

CPSC 8420 Advanced Machine Learning

Week 9: Support Vector Machines

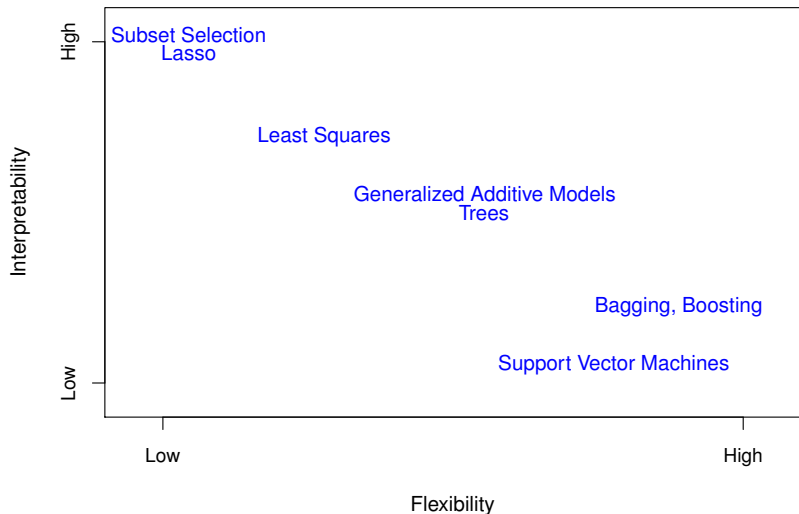
Dr. Kai Liu

October 13, 2020

Support Vector Machines



Comparison of Different Machine Learning Models



Overview

- *Support Vector Machine* (SVM) is an approach for **supervised classification** that was developed in the computer science community in the 1990s.
- SVM is intended for **binary classification**.
- SVM can be considered an extension of the *support vector classifier*, which is an extension of the *maximum margin classifier*.

Vladimir N. Vapnik



Vladimir N. Vapnik

His main contributions are:

- ① Support Vector Machines
- ② Statistical Learning Theory
- ③ Vapnik–Chervonenkis theory
- ④ Kernel Method

he won numerous awards including:

- ① IEEE John von Neumann Medal (2017)
- ② Benjamin Franklin Medal (2012)
- ③ IEEE Neural Networks Pioneer Award (2010)
- ④ Fellow of the U.S. National Academy of Engineering (2006)

“Nothing is more practical than a good theory.”

Hyperplanes

- The maximum margin classifier (and also its extension, the support vector classifier) is based on the idea of a **separating hyperplane**

Hyperplanes

- The maximum margin classifier (and also its extension, the support vector classifier) is based on the idea of a **separating hyperplane**
- In two dimensions, a hyperplane is a flat one-dimensional subspace (i.e., a line), defined by:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$$

Hyperplanes

- The maximum margin classifier (and also its extension, the support vector classifier) is based on the idea of a **separating hyperplane**
- In two dimensions, a hyperplane is a flat one-dimensional subspace (i.e., a line), defined by:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$$

- In three dimensions, a hyperplane is a flat two-dimensional subspace (i.e., a plane), defined by:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 = 0$$

Hyperplanes

- The maximum margin classifier (and also its extension, the support vector classifier) is based on the idea of a **separating hyperplane**
- In two dimensions, a hyperplane is a flat one-dimensional subspace (i.e., a line), defined by:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$$

- In three dimensions, a hyperplane is a flat two-dimensional subspace (i.e., a plane), defined by:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 = 0$$

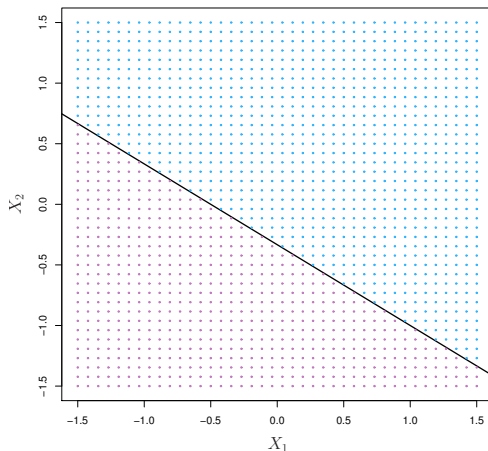
- In p dimensions, a hyperplane is difficult to visualize, but can easily be defined by extending the above equations to:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = 0$$

Hyperplanes

The hyperplane $1 + 2X_1 + 3X_2 = 0$.

