3.1 Recap of last lecture

As in the previous lecture, we assume that \((X, Y) \sim P\), where \(X\) takes values in \(X = \mathbb{R}^d\), and \(Y\) takes values in \(Y = \{-1, 1\}\). Let \(D_n = \{(X^{(i)}, Y^{(i)})\}_{i=1}^n\) be a set of \(n\) i.i.d. samples from \(P\). Each \(\theta \in \mathbb{R}^d\) defines a function \(f_\theta(x) = \langle \theta, x \rangle\). We consider the decision rule

\[
 g(X) = \text{sgn}(f_\theta(X)).
\]

Recall that the margin is defined as

\[
\delta(\theta, D_n) = \min_{1 \leq i \leq n} \frac{Y^{(i)} \langle \theta, X^{(i)} \rangle}{\|\theta\|_2},
\]

and the radius of the data set is defined as

\[
R(D_n) = \max_{1 \leq i \leq n} \|X^{(i)}\|_2.
\]

In the previous lecture, we proved that if \(D_n\) is linearly separable, then the Perceptron Algorithm converges in \(T = R^2 / \delta^2\) steps.

3.2 Motivation for maximizing the margin

The 0-1 loss function is given by

\[
\ell(f_\theta(X), Y) = \begin{cases} 
1 & \text{if } \text{sgn}(f_\theta(X)) \neq y, \\
0 & \text{otherwise}.
\end{cases}
\]

Let \(k\) be chosen uniformly at random from the set \(\{0, 1, \ldots, n-1\}\), and define the truncated data set \(D_{n,k} = \{(X^{(i)}, Y^{(i)})\}_{i=1}^k\). Use the Perceptron Algorithm to compute a classifier \(f_{n,k}\) for the set \(D_{n,k}\). Let \(f_n\) denote the resulting classifier, which depends on the random variables \(k\) and \(D_n\).
To compute the risk of the classifier $f_n$, we take the expectation of the loss over the random variables $k, X, Y, \text{ and } D_n$

\[
E_k E_{X,Y,D_n} \ell(f_n(X), Y) = \frac{1}{n} \sum_{k=0}^{n-1} E_{X,Y,D_n} \ell(f_{n,k}(X), Y) \\
= \frac{1}{n} \sum_{k=0}^{n-1} E_{D_n} \ell(f_{n,k}(X^{(k+1)}), Y^{(k+1)}) \\
\leq \frac{1}{n} \left( \text{# of mistakes} \right) \\
\leq \frac{R^2}{n \delta^2}
\]

Here, the first equality follows from the fact that $k$ is uniformly distributed. Note that $(X^{(k+1)}, Y^{(k+1)}) \sim P$, and the random variable $f_n$ is independent of the data $\{(X^{(i)}, Y^{(i)})\}_{i=k+1}^n$, which was not used in training. Hence, we have the second equality. The last inequality follows from the result we proved last lecture, which said that the total number of mistakes made by the Perceptron Algorithm is bounded by $R^2/\delta^2$. Thus, we conclude that, in this case, maximizing the margin minimizes risk.

### 3.3 Max-margin as an optimization problem

Next, we cast the problem of maximizing the margin as an optimization problem

\[
\max_{\delta \geq 0, \theta \in \mathbb{R}^d} \delta \\
\text{s.t. } \frac{Y^{(i)} \langle \theta, X^{(i)} \rangle}{\|\theta\|_2} \geq \delta, \quad i = 1, \ldots, n
\]

Since the vector $\theta$ is normalized, we can take $\|\theta\|_2 = 1/\delta$ and rewrite our problem as

\[
\min_{\theta \in \mathbb{R}^d} \frac{1}{2} \|\theta\|_2^2 \\
\text{s.t. } Y^{(i)} \langle \theta, X^{(i)} \rangle \geq 1, \quad i = 1, \ldots, n.
\]  

(3.1)

