

Lecture 20: 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 20 – p. 1/15

Questions?

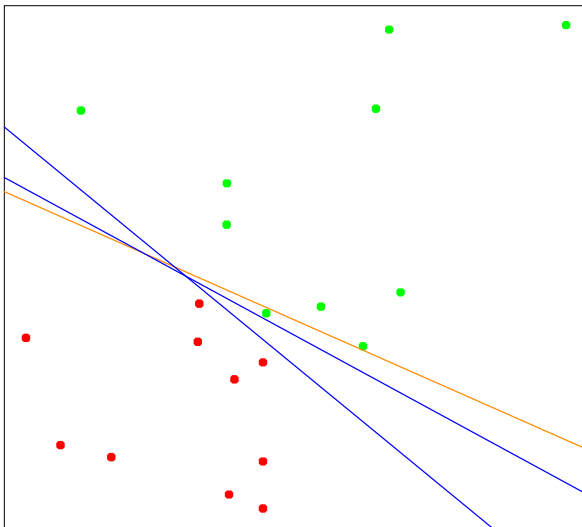
CSI 5v93: Introduction to machine learning, Lecture 20 – p. 2/15

Announcements

- project 4 feedback
- paper presentation: choice, meetings

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.).

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.

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

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

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

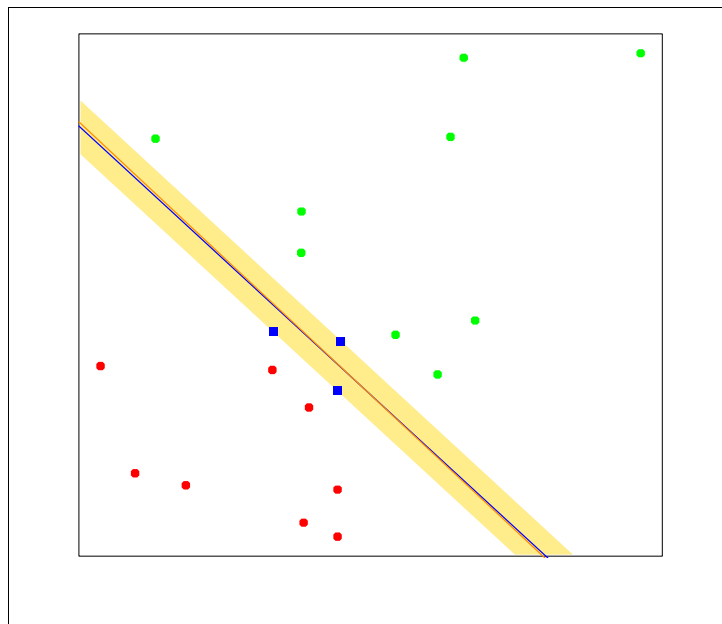
Mathematically defining maximum margin

From your textbook (4.5.2), presented on the board.

- incorporate $\|\beta\| = 1$
- rewrite in terms of $\|\beta\|$
- Lagrange constraint form
- derivatives of Lagrange form
- Karush-Kuhn-Tucker conditions
- support points

CSI 5v93: Introduction to machine learning, Lecture 20 – p. 9/15

Example of support points



CSI 5v93: Introduction to machine learning, Lecture 20 – p. 10/15

Non-separable data (chapter 12)

For separable data, we had (for all $i = 1 \dots n$):

$$y_i(x_i^T \beta + \beta_0) \geq C$$

For non-separable data, we introduce 'slack variables' ξ_i :