- Blue: the set of points for which $1 + 2X_1 + 3X_2 > 0$
- Purple: the set of points for which $1 + 2X_1 + 3X_2 < 0$



Separating Hyperplane

- Suppose that it is possible to construct a hyperplane that separates the training observations perfectly according to their class labels.

Separating Hyperplane

- Suppose that it is possible to construct a hyperplane that separates the training observations perfectly according to their class labels.
- Label the observations above the hyperplane as $y_i = 1$ and those below the hyperplane as $y_i = -1$.

Separating Hyperplane

- Suppose that it is possible to construct a hyperplane that separates the training observations perfectly according to their class labels.
- Label the observations above the hyperplane as $y_i = 1$ and those below the hyperplane as $y_i = -1$.
- Then a separating hyperplane has the property that:

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} > 0 \quad \text{if } y_i = 1 \quad (1)$$

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} < 0 \quad \text{if } y_i = -1 \quad (2)$$

Separating Hyperplane

- Suppose that it is possible to construct a hyperplane that separates the training observations perfectly according to their class labels.
- Label the observations above the hyperplane as $y_i = 1$ and those below the hyperplane as $y_i = -1$.
- Then a separating hyperplane has the property that:

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} > 0 \quad \text{if } y_i = 1 \quad (1)$$

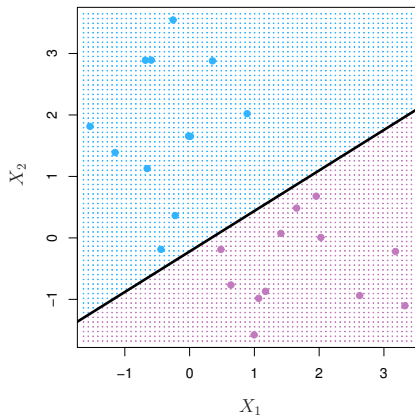
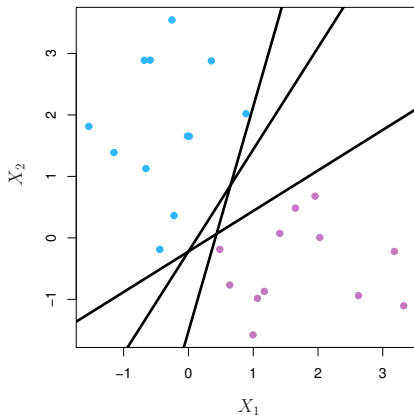
$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} < 0 \quad \text{if } y_i = -1 \quad (2)$$

- This can be written as:

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) > 0 \quad (3)$$

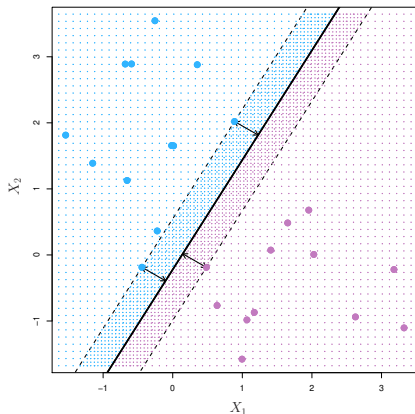
for all $i = 1, \dots, n$.

