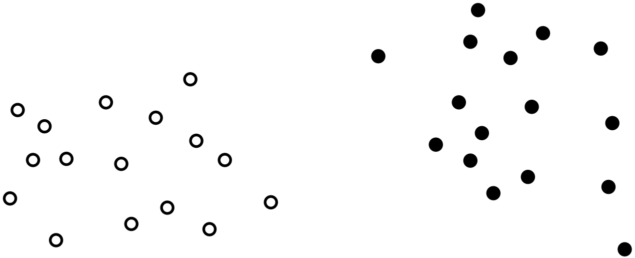
Support Vector Machines

Dit-Yan Yeung

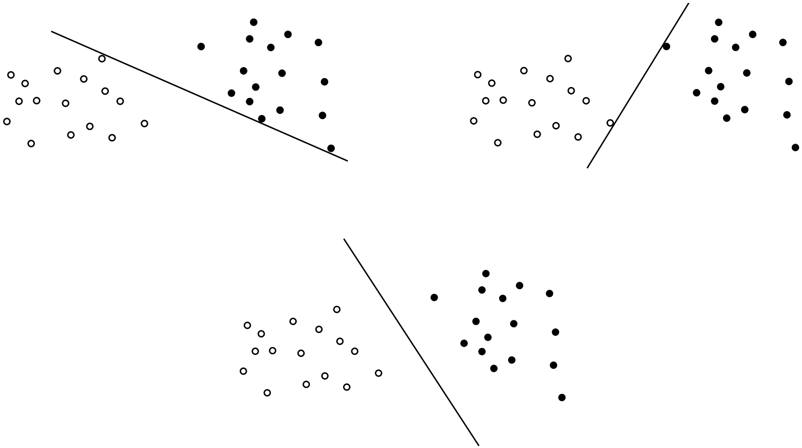
Department of Computer Science and Engineering
Hong Kong University of Science and Technology

COMP 4211: Machine Learning (Fall 2022)

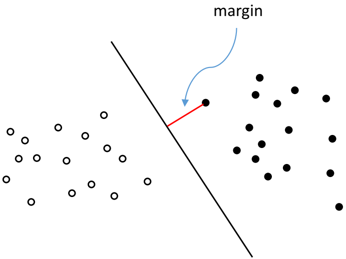
Given a Data Set ...



... Which Separating Hyperplane is the Best?

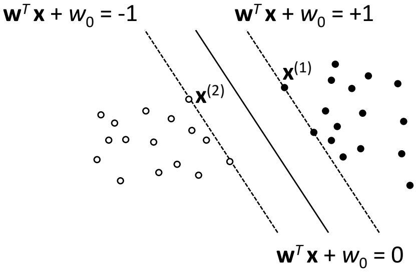


Optimal Separating Hyperplane



- Margin of a separating hyperplane: distance to the separating hyperplane from the data point closest to it.
- Relationship between margin and generalization: There exist theoretical results showing that the separating hyperplane with the largest margin generalizes best (i.e., has smallest generalization error).

Canonical Optimal Separating Hyperplane



- Hard-margin case: data points from the two classes are assumed to be linearly separable.
- With proper scaling of $\mathbf{w}$ and w_0, the points closest to the hyperplane satisfy $|\mathbf{w}^\top\mathbf{x} + w_0| = 1$. Such a hyperplane is called a canonical separating hyperplane.
- The one that maximizes the margin is called the canonical optimal separating hyperplane.

Canonical Optimal Separating Hyperplane (2)

- Let $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ be two closest points, one on each side of the hyperplane.
- Note that $\mathbf{w}$ is a normal vector to the hyperplane (i.e., its direction is perpendicular to that of the hyperplane) and

$$\begin{aligned}\mathbf{w}^\top \mathbf{x}^{(1)} + w_0 &= +1 \\ \mathbf{w}^\top \mathbf{x}^{(2)} + w_0 &= -1,\end{aligned}$$

  which imply

$$\mathbf{w}^\top (\mathbf{x}^{(1)} - \mathbf{x}^{(2)}) = 2.$$

  Hence the margin can be given by

$$\gamma = \frac{1}{2} \frac{\mathbf{w}^\top (\mathbf{x}^{(1)} - \mathbf{x}^{(2)})}{\|\mathbf{w}\|} = \frac{1}{\|\mathbf{w}\|}.$$

- Thus, maximizing the margin is equivalent to minimizing $\|\mathbf{w}\|$.

