Ch 9.3-4: Support Vector Machine

Lecture 20 - CMSE 381

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

Fri, April 8th, 2024

Announcements

Last time:

- 9.2 Support Vector Classifier

This lecture:

- 9.3 Support Vector Machine

Announcements:

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

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

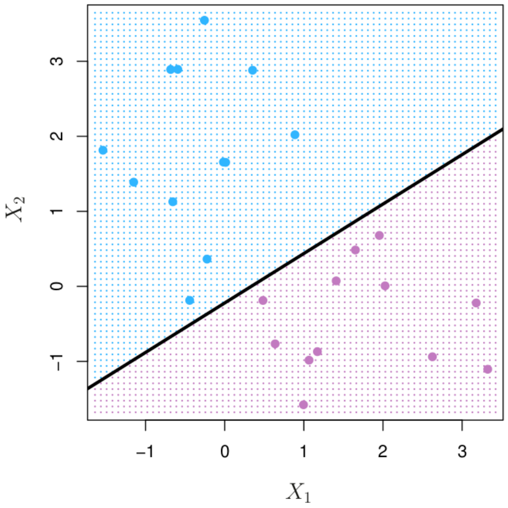
Hyperplane becomes a classifier

If you have a separating hyperplane:

- Check

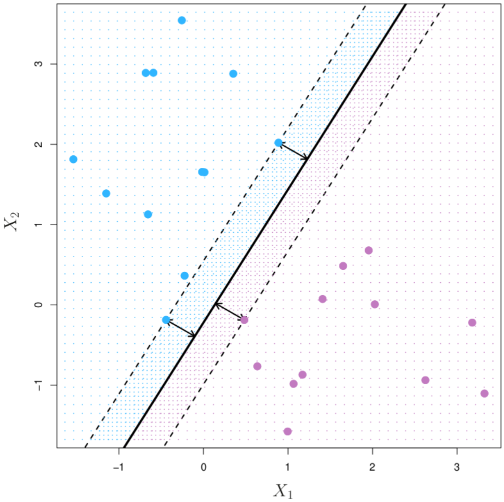
$$f(x^*) = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \cdots + \beta_p x_p^*$$

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



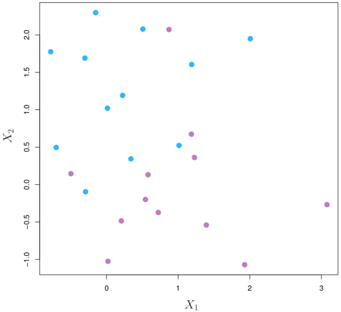
How do we pick? Old version

Maximal margin classifier

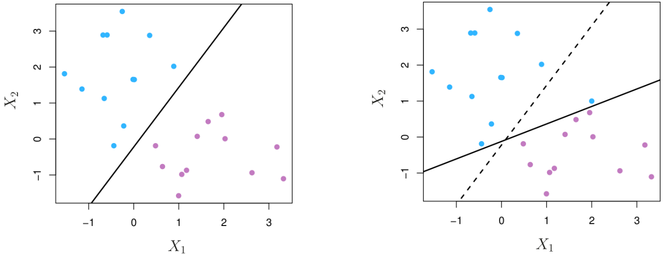


- For a hyperplane, the *margin* is the smallest distance from any data point to the hyperplane.
- Observations that are closest are called *support vectors*.
- The *maximal margin hyperplane* is the hyperplane with the largest margin
- The classifier built from this hyperplane is the *maximal margin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

Issues

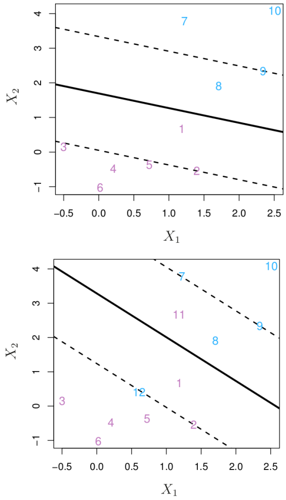


No separating hyperplane exists



Choice of hyperplane is sensitive to new points

Support Vector Classifier



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

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

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

slack variable

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

- Soft margin
- Allow for violations across margin
- Allow for violations across hyperplane

