Support Vector Machines

Ziping Zhao

School of Information Science and Technology
ShanghaiTech University, Shanghai, China

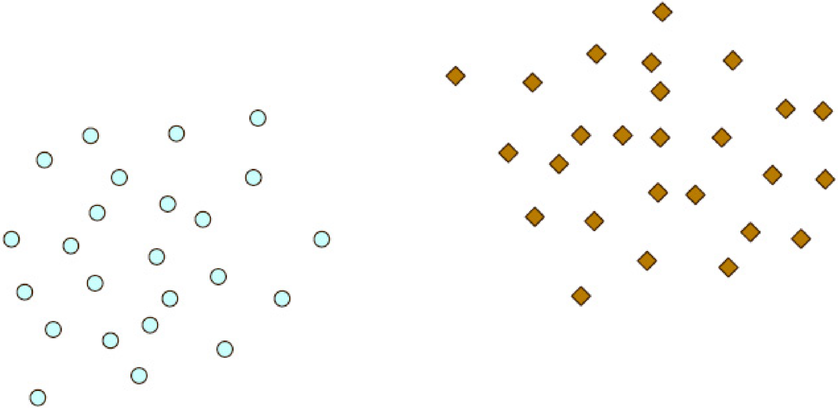
CS182: Introduction to Machine Learning (Fall 2022)
http://cs182.sist.shanghaitech.edu.cn

Ch. 14 of I2ML (Secs. 14.4, 14.7 – 14.9, and 14.11 – 14.14 excluded)

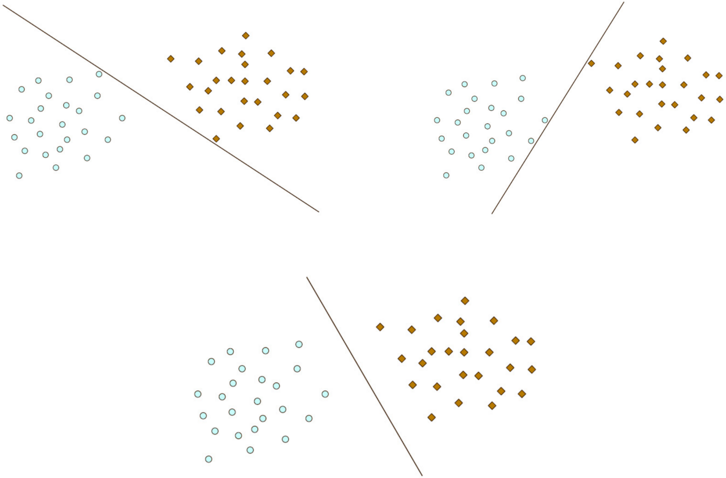
Outline

Outline

Binary Classification given a Sample $\mathcal{X} = \{(\mathbf{x}^t, r^t)\}$ ...

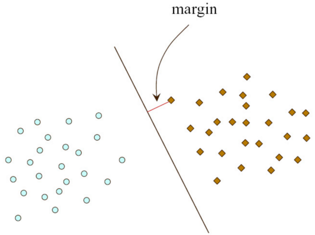


… Which Separating Hyperplane is the Best?



Optimal Separating Hyperplane

- An instance $\mathbf{x}_i$ is represented as a vector in the space.
- We are using the hypothesis class of lines denoted as separating hyperplanes.
- Margin of a separating hyperplane: distance to the separating hyperplane from the data point closest to it on either side.

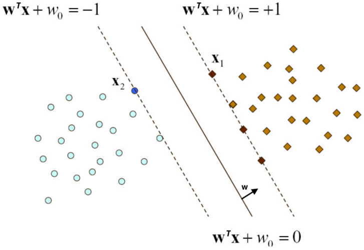


- Relationship between margin and generalization:
  There exist theoretical results from statistical learning theory showing that the separating hyperplane with the largest margin generalizes best (i.e., has smallest generalization error).