Inequality Constraints

- For all data points in the sample $\mathcal{X} = \{(\mathbf{x}^{(\ell)}, y^{(\ell)})\}$, we want $\mathbf{w}$ and w_0 to satisfy

$$\mathbf{w}^\top \mathbf{x}^{(\ell)} + w_0 \begin{cases} \geq +1 & \text{if } y^{(\ell)} = +1 \\ \leq -1 & \text{if } y^{(\ell)} = -1. \end{cases}$$

- Equivalent form of inequality constraints:

$$y^{(\ell)}(\mathbf{w}^\top \mathbf{x}^{(\ell)} + w_0) \geq 1. \tag{1}$$

- Instead of using inequality constraints

$$y^{(\ell)}(\mathbf{w}^\top \mathbf{x}^{(\ell)} + w_0) \geq 0,$$

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

Primal Optimization Problem

- Primal optimization problem:

$$\begin{array}{ll} \text{Minimize} & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{subject to} & y^{(\ell)}(\mathbf{w}^\top \mathbf{x}^{(\ell)} + w_0) \geq 1, \ \forall \ell. \end{array}$$

- This is a quadratic programming (QP) problem, or a quadratic program, which is a type of convex optimization problem.
- In optimization theory, it is very common and sometimes advantageous to turn a 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). The dual problem also makes it easy for a nonlinear extension using kernel functions.

Dual Optimization Problem

- By introducing N Lagrange multipliers, $\{\alpha_\ell\}_{\ell=1}^N$, one for each training data point, we can turn the primal problem into a dual problem (with details of the derivation skipped). The Lagrange multipliers are known as dual variables in the dual problem.
- Dual optimization problem:

$$\begin{array}{ll} \text{Maximize} & \displaystyle\sum_\ell \alpha_\ell - \frac{1}{2}\sum_\ell\sum_{\ell'} \alpha_\ell\alpha_{\ell'} y^{(\ell)} y^{(\ell')} (\mathbf{x}^{(\ell)})^\top \mathbf{x}^{(\ell')} \\ \text{subject to} & \displaystyle\sum_\ell \alpha_\ell y^{(\ell)} = 0 \text{ and } \alpha_\ell \geq 0, \forall \ell. \end{array}$$

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

Support Vectors

- At the optimal solution, most of the dual variables vanish with $\alpha_\ell = 0$. They are points lying beyond the margin with no effect on the hyperplane.
- Support vectors: $\mathbf{x}^{(\ell)}$ with $\alpha_\ell > 0$, hence the name support vector machine (SVM).

Computation of Primal Variables from Dual Variables

- From the (skipped) derivation of the dual problem, we get

$$\mathbf{w} = \sum_{\ell=1}^{N} \alpha_\ell y^{(\ell)} \mathbf{x}^{(\ell)} = \sum_{\mathbf{x}^{(\ell)} \in \mathcal{SV}} \alpha_\ell y^{(\ell)} \mathbf{x}^{(\ell)},$$

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

$$y^{(\ell)}(\mathbf{w}^\top \mathbf{x}^{(\ell)} + w_0) = 1 \quad \text{or} \quad w_0 = y^{(\ell)} - \mathbf{w}^\top \mathbf{x}^{(\ell)}.$$

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

$$w_0 = \frac{1}{|\mathcal{SV}|} \sum_{\mathbf{x}^{(\ell)} \in \mathcal{SV}} \left(y^{(\ell)} - \mathbf{w}^\top \mathbf{x}^{(\ell)} \right).$$

Discriminant Function

- Discriminant function:

$$
\begin{aligned}
g(\mathbf{x}) &= \mathbf{w}^\top \mathbf{x} + w_0 \\
&= \left(\sum_{\mathbf{x}^{(\ell)} \in \mathcal{SV}} \alpha_\ell y^{(\ell)} \mathbf{x}^{(\ell)} \right)^\top \mathbf{x} + \frac{1}{|\mathcal{SV}|} \sum_{\mathbf{x}^{(\ell)} \in \mathcal{SV}} \left(y^{(\ell)} - \mathbf{w}^\top \mathbf{x}^{(\ell)} \right).
\end{aligned}
$$

- Classification rule during testing:

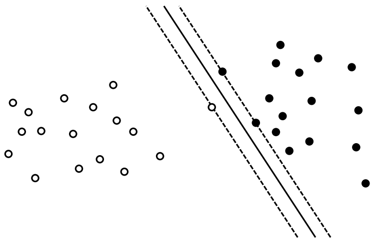
$$
\text{Choose } \begin{cases} C_1 & \text{if } g(\mathbf{x}) > 0 \\ C_2 & \text{otherwise.} \end{cases}
$$

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.
- Let $g_i(\mathbf{x})$ denote $\mathbf{w}_i^\top \mathbf{x} + \mathbf{w}_{i0}$.
- An SVM $g_i(\mathbf{x})$ is learned for each two-class problem.
- Classification rule during testing:

$$\text{Classify to class } j \text{ if } j = \arg\max_{1 \le k \le K} g_k(\mathbf{x}).$$

Relaxing the Constraints



- In practice, a separating hyperplane may not exist, possibly due to a high noise level which causes a large overlap of the classes.
- Even if a separating hyperplane exists, it is not always the best solution to the classification problem when there exist outliers in the data.
- A mislabeled example can become an outlier which affects the location of the separating hyperplane.

