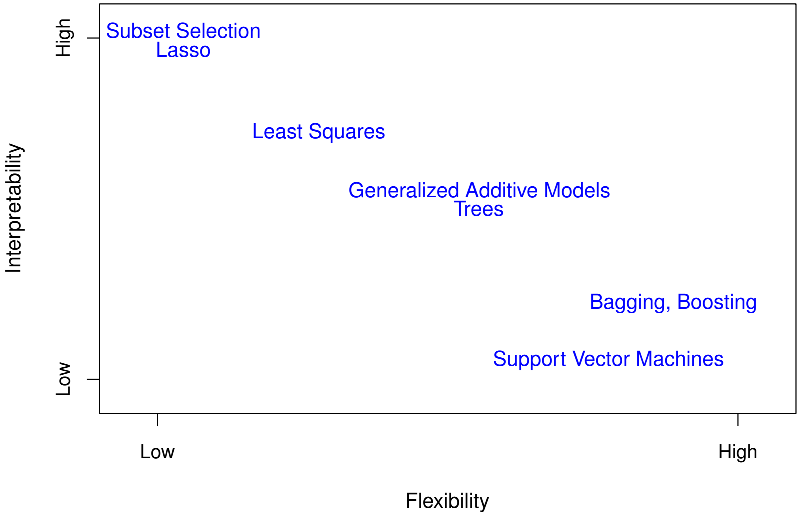
CPSC 8420 Advanced Machine Learning
Week 9: Support Vector Machines

Dr. Kai Liu

October 15, 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:

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

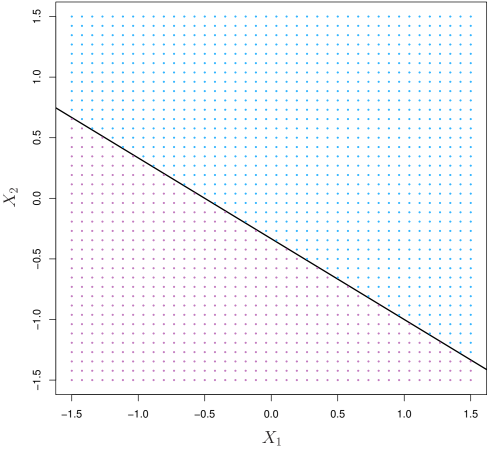
- In p dimensions, a hyperplane is difficult to visualize, but can easily be defined by extending the above equations to:

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

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:

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

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:

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

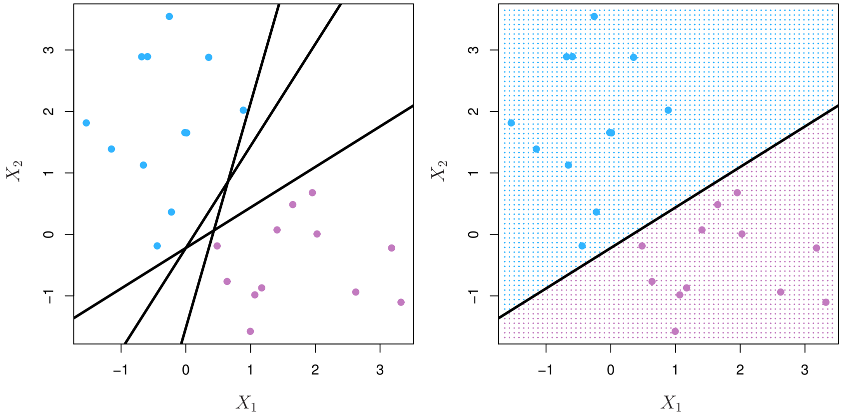
$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} < 0 \quad \text{if } y_i = -1 \tag{2}$$

- This can be written as:

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

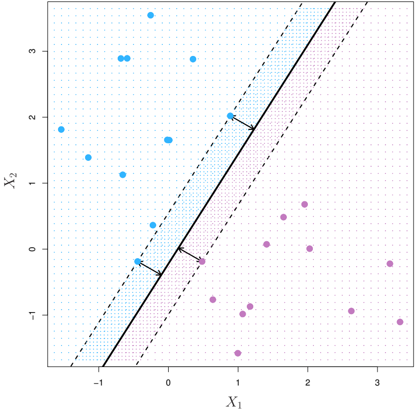
for all $i = 1, \ldots, n$.