Outline

Optimal Canonical Separating Hyperplane – I

- Hard-margin case: data points from the two classes are assumed to be linearly separable.
- Note that $(c\mathbf{w})^T\mathbf{x} + cw_0 = 0$ with $c \neq 0$ defines the same hyperplane as $\mathbf{w}^T\mathbf{x} + w_0 = 0$.
- With proper scaling of $\mathbf{w}$ and w_0, the points closest to the hyperplane satisfy $|\mathbf{w}^T\mathbf{x} + w_0| = 1$. Such a hyperplane is called a canonical separating hyperplane.
- The one that maximizes the margin is called the optimal canonical separating hyperplane.



Optimal Canonical Separating Hyperplane – II

- Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be two closest points, one on each side of the hyperplane.
- Note that

$$\mathbf{w}^T\mathbf{x}_1 + w_0 = +1$$
$$\mathbf{w}^T\mathbf{x}_2 + w_0 = -1$$

  Hence the margin can be given by

$$\gamma = \frac{|\mathbf{w}^T\mathbf{x}_1 + w_0|}{\|\mathbf{w}\|} = \frac{|\mathbf{w}^T\mathbf{x}_2 + w_0|}{\|\mathbf{w}\|} = \frac{1}{\|\mathbf{w}\|}$$

- Maximizing the margin is equivalent to minimizing $\|\mathbf{w}\|$.

Inequality Constraints

- Let us start again with two classes and use labels $+1/-1$ for the two classes.
- The sample is $\mathcal{X} = \{(\mathbf{x}^t, r^t)\}$ where $r^t = +1$ if $\mathbf{x}^t \in C_1$ and $r^t = -1$ if $\mathbf{x}^t \in C_2$.
- For all data points in the sample $\mathcal{X}$, we want $\mathbf{w}$ and w_0 to satisfy

$$\mathbf{w}^T\mathbf{x}^t + w_0 \begin{cases} \geq +1 & \text{if } r^t = +1 \\ \leq -1 & \text{if } r^t = -1 \end{cases}$$

which are equivalent to the following inequality constraints:

$$r^t(\mathbf{w}^T\mathbf{x}^t + w_0) \geq 1, \quad \forall t \tag{1}$$

- Instead of simply using inequality constraints

$$r^t(\mathbf{w}^T\mathbf{x}^t + w_0) \geq 0$$

which only require the data points to lie on the right side of the hyperplane, the constraints in (1) also want them some distance away for better generalization.

Optimization Problem – I

- Optimization problem (the primal problem):

$$\begin{aligned} \underset{\mathbf{w}\in\mathbf{R}^d,\, w_0}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{subject to} \quad & r^t(\mathbf{w}^T\mathbf{x}^t + w_0) \geq 1, \quad \forall t \end{aligned}$$

- This is a convex quadratic programming (QP), the complexity of which depends on d.
  - This QP can be solved directly via QP numerical solving methods to find $\mathbf{w}$ and w_0, i.e., the optimal canonical separating hyperplane.
- On both sides of the hyperplane, there will be instances that are $\frac{1}{\|\mathbf{w}\|}$ away from the hyperplane and the total margin will be $\frac{2}{\|\mathbf{w}\|}$.

Optimization Problem – II

- As discussed in previous lectures, if the classification problem is not linearly separable, instead of fitting a nonlinear function, one trick is to map the problem to a new space $\mathcal{Z}$ by using nonlinear basis functions.
  - It is generally the case that this new space has more dimensions than the original space (i.e., larger than d), and, in such a case, we are interested in a method whose complexity does not depend on the input dimensionality.
- In optimization theory, it is very common and sometimes advantageous to turn the primal problem into a dual problem and then solve the latter instead.
  - In our case, it also turns out to be more convenient to solve the dual problem (whose complexity depends on the sample size N) rather than the primal problem directly (whose complexity depends on the dimensionality d).
- It will be shown that the dual problem also makes it easy for a nonlinear extension using kernel functions.

