Ch 9.3-4: Support Vector Machine

Lecture 28 - CMSE 381

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

Fri, Apr 3, 2026

Announcements

Last time:

- 9.2 Support Vector Classifier

This lecture:

- 9.3 Support Vector Machine

Announcements:

- HW #6 due Sunday 4/12

21	W	3/18	Polynomial & Step Functions	7.1-7.2		
22	F	3/20	Step Functions; Basis functions; Start Splines	7.2-7.4		
23	M	3/23	Regression Splines	7.4		
24	W	3/25	Decision Trees	8.1		Q7
25	F	3/27	Random Forests	8.2.1, 8.2.2	HW #5 Due Sun 3/29	
26	M	3/30	Maximal Margin Classifier	9.1		
27	W	4/1	SVC	9.2		Q8
28	F	4/3	SVM	9.3, 9.4		
29	M	4/6	Single Layer NN	10.1		
30	W	4/8	Multi Layer NN	10.2		Q9
31	F	4/10	CNN	10.3		
32	M	4/13	Unsupervised learning / clustering	12.1, 12.4	HW #6 Due Sun 4/12	Q10
33	W	4/15	Virtual: Project Office Hours			
	F	4/17	***Review***			
	M	4/20	**Midterm #3**			
	W	4/22				
	F	4/24			**Project Due**	

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

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

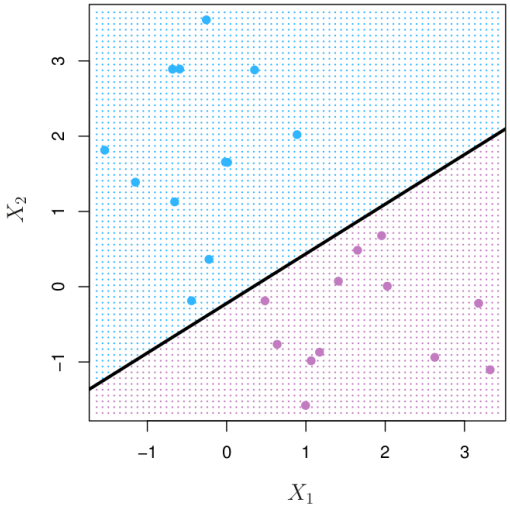
Hyperplane becomes a classifier

If you have a separating hyperplane:

- Check

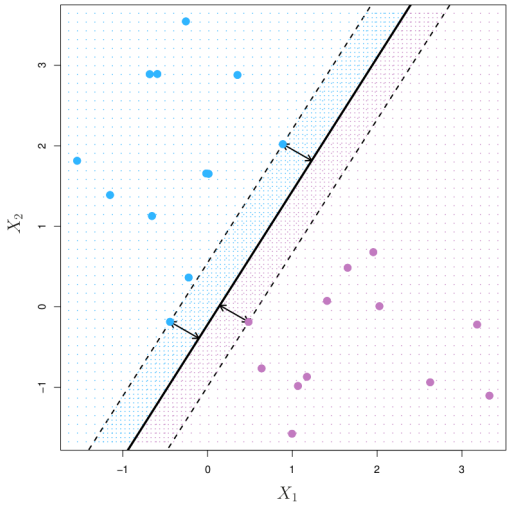
$$f(x^*) = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \cdots + \beta_p x_p^*$$

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



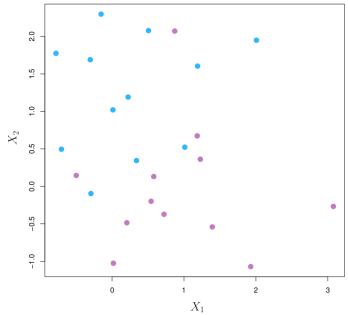
How do we pick? Old version

Maximal margin classifier

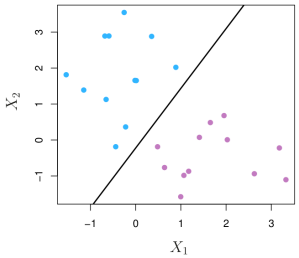


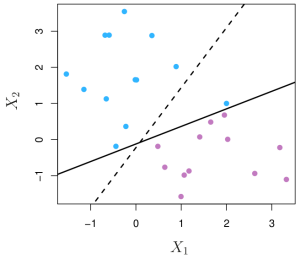
- For a hyperplane, the *margin* is the smallest distance from any data point to the hyperplane.
- Observations that are closest are called *support vectors*.
- The *maximal margin hyperplane* is the hyperplane with the largest margin
- The classifier built from this hyperplane is the *maximal margin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

Issues



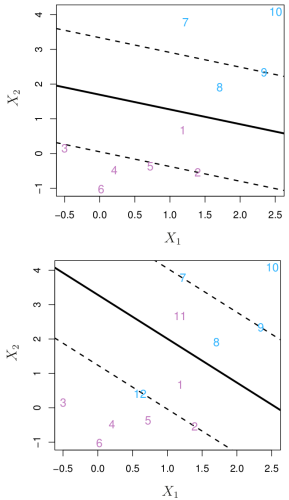
No separating hyperplane exists





Choice of hyperplane is sensitive to new points

Support Vector Classifier



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

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

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

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

Test your understanding in PollEv!

