Ch 9.3-4: Support Vector Machine

Lecture 29 - CMSE 381

Prof. Elizabeth Munch

Michigan State University
::
Dept of Computational Mathematics, Science & Engineering

Fri, Nov 17, 2023

Announcements

Announcements:

- HW #7 due Monday

Lec #	Date			Reading	Homeworks
20	Fri	Oct 27	Dimension Reduction	6.3	
21	Mon	Oct 30	More dimension reduction; High dimensions	6.4	
22	Wed	Nov 1	Polynomial & Step Functions	7.1,7.2	
23	Fri	Nov 3	Step Functions; Basis functions; Start Splines	7.2 - 7.4	
24	Mon	Nov 6	Regression Splines	7.4	~~HW #6 Due~~
25	Wed	Nov 8	Decision Trees	8.1	HW #6 Due
26	Fri	Nov 10	Random Forests	8.2.1, 8.2.2	
27	Mon	Nov 13	Maximal Margin Classifier	9.1	
28	Wed	Nov 15	SVC	9.2	
29	Fri	Nov 17	SVM	9.3, 9.4	
30	Mon	Nov 20	Single layer NN	10.1	HW #7 Due
31	Wed	Nov 22	Virtual: Project office hours		
	Fri	Nov 24	No class - Thanksgiving		
	Mon	Nov 27	**Review**		
	Wed	Nov 29	**Midterm #3**		
32	Fri	Dec 1	Multi Layer NN	10.2	
33	Mon	Dec 4	CNN	10.3	
34	Wed	Dec 6	Unsupervised Learning & Clustering	12.1, 12.4	
35	Fri	Dec 8	Virtual: Project office hours		Project due

Last time:

- 9.2 Support Vector Classifier

This lecture:

- 9.3 Support Vector Machine

Section 1

Last Time

Classification Setup

Data matrix:

$$X = \begin{pmatrix} - & x_1^T & - \\ - & x_2^T & - \\ & \vdots & \\ - & x_n^T & - \end{pmatrix}_{n \times p}$$

Observations in one of two classes,
$y_i \in \{-1, 1\}$

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$x_1 = \begin{pmatrix} x_{11} \\ \vdots \\ x_{1p} \end{pmatrix}, \cdots, x_n = \begin{pmatrix} x_{n1} \\ \vdots \\ x_{np} \end{pmatrix}$$

Separate out a test observation

$$x^* = (x_1^* \ \cdots \ x_p^*)^T$$

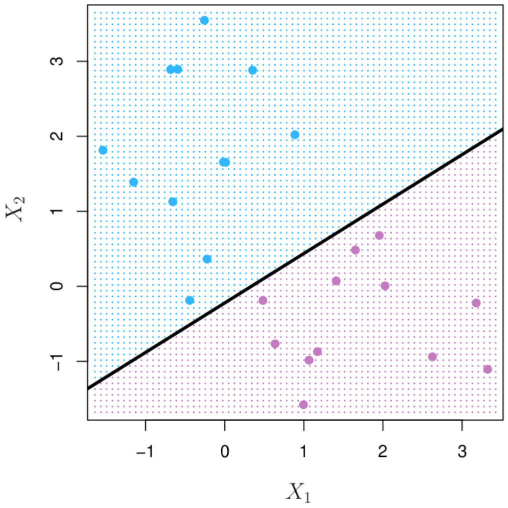
Hyperplane becomes a classifier

If you have a separating hyperplane:

- Check

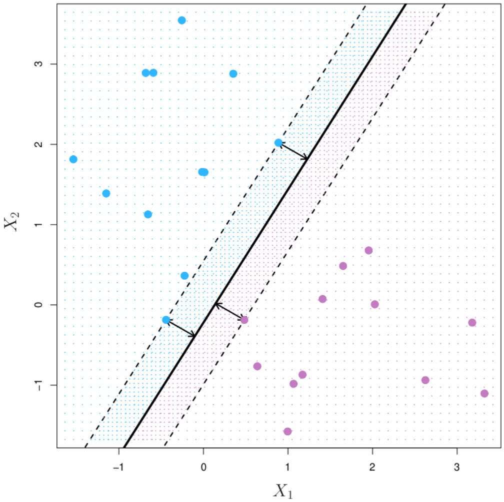
$$f(x^*) = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \cdots + \beta_p x_p^*$$

