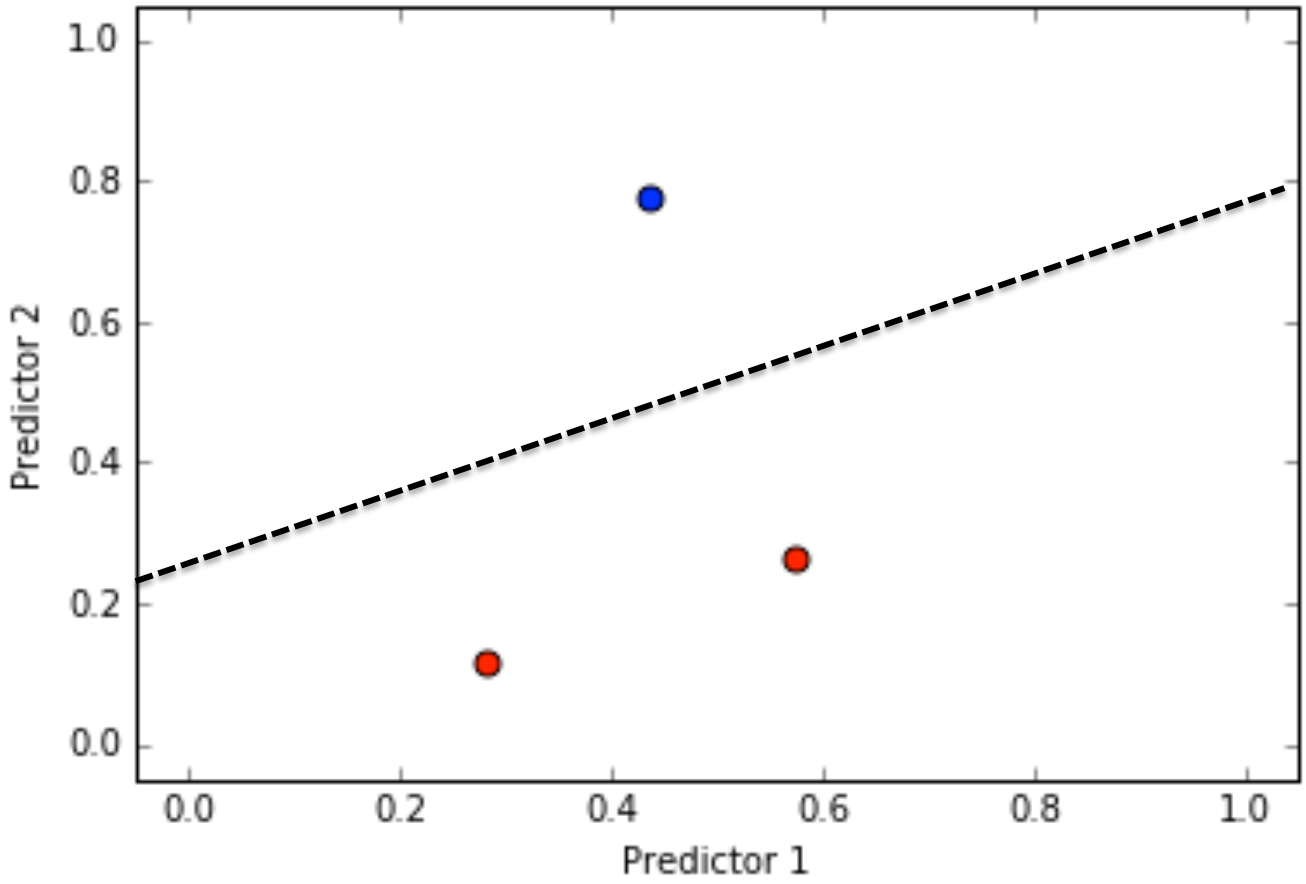
Revisiting Example 1



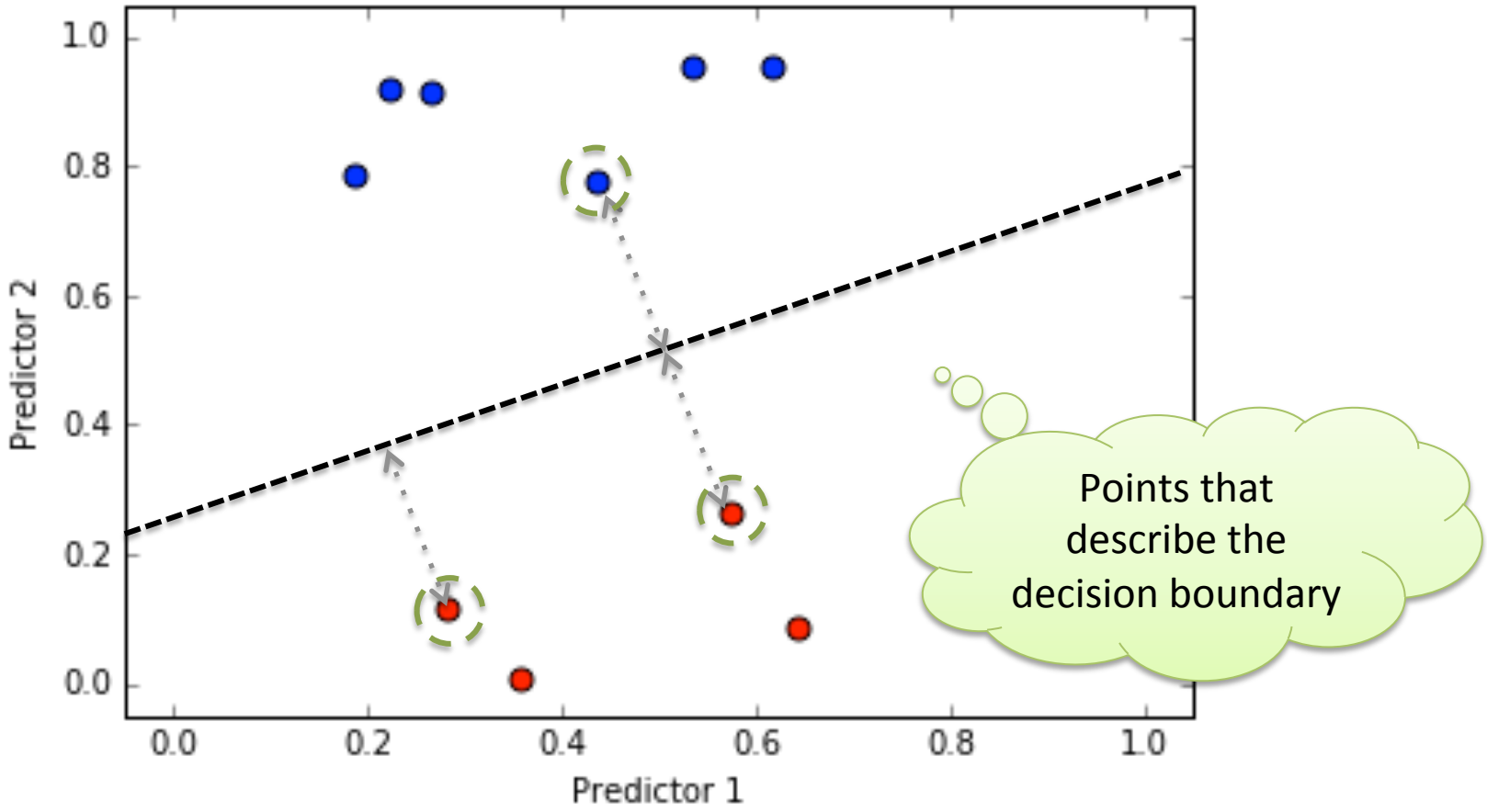
Revisiting Example 1



Revisiting Example 1



Revisiting Example 1



Support vectors!

SVM Model Fitting

- Search for right subset of support vectors

- Intuitively, SVM searches over subsets of points, and picks the one that describes a decision boundary with minimum loss

Questions

- How do the support vectors effect the decision boundary?

- How does the parameter '*C*' influence the decision boundary?