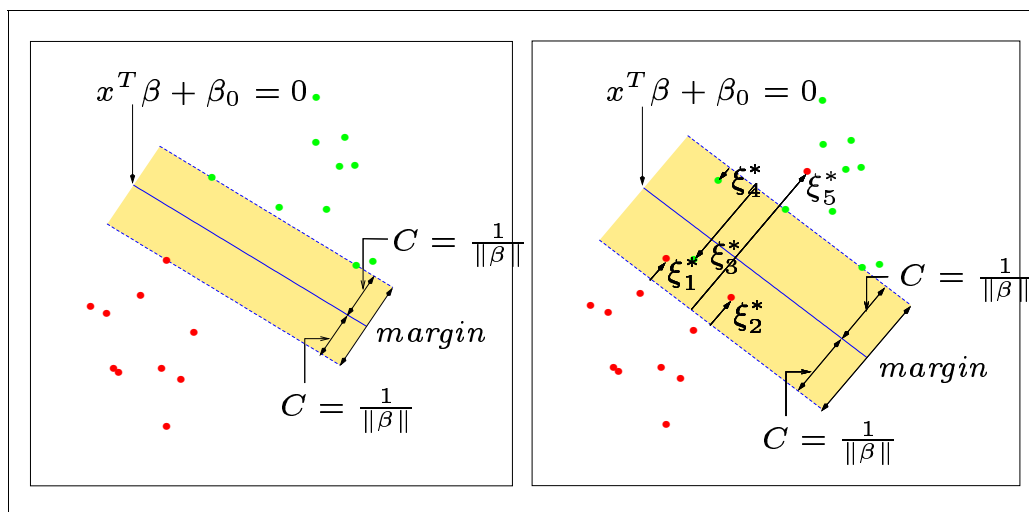
$$y_i(x_i^T \beta + \beta_0) \geq C(1 - \xi_i)$$

where

$$\xi_i \geq 0, \quad \sum_{i=1}^n \xi_i \leq \text{constant}$$

Separable and Non-separable data

$$y_i(x_i^T \beta + \beta_0) \geq C(1 - \xi_i), \quad \xi_i \geq 0, \quad \sum_{i=1}^n \xi_i \leq \text{constant}$$



Support vector classifier

$$y_i(x_i^T \beta + \beta_0) \geq C(1 - \xi_i), \quad \xi_i \geq 0, \quad \sum_{i=1}^n \xi_i \leq \text{constant}$$

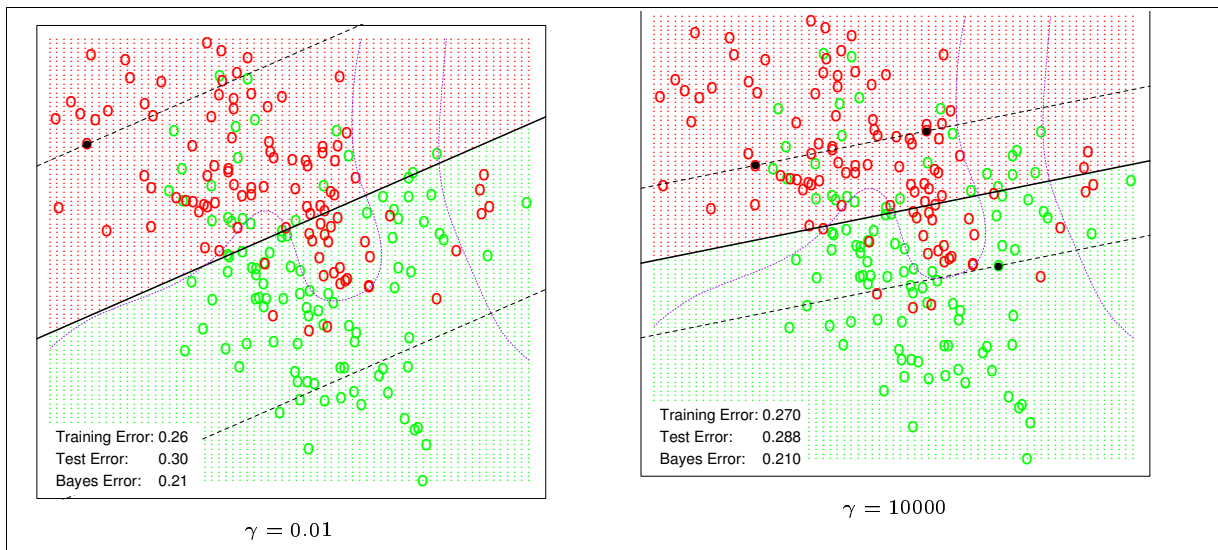
After many Lagrangian formulations and re-formulations (see 12.2.1), we get

$$\hat{\beta} = \sum_{i=1}^n \hat{\alpha}_i y_i x_i$$

The x_i that have corresponding $\alpha_i > 0$ are called support vectors, and define $\hat{\beta}$.

Remaining x_i do not have any influence on $\hat{\beta}$.

Support vector classifier example



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