

# Lecture 21: 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 21 – p. 1/14

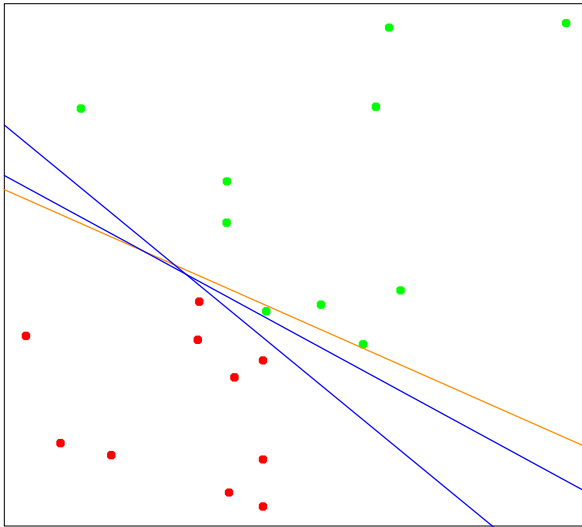
## Questions?

CSI 5v93: Introduction to machine learning, Lecture 21 – p. 2/14

## 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.

CSI 5v93: Introduction to machine learning, Lecture 21 – p. 3/14

## 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?

CSI 5v93: Introduction to machine learning, Lecture 21 – p. 4/14

## Mathematically defining maximum margin

---

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

subject to

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

CSI 5v93: Introduction to machine learning, Lecture 21 – p. 5/14

## Mathematically defining maximum margin

---

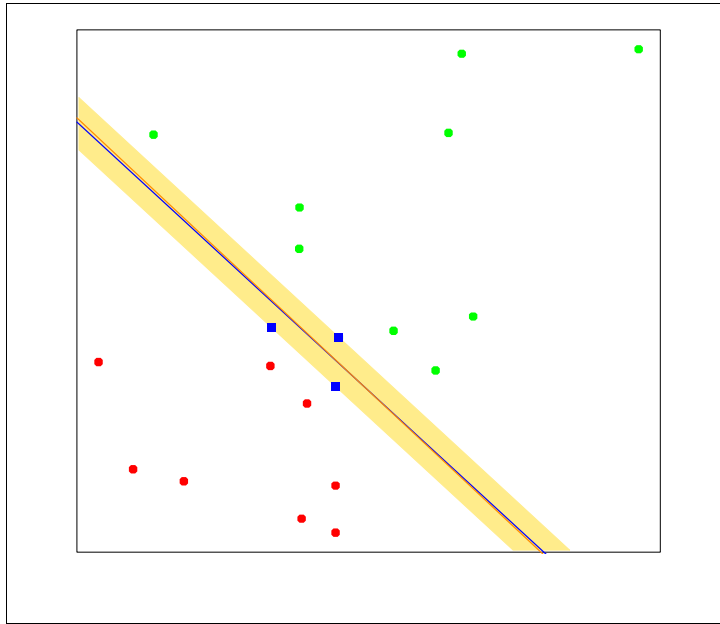
From your textbook (4.5.2), presented on the board.