Maximum Margin Classifier



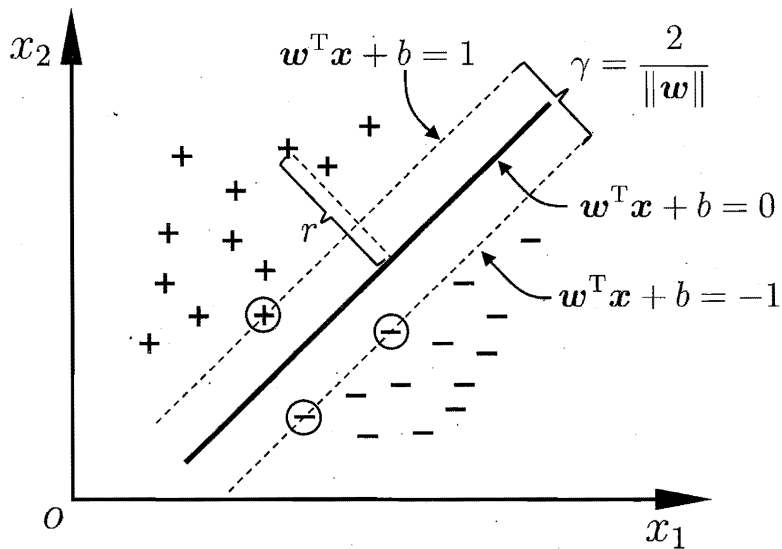
Example of data that can be perfectly separated using a hyperplane.
Problem: there exists an infinite number of such hyperplanes.

Maximum Margin Classifier



- The **maximum margin classifier** uses the separating hyperplane that is farthest from the training observations (that is, has the largest margin).
- The three observations that lie along the dashed lines indicating the width of the margin are called **support vectors**.

Maximum Margin Classifier



Objective

The maximum margin hyperplane is the solution to the following optimization problem:

$$\max \gamma = \frac{y(w^T x + b)}{\|w\|_2} \quad s.t \quad y_i(w^T x_i + b) = \gamma'^{(i)} \geq \gamma' \quad (i = 1, 2, \dots, m) \quad (4)$$

usually we set $\gamma' = 1$, then it is equivalent to:

$$\begin{aligned} \max \quad & \frac{1}{\|w\|_2} \quad s.t \quad y_i(w^T x_i + b) \geq 1 \quad (i = 1, 2, \dots, m) \\ \min \quad & \frac{1}{2} \|w\|_2^2 \quad s.t \quad y_i(w^T x_i + b) \geq 1 \quad (i = 1, 2, \dots, m) \end{aligned} \quad (5)$$

Optimization

$$L(w, b, \alpha) = \frac{1}{2} \|w\|_2^2 - \sum_{i=1}^m \alpha_i [y_i (w^T x_i + b) - 1] \quad \text{s.t.} \quad \alpha_i \geq 0$$
$$\underbrace{\min}_{w, b} \underbrace{\max}_{\alpha_i \geq 0} L(w, b, \alpha) \tag{6}$$
$$\underbrace{\max}_{\alpha_i \geq 0} \underbrace{\min}_{w, b} L(w, b, \alpha)$$

we can first optimize L w.r.t. w and b , namely $\underbrace{\min}_{w, b} L(w, b, \alpha)$ by taking the derivative.

$$\begin{aligned} L(w, b, \alpha) &= \frac{1}{2} \|w\|_2^2 - \sum_{i=1}^m \alpha_i [y_i (w^T x_i + b) - 1] \quad \text{s.t.} \quad \alpha_i \geq 0 \\ \frac{\partial L}{\partial w} = 0 &\Rightarrow w = \sum_{i=1}^m \alpha_i y_i x_i \\ \frac{\partial L}{\partial b} = 0 &\Rightarrow \sum_{i=1}^m \alpha_i y_i = 0 \end{aligned} \tag{7}$$

Optimization

Now define $\psi(\alpha) = \underbrace{\min}_{w, b} L(w, b, \alpha)$, we have:

$$\begin{aligned}\psi(\alpha) &= \frac{1}{2} \|w\|_2^2 - \sum_{i=1}^m \alpha_i [y_i (w^T x_i + b) - 1] \\&= \frac{1}{2} w^T \sum_{i=1}^m \alpha_i y_i x_i - w^T \sum_{i=1}^m \alpha_i y_i x_i - \sum_{i=1}^m \alpha_i y_i b + \sum_{i=1}^m \alpha_i \\&= -\frac{1}{2} w^T \sum_{i=1}^m \alpha_i y_i x_i - \sum_{i=1}^m \alpha_i y_i b + \sum_{i=1}^m \alpha_i \\&= -\frac{1}{2} \left(\sum_{i=1}^m \alpha_i y_i x_i \right)^T \left(\sum_{i=1}^m \alpha_i y_i x_i \right) - b \sum_{i=1}^m \alpha_i y_i + \sum_{i=1}^m \alpha_i \\&= -\frac{1}{2} \sum_{i=1}^m \alpha_i y_i x_i^T \sum_{i=1}^m \alpha_i y_i x_i + \sum_{i=1}^m \alpha_i \\&= \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1, j=1}^m \alpha_i \alpha_j y_i y_j x_i^T x_j\end{aligned}\tag{8}$$