Slack Variables

- A soft-margin SVM allows for the possibility of violating the inequality constraints

$$y^{(\ell)}(\mathbf{w}^\top \mathbf{x}^{(\ell)} + w_0) \geq 1$$

by introducing slack variables

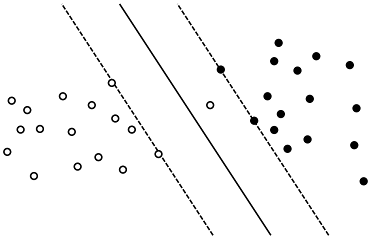
$$\xi_\ell \geq 0, \;\; \ell = 1, \ldots, N,$$

which store the deviation from the margin.

- Relaxed separation constraints:

$$y^{(\ell)}(\mathbf{w}^\top \mathbf{x}^{(\ell)} + w_0) \geq 1 - \xi_\ell.$$

Relaxed Separation Constraints



- One data point is inside the margin on the wrong side of the hyperplane:

Penalty

- By making ξ_ℓ large enough, the constraint on $(\mathbf{x}^{(\ell)}, y^{(\ell)})$ can always be met. In order not to obtain the trivial solution where all ξ_ℓ take on large values, we should penalize them in the objective function.
- Four cases:
  - $\xi_\ell = 0$: no problem with $\mathbf{x}^{(\ell)}$ (no penalty)
  - $0 < \xi_\ell < 1$: $\mathbf{x}^{(\ell)}$ lies on the right side of the hyperplane but in the margin (small penalty)
  - $\xi_\ell = 1$: $\mathbf{x}^{(\ell)}$ lies at the hyperplane (penalty between the previous and next cases)
  - $\xi_\ell > 1$: $\mathbf{x}^{(\ell)}$ lies on the wrong side of the hyperplane (large penalty)
- Number of misclassifications: $\#\{\xi_\ell > 1\}$
- Number of nonseparable instances: $\#\{\xi_\ell > 0\}$
- Soft error as additional penalty term:

$$\sum_\ell \xi_\ell$$

Primal and Dual Optimization Problems

- Primal optimization problem:

$$\begin{aligned} \text{Minimize} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_\ell \xi_\ell \\ \text{subject to} \quad & y^{(\ell)}(\mathbf{w}^\top \mathbf{x}^{(\ell)} + w_0) \geq 1 - \xi_\ell, \ \forall \ell \\ & \xi_\ell \geq 0, \ \forall \ell. \end{aligned}$$

- Dual optimization problem:

$$\begin{aligned} \text{Maximize} \quad & \sum_\ell \alpha_\ell - \frac{1}{2}\sum_\ell \sum_{\ell'} \alpha_\ell \alpha_{\ell'} y^{(\ell)} y^{(\ell')} (\mathbf{x}^{(\ell)})^\top \mathbf{x}^{(\ell')} \\ \text{subject to} \quad & \sum_\ell \alpha_\ell y^{(\ell)} = 0 \text{ and } 0 \leq \alpha_\ell \leq C, \forall \ell. \end{aligned}$$

Alternative View as Unconstrained Optimization Problems

- An alternative view is to formulate the learning of hard-margin and soft-margin SVMs as unconstrained optimization problems.
- Each optimization problem minimizes a loss function with a regularizer, i.e., a regularized loss function.
- Different loss functions are used for the hard-margin and soft-margin SVMs.

Alternative Formulation for Hard-Margin SVM

- Objective function for minimization:

$$\sum_{\ell=1}^{N} E_\infty\left(y^{(\ell)} g(\mathbf{x}^{(\ell)}) - 1\right) + \lambda \|\mathbf{w}\|^2,$$

where

$$E_\infty(z) = \begin{cases} 0 & z \geq 0 \\ \infty & \text{otherwise.} \end{cases}$$