Lagrangian

- Lagrangian:

$$
\begin{aligned}
\mathcal{L}(\mathbf{w}, w_0, \{\alpha_t\}) =& \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{t=1}^{N} \alpha_t \left[r^t(\mathbf{w}^T\mathbf{x}^t + w_0) - 1 \right] \\
=& \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{t=1}^{N} \alpha_t r^t(\mathbf{w}^T\mathbf{x}^t + w_0) + \sum_{t=1}^{N} \alpha_t \\
=& \frac{1}{2}\mathbf{w}^T\mathbf{w} - \mathbf{w}^T \sum_{t=1}^{N} \alpha_t r^t \mathbf{x}^t - w_0 \sum_{t=1}^{N} \alpha_t r^t + \sum_{t=1}^{N} \alpha_t
\end{aligned}
$$

  with Lagrange multipliers $\alpha_t \geq 0$.
- The optimal solution is a saddle point which minimizes $\mathcal{L}$ w.r.t. the primal variables $\mathbf{w}, w_0$ and maximizes $\mathcal{L}$ w.r.t. the dual variables α_t.

Eliminating Primal Variables

- Setting the derivatives of $\mathcal{L}$ w.r.t. $\mathbf{w}$ and w_0 to $\mathbf{0}$:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{0} \quad \Rightarrow \quad \mathbf{w} = \sum_{t=1}^{N} \alpha_t r^t \mathbf{x}^t \tag{2}$$

$$\frac{\partial \mathcal{L}}{\partial w_0} = 0 \quad \Rightarrow \quad \sum_{t=1}^{N} \alpha_t r^t = 0 \tag{3}$$

- Plugging (2) and (3) into $\mathcal{L}$ gives the objective function G for the dual problem:

$$\begin{aligned} G(\{\alpha_t\}) &= -\frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{t=1}^{N} \alpha_t \\ &= -\frac{1}{2}\sum_{t=1}^{N}\sum_{t'=1}^{N} \alpha_t \alpha_{t'} r^t r^{t'} (\mathbf{x}^t)^T \mathbf{x}^{t'} + \sum_{t=1}^{N} \alpha_t \end{aligned}$$

Dual Optimization Problem – I

- Dual optimization problem:

$$\begin{aligned}
\underset{\{\alpha_t\}}{\text{maximize}} \quad & \sum_{t=1}^{N} \alpha_t - \frac{1}{2} \sum_{t=1}^{N} \sum_{t'=1}^{N} \alpha_t \alpha_{t'} r^t r^{t'} (\mathbf{x}^t)^T \mathbf{x}^{t'} \\
\text{subject to} \quad & \sum_{t=1}^{N} \alpha_t r^t = 0 \\
& \alpha_t \geq 0, \quad \forall t
\end{aligned}$$

- This is also a QP problem, and its complexity depends on the sample size N (rather than the input dimensionality d):
  - Time complexity: $O(N^3)$ (for generic QP solvers)
  - Space complexity: $O(N^2)$

Dual Optimization Problem – II

- Define

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{bmatrix}, \quad \mathbf{r} = \begin{bmatrix} r^1 \\ \vdots \\ r^N \end{bmatrix},$$

  and the symmetric matrix $\mathbf{H} \in \mathbb{R}^{N \times N}$ with $h_{ij} = r^i r^j (\mathbf{x}^i)^T \mathbf{x}^j$.
- We get the equivalent reformulation

$$\begin{aligned} \underset{\boldsymbol{\alpha}}{\text{maximize}} \quad & \boldsymbol{\alpha}^T \mathbf{1} - \frac{1}{2} \boldsymbol{\alpha}^T \mathbf{H} \boldsymbol{\alpha} \\ \text{subject to} \quad & \boldsymbol{\alpha}^T \mathbf{r} = 0 \\ & \boldsymbol{\alpha} \geq \mathbf{0} \end{aligned}$$