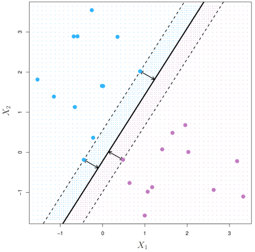
Two formulations side by side

Maximal Margin Classifier

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

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

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



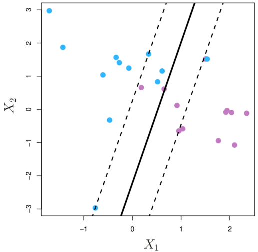
Support Vector Classifier

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

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

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

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



So many variables

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

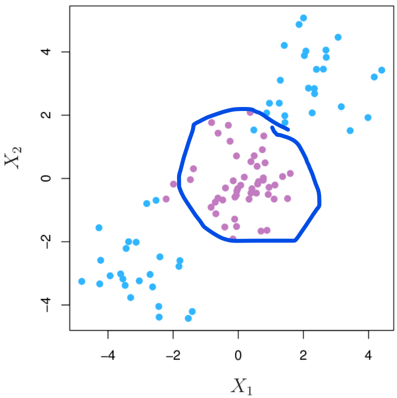
$$\text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 = 1,$$

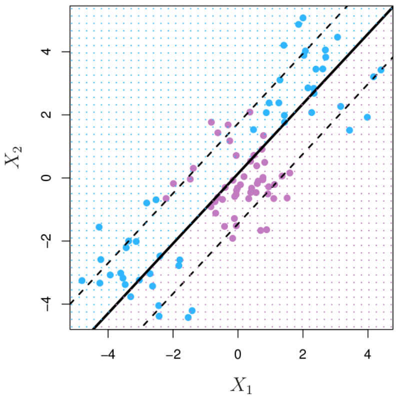
$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

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

- C is nonnegative tuning parameter
  - Bounds sum of ε_i; number & severity of violating margin (budget)
  - $C = 0$ means no violations allowed
  - $C > 0$ means at most C observations can be on wrong side of hyperplane
- M is the width of the margin
- $\varepsilon_1, \cdots, \varepsilon_n$ are slack variables allowing observations to go to the other side
  - If $\varepsilon_i = 0$, then on correct side of margin
  - If $\varepsilon_i > 0$ then on the wrong side of margin (Violated margin)
  - If $\varepsilon_i > 1$ then on the wrong side of hyperplane

Limiting factor of SVC





- Requires linear boundaries
- "Non-linear" is a lot of things, how do you choose features to learn?
- Right side is result from 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:

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

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

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

This becomes unwieldly if we have to check, say every degree 2 monomoial X_iX_j, so need something more efficient

Kernels

- Main idea: enlarge feature space like above
- But want it to be computationally efficient

Inner products

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

- Example
  $\langle (1, 2, 3), (5, 0, 2) \rangle = 5 + 0 + 6 = 11$

Quick computations

What are the following?

- $\langle (1,1),(0,3) \rangle = 0 + 3 = 3$
- $\langle (1,1),(3,2) \rangle = 3 + 2 = 5$
- $\langle (2,3),(0,3) \rangle = 0 + 9 = 9$
- $\langle (2,3),(3,2) \rangle = 6 + 6 = 12$

SVC via inner products

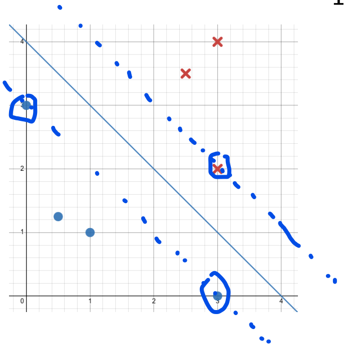
$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle$$

- Via some magic, there are coefficients α_i which give the linear support vector classifier
- In this notation, the x_i's are all the training points
- How to actually get it is outside the scope of this class
- To estimate the paramters $\alpha_1, \cdots, \alpha_n$ and β_0, need $\binom{n}{2} = n(n-1)/2$ inner products $\langle x_i, x_{i'} \rangle$
- Turns out α_i are only nonzero for support vectors

Example

Given the SVC: $-2\sqrt{2}+\frac{\sqrt{2}}{2}X_1+\frac{\sqrt{2}}{2}X_2=0$ Find its inner product expression.

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



- $f(1,1)=-2\sqrt{2}+\frac{\sqrt{2}}{18}\langle (1,1),(0,3)\rangle+\frac{\sqrt{2}}{6}\langle (1,1),(3,2)\rangle$
- $=-2\sqrt{2}+\frac{\sqrt{2}}{18}\cdot 3+\frac{\sqrt{2}}{6}\cdot 5$
- $=(-2+\frac{3}{18}+\frac{5}{6})\sqrt{2}=-\sqrt{2}$

What are the α_is?