Maximum Margin Classifier



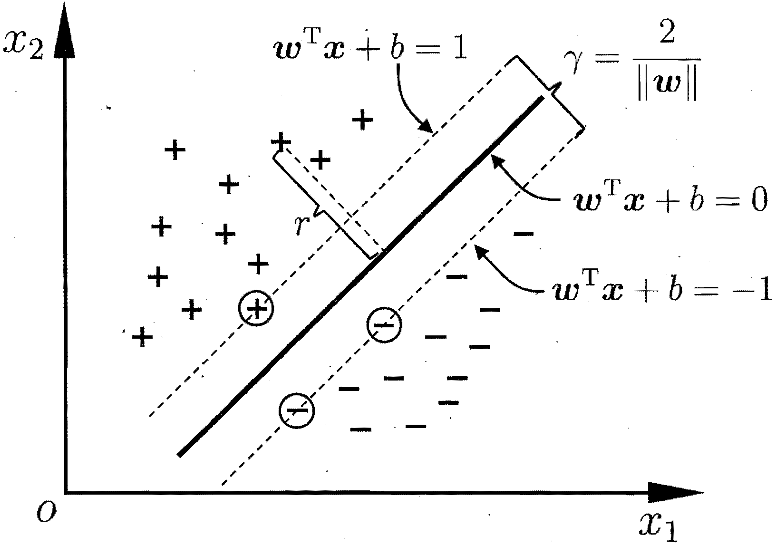
Example of data that can be perfectly separated using a hyperplane.
Problem: there exists an infinite number of such hyperplanes.

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:

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

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

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

Optimization

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

$$\underbrace{min}_{w,b} \underbrace{max}_{\alpha_i \geq 0} L(w,b,\alpha) \tag{6}$$

$$\underbrace{max}_{\alpha_i \geq 0} \underbrace{min}_{w,b} L(w,b,\alpha)$$

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

Optimization

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

Optimization

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

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

Optimization

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

which is equivalent to:

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

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 arbitary (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 conidtion: $\alpha_i^*(y_i(w^T x_i + b) - 1) = 0$, 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 $f(x) = sign(\langle w^*, x \rangle + b^*)$.

More Flexible Case

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

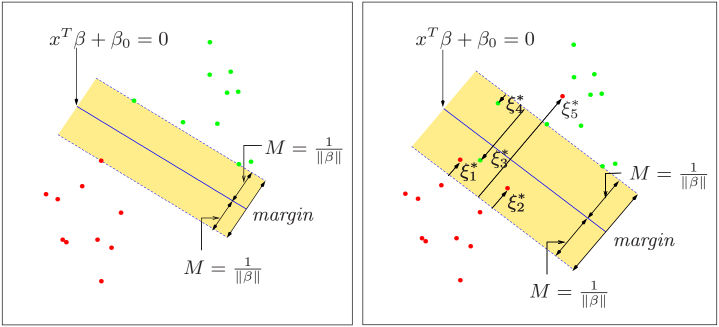


FIGURE 12.1. *Support vector classifiers. The left panel shows the separable case. The decision boundary is the solid line, while broken lines bound the shaded maximal margin of width $2M = 2/\|\beta\|$. The right panel shows the nonseparable (overlap) case. The points labeled ξ_j^* are on the wrong 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 $\sum \xi_i \leq$ constant. Hence $\sum \xi_j^*$ is the total distance of points on the wrong side of their margin.*

New Objective

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

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.

Optimization

If we denote $J(w,b,\xi) = \frac{1}{2}||w||_2^2 + C\sum_{i=1}^{m}\xi_i$, then we have:

$$J = \underbrace{max}_{\alpha_i \geq 0, \mu_i \geq 0,} \quad L(w,b,\alpha,\xi,\mu), \tag{13}$$

therefore, $\min J(w,b,\xi) = \underbrace{min}_{w,b,\xi} \quad \underbrace{max}_{\alpha_i \geq 0, \mu_i \geq 0,} \quad L(w,b,\alpha,\xi,\mu)$. According to KKT condition, it is equivalent to $\underbrace{max}_{\alpha_i \geq 0, \mu_i \geq 0,} \quad \underbrace{min}_{w,b,\xi} L(w,b,\alpha,\xi,\mu)$.

