

Lecture 19: Support Vector Machines

CSI 5v93: Introduction to machine learning

Baylor University
Computer Science Department

Dr. Greg Hamerly
<http://cs.baylor.edu/~hamerly/>

CSI 5v93: Introduction to machine learning, Lecture 19 – p. 1/15

Questions?

CSI 5v93: Introduction to machine learning, Lecture 19 – p. 2/15

Announcements

- choose a paper to present, email me your choice

The next topic

Leaving naive Bayes, we now (re)consider real-valued data.

We are still concerned with classification.

We will be considering linear decision boundaries.

Ultimately we will learn about the Support Vector Machine, which is one of the most important and powerful recent developments in machine learning.

Separating hyperplanes

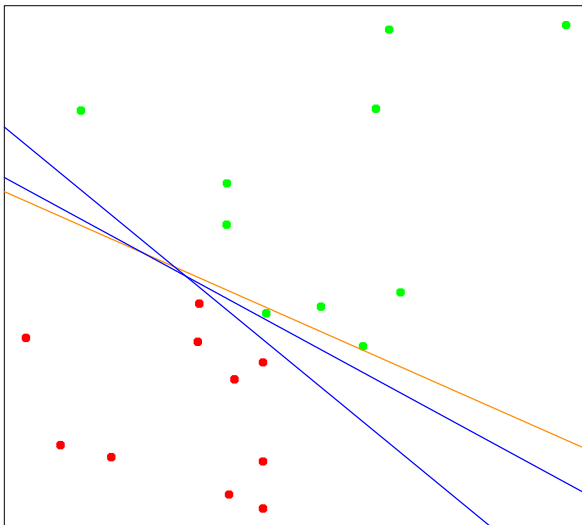
Remember: linear discriminant analysis and logistic regression.

These classifiers use linear boundaries between different classes, but the linear decision boundaries were artifacts of the models used for the data.

We now turn to the topic of separating hyperplanes, where the goal is not to model the data within classes, but to find the best (linear) separation boundary between classes.

Separating hyperplanes

Here is a two-class classification problem, with several linear boundaries:



Note that there are many solutions to use a linear boundary to separate the classes.

Some linear algebra for hyperplanes

Like in linear regression, we have a vector of parameters $\beta \in \mathbb{R}^d$

The β and β_0 define a hyperplane (actually affine set) as

$$f(x) = \beta_0 + \beta^T x = 0$$

The set of all x that fulfill this equation define the affine set, which we'll call L .

If β and x are scalars, then the linear boundary L is a line.

If β and x are 2-vectors, then L is a 2-d plane (etc.).

Hyperplane properties

For any two points x_1 and x_2 that lie on L ,

$$\beta^T (x_1 - x_2) = 0$$

Also, β is the vector that is normal to the surface of L , and the unit-length normal vector is

$$\beta^* = \beta / \|\beta\|$$

Hyperplane properties

For any point x_0 in L ,

$$\beta^T x_0 = -\beta_0$$

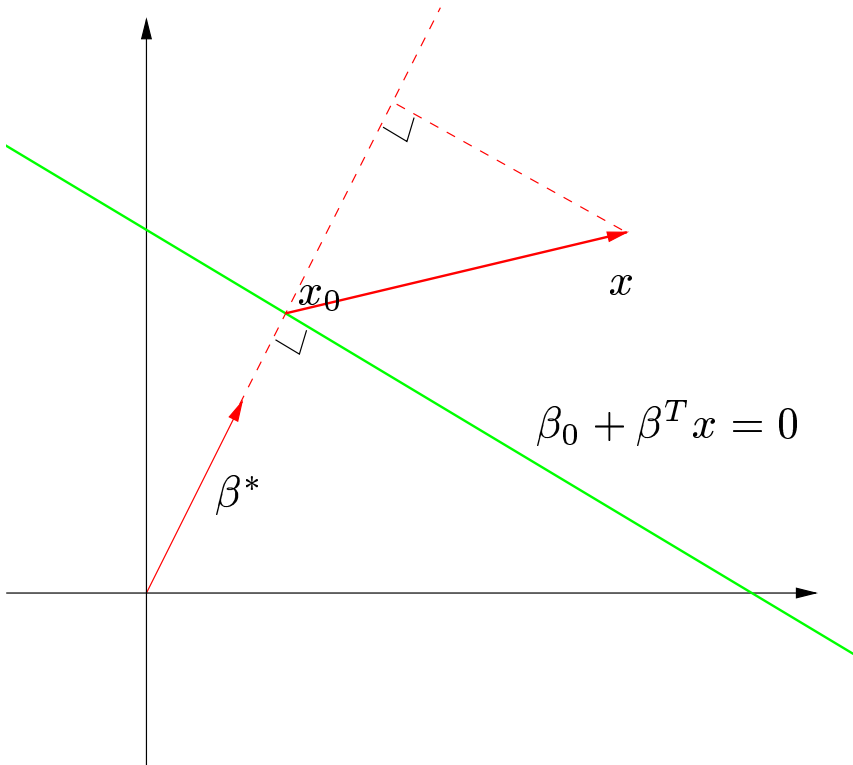
Thus β_0 is the distance of L (the linear boundary) from the origin (if β is unit length; i.e. $\beta = \beta^*$).

Hyperplane properties

The signed distance of any point x to L is

$$\begin{aligned}\beta^{*T}(x - x_0) &= \frac{1}{\|\beta\|}(\beta^T x + \beta_0) \\ &= \frac{1}{\|f'(x)\|} f(x)\end{aligned}$$

Illustration of hyperplane properties



CSI 5v93: Introduction to machine learning, Lecture 19 – p. 11/15

Basic point: hyperplanes for classification

We use the hyperplane equation to classify points to +/- class:

$$f(x) = \beta^T x + \beta_0$$

If $f(x) > 0$, then classify as + class.

Otherwise, classify as - class.

$$g(x) = \text{sign}(f(x))$$

CSI 5v93: Introduction to machine learning, Lecture 19 – p. 12/15

Optimal separating hyperplanes

The optimal separating hyperplane (Vapnik, 1996)

- separates the two classes
- maximizes the distance to the closest point from either class

This second step is called *maximizing the margin*.

What are the advantages of maximizing the margin?

Mathematically defining maximum margin

$$\max_{\beta, \beta_0, \|\beta\|=1} C$$

subject to

$$y_i(x_i^T \beta + \beta_0) \geq C, i = 1, \dots, n$$

2-minute journal

Please write a response to the following on a piece of paper and hand it in immediately. Please make it anonymous (no names). Write about:

- major points you learned today
- areas not understood or requiring clarification