Data point	α_i
$(3, 4)$	0
$(2.5, 3.5)$	0
$(3, 2)$	$\sqrt{2}/6$
$(3, 0)$	0
$(0, 3)$	$\sqrt{2}/18$
$(1, 1)$	0
$(0.5, 1.25)$	0

What α_i's are needed to write the hyperplane

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

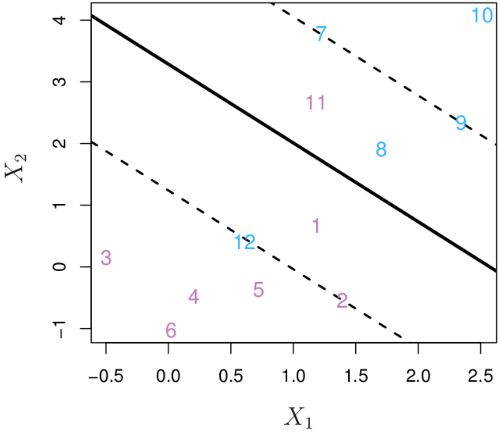
of the previous page in the form

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

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

similarity of
x and x_i

- Turns out α_i are only nonzero for support vectors
- To estimate the paramters $\alpha_1, \cdots, \alpha_n$ and β_0, need $\binom{n}{2} = n(n-1)/2$ inner products $\langle x_i, x_{i'} \rangle$



The point: representing linear classifier $f(x)$ just needs inner products

The kernel

$$K(x_i, x_i')$$

- Swap out my inner product $\langle x, x_i \rangle$ for something potentially more complicated
- $\langle x, x_i \rangle$ is known as a linear kernel

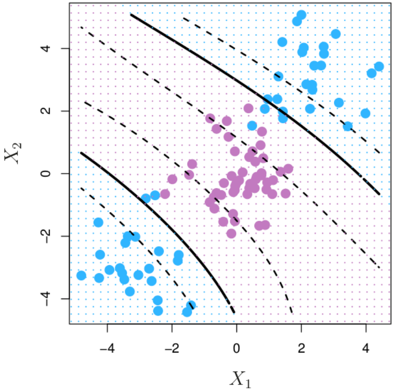
$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}} \alpha_i K(x, x_i)$$

- The function defined is as above with whatever choice of K
- $K(x_i, x_i')$ can be thought of as similarity function.
- When the support vector classifier is combined with a non-linear kernel the resulting classifier is known as a support vector machine.

A polynomial kernel

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

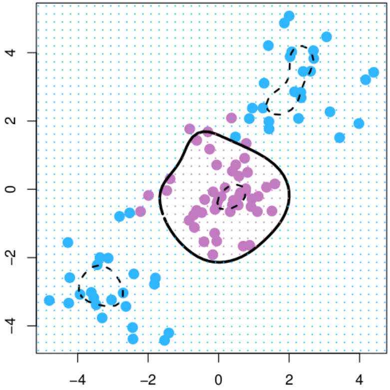
- much more flexible decision boundary.
- amounts to fitting a support vector classifier in a higher-dimensional space involving polynomials of degree *d*, rather than in the original feature space.
- Right: degree 3 polynomial kernel



A radial kernel

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

- Draw picture of circular values around a point
- big distance makes $\sum(x_{ij} - x_{i'j})^2$ big, but then e^{-big} is tiny.
- In $f(x)$, this means that x_i plays little role if it's far away.
- Radial kernel has local behavior



Support Vector Machine

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

- Learning function above
- Choose K in advance

Section 3

SVM with more than two classes

One-Vs-One Classification

Also called all-pairs

- Predict for $K > 2$ classes
- Construct $\binom{K}{2}$ SVMs, each of which compairs a pair of classes
- Write example with $\{$ apples, bananas, oranges, strawberries$\}$
- Assign the observation to the class which was most frequently assigned

One-Vs-All Classification

- Fit K SVMs each time comparing one class to remaining $K-1$ classes.
- Again, do $\{$ apples, bananas, oranges, strawberries$\}$
- Figure out coefficients for that class, and determine $f_k(x) = \beta_{0k} + \beta_{1k}x_1^* + \cdots + \beta_{pk}x_p^*$
- Assign observation to the class for which $f_k(x)$ is largest

Other dissimilarity measure

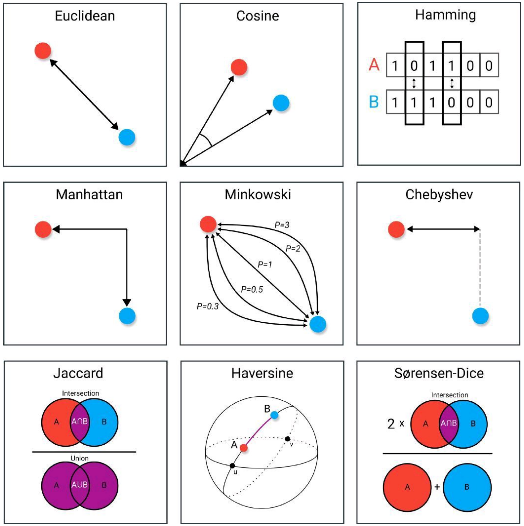


Photo Credit Link

Summary

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

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

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