Karush-Kuhn-Tucker (KKT) conditions

KKT is an extension for Lagrangian Multipliers, where there are inequality as well as equality constriants. assume we are given:

$$\begin{aligned} \min_w \quad & f(w) \\ \text{s.t.} \quad & g_i(w) \leq 0, \quad i = 1, \dots, k \\ & h_i(w) = 0, \quad i = 1, \dots, l. \end{aligned}$$

To solve it, we start by defining the generalized Lagrangian:

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w).$$

Karush-Kuhn-Tucker (KKT) conditions

KKT will tell us the necessary condition of the minimizers:

$$
\begin{aligned}
\frac{\partial}{\partial w_i}\mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \ldots, d \\
\frac{\partial}{\partial \beta_i}\mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \ldots, l \\
\alpha_i^* g_i(w^*) &= 0, \quad i = 1, \ldots, k \\
g_i(w^*) &\leq 0, \quad i = 1, \ldots, k \\
\alpha^* &\geq 0, \quad i = 1, \ldots, k
\end{aligned}
$$

Karush-Kuhn-Tucker (KKT) Practice

$$\min \frac{1}{2}(x^2 + y^2), \quad s.t. \quad x + y \geq 4 \tag{14}$$

Karush-Kuhn-Tucker (KKT) Practice

$$\min \frac{1}{2}(x^2 + y^2), \quad s.t. \quad x + 2y \leq 4 \tag{15}$$

Optimization

For $\underbrace{max}_{\alpha_i \geq 0, \mu_i \geq 0,} \underbrace{min}_{w,b,\xi} L(w, b, \alpha, \xi, \mu)$, now we can first optimize $\underbrace{min}_{w,b,\xi} L(w, b, \alpha, \xi, \mu)$ by taking the derivative and set it to be zero:

$$\begin{aligned}
\frac{\partial L}{\partial w} &= 0 \ \Rightarrow w = \sum_{i=1}^{m} \alpha_i y_i x_i \\
\frac{\partial L}{\partial b} &= 0 \ \Rightarrow \sum_{i=1}^{m} \alpha_i y_i = 0 \\
\frac{\partial L}{\partial \xi} &= 0 \ \Rightarrow C - \alpha_i - \mu_i = 0
\end{aligned} \tag{16}$$

Optimization

Now pluggin the above to L:

$$L(w,b,\xi,\alpha,\mu) = \frac{1}{2}||w||_2^2 + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i[y_i(w^Tx_i+b)-1+\xi_i] - \sum_{i=1}^{m}\mu_i\xi_i \tag{17}$$

we have:

$$\begin{aligned} L(w,b,\xi,\alpha,\mu) &= \frac{1}{2}||w||_2^2 + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i[y_i(w^Tx_i+b)-1+\xi_i] - \sum_{i=1}^{m}\mu_i\xi_i \\ &= \frac{1}{2}||w||_2^2 - \sum_{i=1}^{m}\alpha_i[y_i(w^Tx_i+b)-1+\xi_i] + \sum_{i=1}^{m}\alpha_i\xi_i \\ &= \frac{1}{2}||w||_2^2 - \sum_{i=1}^{m}\alpha_i[y_i(w^Tx_i+b)-1] \\ &= \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1,j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j \end{aligned} \tag{18}$$

Optimization

$$\begin{aligned}
\underbrace{max}_{\alpha} \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1,j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j \\
s.t.\ \sum_{i=1}^{m}\alpha_i y_i = 0 \\
C - \alpha_i - \mu_i = 0 \\
\alpha_i \geq 0\ (i = 1, 2, ..., m) \\
\mu_i \geq 0\ (i = 1, 2, ..., m)
\end{aligned} \tag{19}$$

which is equivalent to:

$$\begin{aligned}
\underbrace{min}_{\alpha} \frac{1}{2}\sum_{i=1,j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j - \sum_{i=1}^{m}\alpha_i \\
s.t.\ \sum_{i=1}^{m}\alpha_i y_i = 0 \\
0 \leq \alpha_i \leq C
\end{aligned} \tag{20}$$

Classification Discussion

According to the complementary slackness condition, we have:

$$\begin{aligned} \alpha_i^*(y_i(w^T x_i + b) - 1 + \xi_i^*) &= 0 \\ \mu_i \xi_i &= 0 \end{aligned} \tag{21}$$

also, since $C = \alpha_i + \mu_i$, we now discuss different cases:

1. $\alpha_i = 0 \implies \mu_i = C \implies \xi_i = 0$.
2. $0 < \alpha_i < C \implies y_i(w^T x_i + b) - 1 + \xi_i^* = 0$, also $\mu_i > 0 \implies \xi_i = 0$, then $y_i(w^T x_i + b) - 1 = 0$, which is the support vector.
3. $\alpha_i = C$
    1. $0 < \xi_i < 1$, it is correctly classified but on the wrong side of its margin.
    2. $\xi_i = 1$, it lies exactly on its margin.
    3. $\xi_i > 1$, misclassified.

Classification Discussion