- If positive, assign $\hat{y} = 1$
- If negative, assign $\hat{y} = -1$



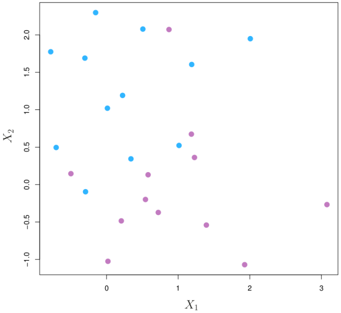
How do we pick? Old version

Maximal margin classifier

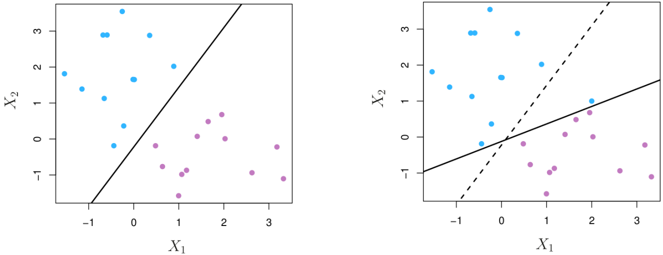


- For a hyperplane, the *margin* is the smallest distance from any data point to the hyperplane.
- Observations that are closest are called *support vectors*.
- The *maximal margin hyperplane* is the hyperplane with the largest margin
- The classifier built from this hyperplane is the *maximal margin classifier*.

Issues

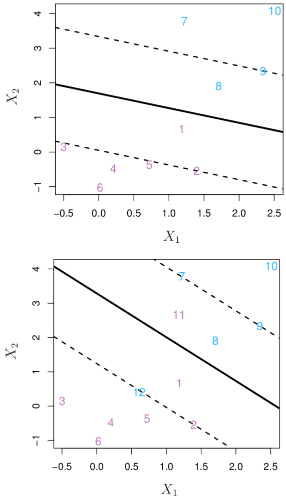


No separating hyperplane exists



Choice of hyperplane is sensitive to new points

Support Vector Classifier



$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n,\,M}{\text{maximize}} \quad M$$

$$\text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

$$\epsilon_i \geq 0, \quad \sum_{i=1}^{n} \epsilon_i \leq C,$$

Two formulations side by side

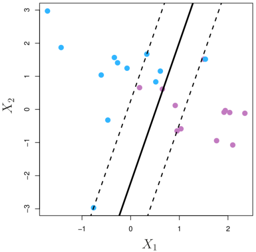
Support Vector Classifier

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n,\,M}{\text{maximize}} \quad M$$

$$\text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

$$\epsilon_i \geq 0, \quad \sum_{i=1}^{n} \epsilon_i \leq C,$$

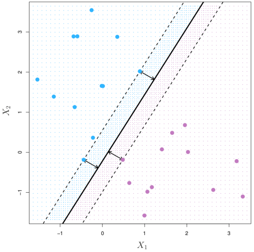


Maximal Margin Classifier

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,M}{\text{maximize}} \; M$$

$$\text{subject to} \; \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \geq M \quad \forall \; i = 1, \ldots, n$$



So many variables

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n,\,M}{\text{maximize}} \quad M$$

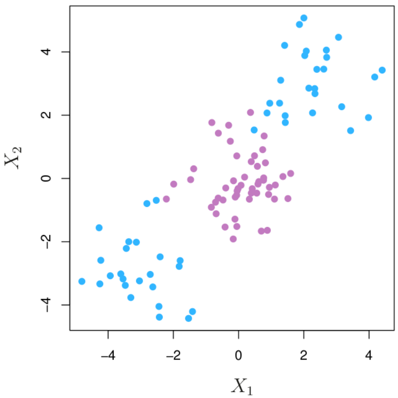
$$\text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 = 1,$$