$$\underbrace{\max}_{\alpha} -\frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum_{i=1}^m \alpha_i \quad (9)$$

which is equivalent to:

$$\begin{aligned} \underbrace{\min}_{\alpha} \quad & \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle - \sum_{i=1}^m \alpha_i \\ \text{s.t.} \quad & \sum_{i=1}^m \alpha_i y_i = 0 \\ & \alpha_i \geq 0 \quad i = 1, 2, \dots, m \end{aligned} \quad (10)$$

Optimization

We can make use of **Sequential Minimal Optimization (SMO)** to obtain α^* , which optimizes α pair-wise. By the definition

$w = \sum_{i=1}^m \alpha_i y_i x_i$, we have: $w^* = \sum_{i=1}^m \alpha_i^* y_i x_i$. To determine b^* , for

arbitrary (x_s, y_s) : $y_s(w^T x_s + b) = y_s(\sum_{i=1}^m \alpha_i y_i x_i^T x_s + b) = 1$, that is

$b_s^* = y_s - \sum_{i=1}^m \alpha_i y_i x_i^T x_s$. To determine the support vector, according to

KKT complementary condition: $\alpha_i^*(y_i(w^T x_i + b) - 1) = 0$, if $\alpha_i > 0$, then $y_i(w^T x_i + b) = 1$. Therefore the hyperplane is: $w^* \bullet x + b^* = 0$, and a new given data will be classified as $f(x) = \text{sign}(\langle w^*, x \rangle + b^*)$.

More Flexible Case

For non-separable case, we introduce slack variable $\epsilon_i \geq 0$ for each data (x_i, y_i) , such that: $y_i(w \bullet x_i + b) \geq 1 - \xi_i$.

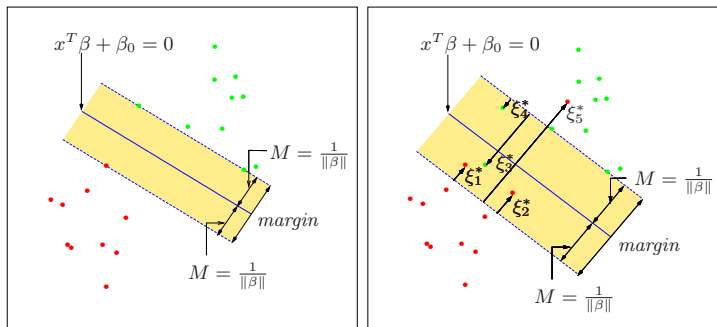


FIGURE 12.1. Support vector classifiers. The left panel shows the separable case. The decision boundary is the solid line, while broken lines bound the shaded maximal margin of width $2M = 2/\|\beta\|$. The right panel shows the nonseparable (overlap) case. The points labeled ξ_j^* are on the wrong side of their margin by an amount $\xi_j^* = M \xi_j$; points on the correct side have $\xi_j^* = 0$. The margin is maximized subject to a total budget $\sum \xi_i \leq \text{constant}$. Hence $\sum \xi_j^*$ is the total distance of points on the wrong side of their margin.

New Objective

$$\begin{aligned} \min \quad & \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^m \xi_i \\ \text{s.t.} \quad & y_i(w^T x_i + b) \geq 1 - \xi_i \quad (i = 1, 2, \dots, m) \\ & \xi_i \geq 0 \quad (i = 1, 2, \dots, m) \end{aligned} \tag{11}$$

Similar to separable case above following Lagrangian Multipliers, we have:

$$L(w, b, \xi, \alpha, \mu) = \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^m \xi_i - \sum_{i=1}^m \alpha_i [y_i(w^T x_i + b) - 1 + \xi_i] - \sum_{i=1}^m \mu_i \xi_i \tag{12}$$

where $\mu_i \geq 0, \alpha_i \geq 0$ are Lagrangian variables.