Support Vectors

- Based on KKT complementarity slackness condition, we have the following results.
- For points lying beyond the margin (sufficiently away from the hyperplane), i.e., $r^t(\mathbf{w}^T\mathbf{x}^t + w_0) > 1$, they have no effect on the hyperplane. The corresponding dual variables vanish with $\alpha_t = 0$.
  - Even if any subset of them are removed or moved around, we would still get the same solution.
  - It is possible to use a simpler classifier to filter out a large portion of such instances, i.e., decreasing N, thereby decreasing the complexity of the optimization.
- Support vectors (SVs): $\mathbf{x}^t$ with $\alpha_t > 0$, i.e., $r^t(\mathbf{w}^T\mathbf{x}^t + w_0) = 1$ (exactly on the hyperplane), hence the name support vector machine (SVM).
  - Solution is determined by the data on the margin.

Computation of Primal Variables

- From (2) we get

$$\mathbf{w} = \sum_{t=1}^{N} \alpha_t r^t \mathbf{x}^t = \sum_{\mathbf{x}^t \in \mathcal{SV}} \alpha_t r^t \mathbf{x}^t$$

  where $\mathcal{SV}$ denotes the set of support vectors.
- The support vectors must lie on the margin, so they should satisfy

$$r^t(\mathbf{w}^T \mathbf{x}^t + w_0) = 1.$$

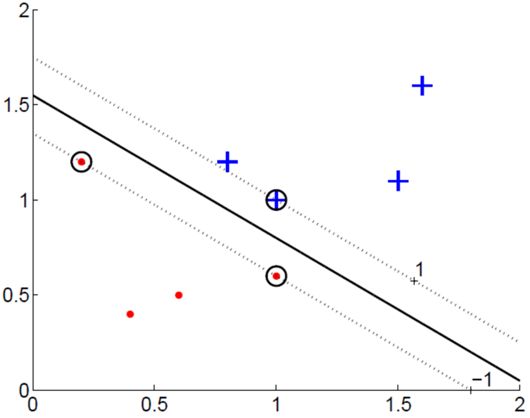
  Then, we have

$$w_0 = r^t - \mathbf{w}^T \mathbf{x}^t.$$

  – For numerical stability, in practice all support vectors are used to compute w_0:

$$w_0 = \frac{1}{|\mathcal{SV}|} \sum_{\mathbf{x}^t \in \mathcal{SV}} (r^t - \mathbf{w}^T \mathbf{x}^t)$$

Hard-Margin Support Vector Machine



Discriminant function

- Discriminant function:

$$
\begin{aligned}
g(\mathbf{x}) &= \mathbf{w}^T \mathbf{x} + w_0 \\
&= \left[\sum_{\mathbf{x}^t \in \mathcal{SV}} \alpha_t r^t \mathbf{x}^t \right]^T \mathbf{x} + \frac{1}{|\mathcal{SV}|} \sum_{\mathbf{x}^t \in \mathcal{SV}} (r^t - \mathbf{w}^T \mathbf{x}^t)
\end{aligned}
$$

- During testing, we do not enforce a margin and obtain the classification rule :

$$
\text{Choose } \begin{cases} C_1 & \text{if } g(\mathbf{x}) > 0 \\ C_2 & \text{otherwise} \end{cases}
$$

Generalization to $K > 2$ Classes

- One way to handle multiple classes is to define K two-class problems, each separating one class from all other classes combined, i.e., the one-vs.-all approach.
- An SVM $g_i(\mathbf{x})$ is learned for each two-class problem.
- Classification rule during testing:

$$\text{Choose } C_j \text{ if } j = \arg\max_k g_k(\mathbf{x})$$

- We can also define pairwise separation of classes by training $\frac{K(K-1)}{2}$ SVMs, i.e., the one-vs.-one approach.