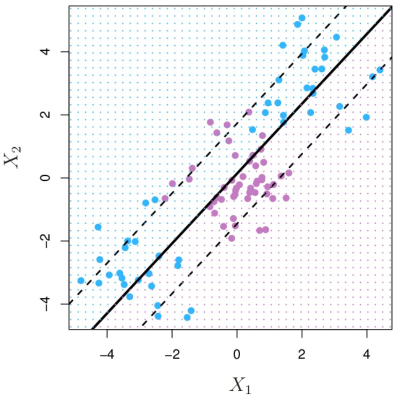
$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

$$\epsilon_i \geq 0, \quad \sum_{i=1}^{n} \epsilon_i \leq C,$$

- C is nonnegative tuning parameter
- M is the width of the margin
- $\varepsilon_1, \cdots, \varepsilon_n$ are slack variables allowing observations to go to the other side

Limiting factor of SVC





Section 2

Support Vector Machine

Example of using more features

Want $2p$ features:

$$X_1, X_1^2, X_2, X_2^2, \cdots, X_p, X_p^2$$

Optimization becomes:

$$\underset{\beta_0,\beta_{11},\beta_{12},\ldots,\beta_{p1},\beta_{p2},\epsilon_1,\ldots,\epsilon_n,\, M}{\text{maximize}} \quad M$$

$$\text{subject to } y_i \left(\beta_0 + \sum_{j=1}^{p} \beta_{j1} x_{ij} + \sum_{j=1}^{p} \beta_{j2} x_{ij}^2 \right) \geq M(1 - \epsilon_i),$$

$$\sum_{i=1}^{n} \epsilon_i \leq C, \quad \epsilon_i \geq 0, \quad \sum_{j=1}^{p} \sum_{k=1}^{2} \beta_{jk}^2 = 1.$$

Kernels

Inner products

$$\langle a, b \rangle = \sum_{i=1}^{r} a_i b_i$$

Quick computations

What are the following?

- $\langle (1,1),(0,3) \rangle$
- $\langle (1,1),(3,2) \rangle$
- $\langle (2,3),(0,3) \rangle$
- $\langle (2,3),(3,2) \rangle$

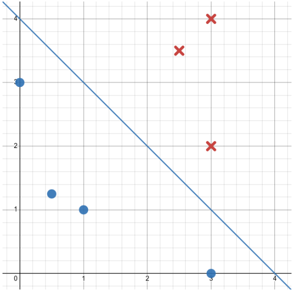
SVC via inner products

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle$$

Example

$$-2\sqrt{2} + \frac{\sqrt{2}}{2}X_1 + \frac{\sqrt{2}}{2}X_2 = 0$$

$$-2\sqrt{2} + \frac{\sqrt{2}}{18}\langle (X_1, X_2), (0,3)\rangle + \frac{\sqrt{2}}{6}\langle (X_1, X_2), (3,2)\rangle = 0$$



What are the α_is?

Data point	α_i
$(3,4)$	
$(2.5,3.5)$	
$(3,2)$	
$(3,0)$	
$(0,3)$	
$(1,1)$	
$(0.5,1.25)$	

What α_i's are needed to write the hyperplane

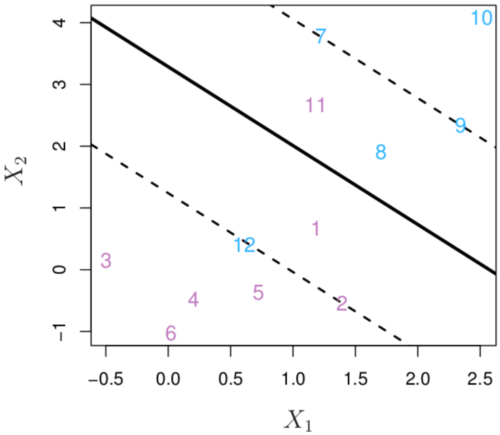
$$-2\sqrt{2}+\frac{\sqrt{2}}{18}\langle(X_1,X_2),(0,3)\rangle+\frac{\sqrt{2}}{6}\langle(X_1,X_2),(3,2)\rangle$$