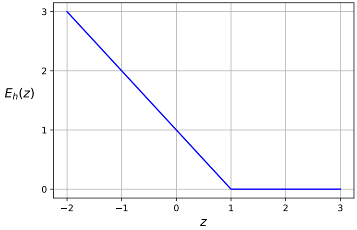
- Equivalent formulation that is more consistent with that of the soft-margin SVM:

$$\sum_{\ell=1}^{N} E_\infty\left(y^{(\ell)} g(\mathbf{x}^{(\ell)})\right) + \lambda \|\mathbf{w}\|^2,$$

where

$$E_\infty(z) = \begin{cases} 0 & z \geq 1 \\ \infty & \text{otherwise.} \end{cases}$$

Alternative Formulation for Soft-Margin SVM



- Objective function for minimization:

$$\sum_{\ell=1}^{N} E_h\left(y^{(\ell)} g(\mathbf{x}^{(\ell)})\right) + \lambda \|\mathbf{w}\|^2,$$

where

$$E_h(z) = [1 - z]_+ = \max(0, 1 - z)$$

is called the hinge loss.

Key Ideas of Kernel Methods

- Instead of defining a nonlinear model in the original (input) space, the problem is mapped to a new (feature) space by performing a nonlinear transformation using suitably chosen basis functions.
- A linear model is then applied in the new space.
- The basis functions are often defined implicitly via defining kernel functions directly.

Basis Functions

- Basis functions:

$$\mathbf{z} = \phi(\mathbf{x}) \quad \text{where } z_j = \phi_j(\mathbf{x}),\ j = 1, \ldots, k.$$

- Discriminant function:

$$g(\mathbf{z}) = \mathbf{w}^\top \mathbf{z}$$

$$g(\mathbf{x}) = \mathbf{w}^\top \phi(\mathbf{x}) = \sum_{j=1}^{k} w_j \phi_j(\mathbf{x}),$$

  where we do not use a separate w_0 but assume that $z_1 = \phi_1(\mathbf{x}) \equiv 1$.

- Usually, $k \gg d, N$ (in fact k can even be infinite). The dual form is preferred because its complexity depends on N but that of the primal form depends on k.

Kernel Functions

- Dual optimization problem:

$$
\begin{aligned}
\text{Maximize} \quad & \sum_{\ell} \alpha_\ell - \frac{1}{2} \sum_{\ell} \sum_{\ell'} \alpha_\ell \alpha_{\ell'} y^{(\ell)} y^{(\ell')} \phi(\mathbf{x}^{(\ell)})^\top \phi(\mathbf{x}^{(\ell')}) \\
\text{subject to} \quad & \sum_{\ell} \alpha_\ell y^{(\ell)} = 0 \text{ and } 0 \leq \alpha_\ell \leq C, \forall \ell,
\end{aligned}
$$

or

$$
\begin{aligned}
\text{Maximize} \quad & \sum_{\ell} \alpha_\ell - \frac{1}{2} \sum_{\ell} \sum_{\ell'} \alpha_\ell \alpha_{\ell'} y^{(\ell)} y^{(\ell')} K(\mathbf{x}^{(\ell)}, \mathbf{x}^{(\ell')}) \\
\text{subject to} \quad & \sum_{\ell} \alpha_\ell y^{(\ell)} = 0 \text{ and } 0 \leq \alpha_\ell \leq C, \forall \ell,
\end{aligned}
$$

where $K(\mathbf{x}^{(\ell)}, \mathbf{x}^{(\ell')}) \equiv \phi(\mathbf{x}^{(\ell)})^\top \phi(\mathbf{x}^{(\ell')})$ is a kernel function defined directly on the inputs $\mathbf{x}^{(\ell)}$ and $\mathbf{x}^{(\ell')}$.

Some Common Kernel Functions: Polynomial Kernel

- Polynomial kernel:

$$K(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^\top \mathbf{x}' + 1)^q,$$

  where q is the degree.

- E.g., when $q = 2$ and $d = 2$,

$$\begin{aligned} K(\mathbf{x}, \mathbf{x}') &= (\mathbf{x}^\top \mathbf{x}' + 1)^2 \\ &= (x_1 x_1' + x_2 x_2' + 1)^2 \\ &= 1 + 2x_1 x_1' + 2x_2 x_2' + 2x_1 x_2 x_1' x_2' + (x_1)^2 (x_1')^2 + (x_2)^2 (x_2')^2, \end{aligned}$$

  which corresponds to the inner product of the basis function

$$\phi(\mathbf{x}) = \left(1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1 x_2, (x_1)^2, (x_2)^2\right)^\top.$$

Some Common Kernel Functions: RBF Kernel

- Radial basis function (RBF) kernel:

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2s^2}\right).$$

- It can be generalized to

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\mathcal{D}(\mathbf{x}, \mathbf{x}')}{2s^2}\right),$$

  where $\mathcal{D}(\cdot, \cdot)$ is some distance function.

To Learn More...

- One-class support vector machines