- 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 21 – p. 6/14

## Example of support points

---



CSI 5v93: Introduction to machine learning, Lecture 21 – p. 7/14

## 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'  $\xi_i$ :

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

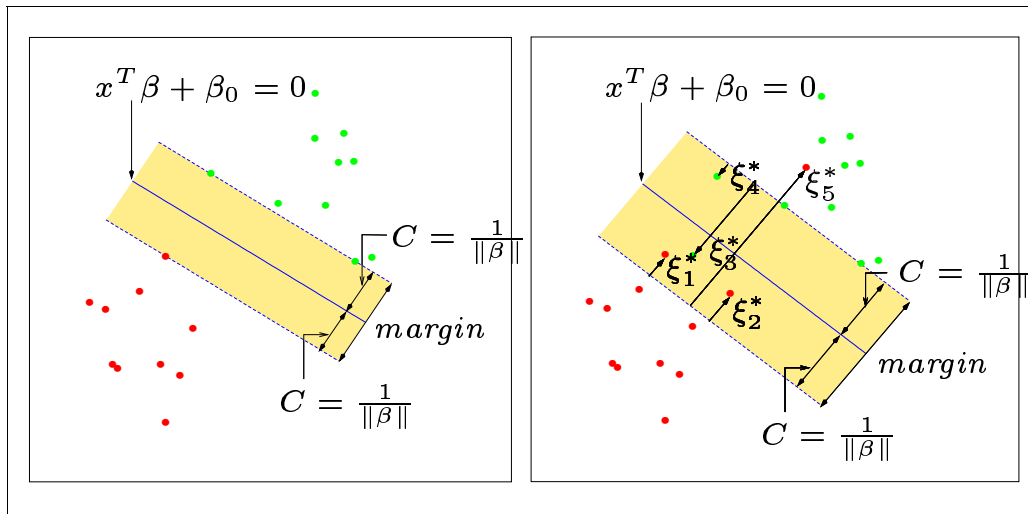
where

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

CSI 5v93: Introduction to machine learning, Lecture 21 – p. 8/14

## Separable and Non-separable data

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



CSI 5v93: Introduction to machine learning, Lecture 21 – p. 9/14

## 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  $\hat{\beta}$ .

CSI 5v93: Introduction to machine learning, Lecture 21 – p. 10/14

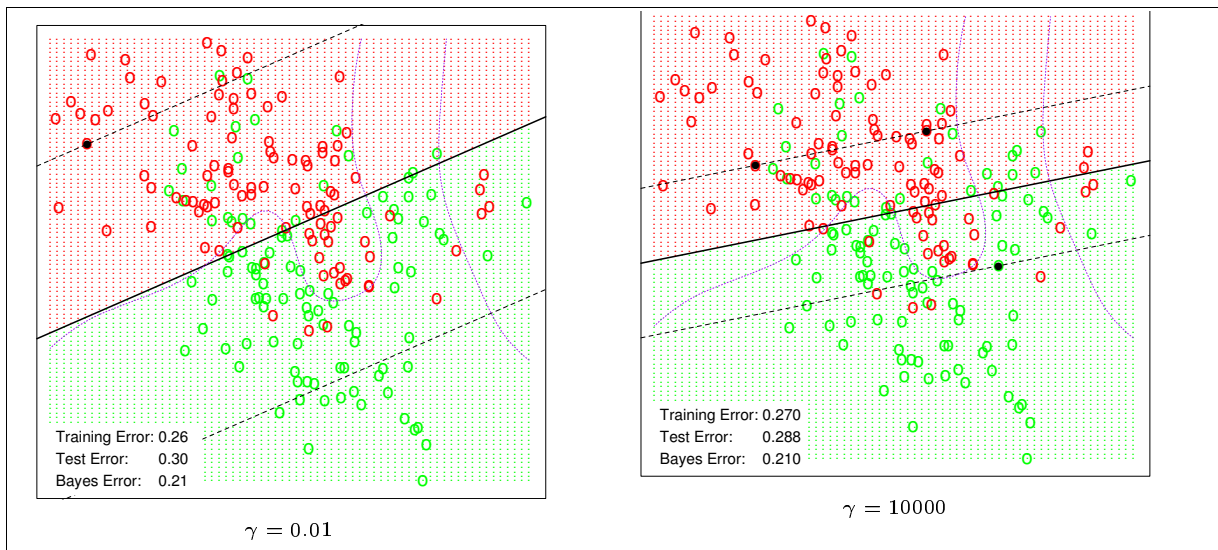
## Maximum margin for non-separable data

From your textbook (12.2.1), presented on the board.

- incorporate  $\xi_i$  constraints
- Lagrange constraint form
- derivatives of Lagrange form
- dual form
- Karush-Kuhn-Tucker conditions
- parameter  $\gamma$

CSI 5v93: Introduction to machine learning, Lecture 21 – p. 11/14

## Support vector classifier example



CSI 5v93: Introduction to machine learning, Lecture 21 – p. 12/14

## Next: flexible (non-linear) support vector machines

---

Question: How do we better handle data that is not linearly separable with support vector machines?

Answer: Transform the data into a higher dimension where the data is more easily separated!

## 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