of the previous page in the form

$$f(x)=\beta_0+\sum_{i\in\mathcal{S}}\alpha_i\langle x,x_i\rangle?$$

SVC via inner products of support vectors

$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}} \alpha_i \langle x, x_i \rangle$$



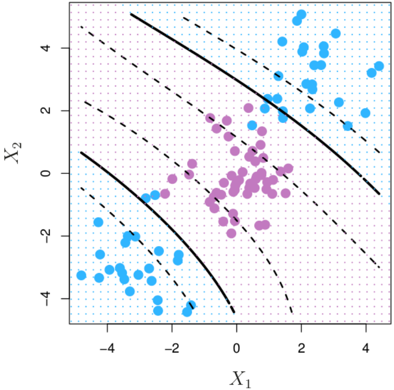
The kernel

$$K(x_i, x_i')$$

$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}} \alpha_i K(x, x_i)$$

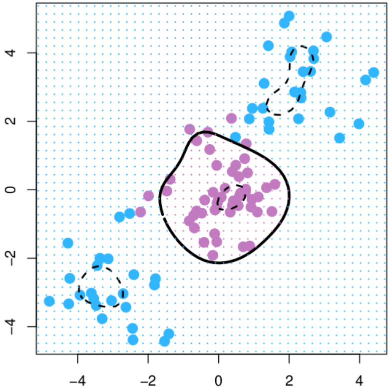
A polynomial kernel

$$K(x_i, x_{i'}) = \left(1 + \sum_{j=1}^{p} x_{ij} x_{i'j}\right)^d$$



A radial kernel

$$K(x_i, x_i') = \exp\left(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2\right)$$



Support Vector Machine

$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}} \alpha_i K(x, x_i)$$

Coding

Section 3

SVM with more than two classes

One-Vs-One Classification

Also called all-pairs

One-Vs-All Classification

TL;DR

$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}} \alpha_i K(x, x_i)$$

Kernels

- Linear

$$K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij} x_{i'j}$$

- Polynomial

$$K(x_i, x_{i'}) = \left(1 + \sum_{j=1}^{p} x_{ij} x_{i'j}\right)^d$$

- Radial

$$K(x_i, x_i') = \exp\left(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2\right)$$

Next time

Lec #	Date			Reading	Homeworks
20	Fri	Oct 27	Dimension Reduction	6.3	
21	Mon	Oct 30	More dimension reduction; High dimensions	6.4	
22	Wed	Nov 1	Polynomial & Step Functions	7.1,7.2	
23	Fri	Nov 3	Step Functions; Basis functions; Start Splines	7.2 - 7.4	
24	Mon	Nov 6	Regression Splines	7.4	~~HW #6 Due~~
25	Wed	Nov 8	Decision Trees	8.1	HW #6 Due
26	Fri	Nov 10	Random Forests	8.2.1, 8.2.2	
27	Mon	Nov 13	Maximal Margin Classifier	9.1	
28	Wed	Nov 15	SVC	9.2	
29	Fri	Nov 17	SVM	9.3, 9.4	
30	Mon	Nov 20	Single layer NN	10.1	HW #7 Due
31	Wed	Nov 22	Virtual: Project office hours		
	Fri	Nov 24	No class - Thanksgiving		
	Mon	Nov 27	**Review**		
	Wed	Nov 29	**Midterm #3**		
32	Fri	Dec 1	Multi Layer NN	10.2	
33	Mon	Dec 4	CNN	10.3	
34	Wed	Dec 6	Unsupervised Learning & Clustering	12.1, 12.4	
35	Fri	Dec 8	Virtual: Project office hours		Project due