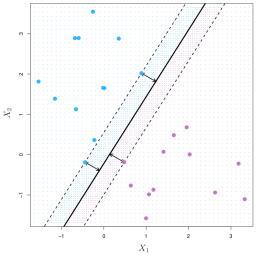
Two formulations side by side

Maximal Margin Classifier

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

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

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



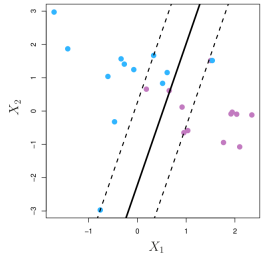
Support Vector Classifier

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

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

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

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



So many variables

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

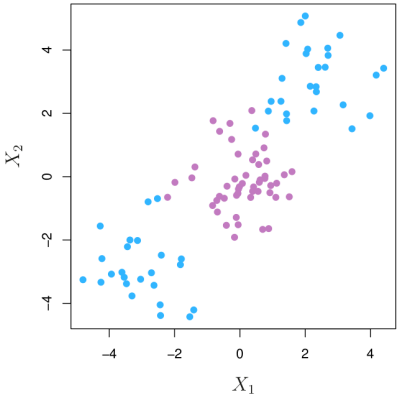
$$\text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 = 1,$$

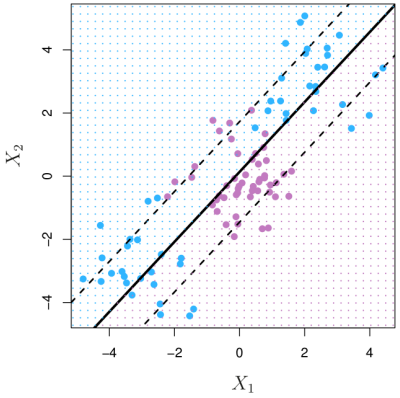
$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

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

- 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





What will you learn today?

- How is nonlinearity introduced in Support Vector Machine?
  - You should be able to describe how kernels work conceptually and mathematically.
- What type of kernel makes Support Vector Machine equivalent to Support Vector Classifier?
- What other kernels are there?
- What type of kernel leads to more local behavior?
- How to use Support Vector Machine in Python?
  - You should be able to fit a Support Vector Machine using different kernels covered in class.
  - You should be able select the appropriate hyperparameter that optimize bias-variance trade off.

Section 2

Support Vector Machine

How to bend the plane: polynomial features?

Want $2p$ features:

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

Optimization becomes:

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

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

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

Kernels - "There is no plane"

$$y_i f(x_i) \geq M(1 - \epsilon_i)$$

$f(x_i)$ should give you some more general way of talking about how far is point x_i away from the decision boundary.

The kernel - replacing global coordinate with local similarity

$y_i f(x_i) \geq M(1 - \epsilon_i)$, where

$K(x_i, x_i')$

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

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

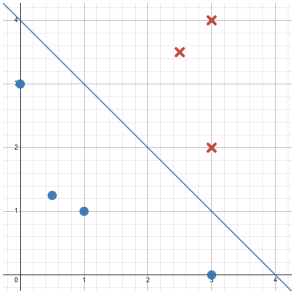
$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle$$

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

Example

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

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



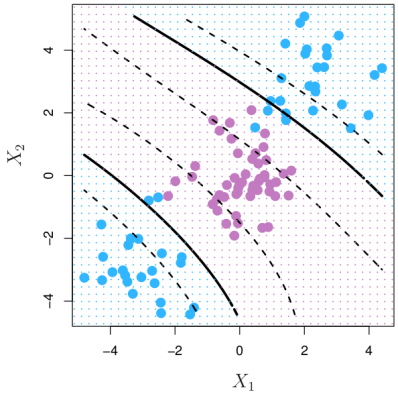
The kernel - more generally

$K(x_i, x_i')$

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

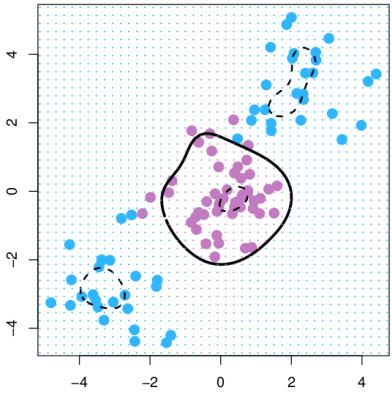
A polynomial kernel

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



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

21	W	3/18	Polynomial & Step Functions	7.1-7.2		
22	F	3/20	Step Functions; Basis functions; Start Splines	7.2-7.4		
23	M	3/23	Regression Splines	7.4		
24	W	3/25	Decision Trees	8.1		Q7
25	F	3/27	Random Forests	8.2.1, 8.2.2	HW #5 Due Sun 3/29	
26	M	3/30	Maximal Margin Classifier	9.1		
27	W	4/1	SVC	9.2		Q8
28	F	4/3	SVM	9.3, 9.4		
29	M	4/6	Single Layer NN	10.1		
30	W	4/8	Multi Layer NN	10.2		Q9
31	F	4/10	CNN	10.3		
32	M	4/13	Unsupervised learning / clustering	12.1, 12.4	HW #6 Due Sun 4/12	
33	W	4/15	Virtual: Project Office Hours			Q10
	F	4/17	***Review***			
	M	4/20	**Midterm #3**			
	W	4/22				
	F	4/24			**Project Due**	