Because the objective function is quadratic and the constraints are affine, this optimization problem is called a quadratic program (QP). We refer to (3.1) as the primal problem. A point $\theta \in \mathbb{R}^d$ is feasible for the problem (3.1) if it satisfies all the constraints. Note that a feasible point exists if and only if the data are linearly separable. In the following derivation, we assume that the data are linearly separable.
3.3.1 The dual formulation

Define the Lagrangian

\[ L(\theta, \alpha) = \frac{1}{2}\|\theta\|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - Y^{(i)} \langle \theta, X^{(i)} \rangle) , \]

where \( \alpha \geq 0 \) (component-wise). The \( \alpha_i \) are called the dual variables. If \( \hat{\theta} \) is a feasible point for problem (3.1), then \( (1 - Y^{(i)} \langle \hat{\theta}, X^{(i)} \rangle) \leq 0, \) for \( i = 1, \ldots, n \). Hence,

\[ \sup_{\alpha \geq 0} L(\hat{\theta}, \alpha) = \begin{cases} \frac{1}{2}\|\hat{\theta}\|_2^2, & \text{if } \hat{\theta} \text{ is feasible} \\ +\infty & \text{otherwise.} \end{cases} \quad (3.2) \]

Therefore, computing

\[ p^* \triangleq \inf_{\theta \in \mathbb{R}^d} \sup_{\alpha \geq 0} L(\theta, \alpha), \]

is the same as solving the primal problem (3.1).

By equation (3.2), the following relations hold for all feasible \( \hat{\theta} \in \mathbb{R}^d \) and all \( \alpha \geq 0 \):

\[ \sup_{\alpha \geq 0} L(\hat{\theta}, \alpha) = \frac{1}{2}\|\hat{\theta}\|_2^2 \geq L(\hat{\theta}, \alpha) \geq \inf_{\theta \in \mathbb{R}^d} L(\theta, \alpha). \quad (3.3) \]

Since this holds for all feasible \( \hat{\theta} \) and all \( \alpha \geq 0 \), we can take the infimum on the left and the supremum on the right to get

\[ p^* = \inf_{\theta \in \mathbb{R}^d} \sup_{\alpha \geq 0} L(\theta, \alpha) \geq \sup_{\alpha \geq 0} \inf_{\theta \in \mathbb{R}^d} L(\theta, \alpha) \overset{\Delta}{=} q^* \quad (3.4) \]

The optimization

\[ q^* = \sup_{\alpha \geq 0} \inf_{\theta \in \mathbb{R}^d} L(\theta, \alpha) \quad (3.5) \]

is known as the dual problem, and the inequality (3.4) is a result known as weak duality. Because the original problem (3.1) satisfies Slater’s condition, we actually have strong duality (i.e., \( p^* = q^* \)).

The Lagrangian \( L(\theta, \alpha) \) is a convex function, so the \( \theta \) that minimizes \( L(\theta, \alpha) \) is given by

\[ \frac{\partial L(\theta, \alpha)}{\partial \theta} = \theta + \sum_{i=1}^{n} \alpha_i (-Y^{(i)} X^{(i)}) = 0, \]

which yields \( \theta = \sum \alpha_i Y^{(i)} X^{(i)} \). Substituting this value of \( \theta \) into \( L(\theta, \alpha) \) gives

\[ q(\alpha) \overset{\Delta}{=} \inf_{\theta} L(\theta, \alpha) = -\frac{1}{2} \alpha^T (K \odot Y) \alpha + \langle \alpha, e \rangle, \]

where \( K, Y \in \mathbb{R}^{n \times n} \) are the Gram matrices given by \( K_{ij} = \langle X^{(i)}, X^{(j)} \rangle \) and \( Y_{ij} = Y^{(i)} Y^{(j)} \), and \( e \in \mathbb{R}^n \) is the vector of ones. The symbol \( \odot \) denotes the Hadamard (element-by-element) product.
3.3.2 Interpretation

Suppose that \((\theta^*, \alpha^*)\) is an optimum for the Lagrangian. From equations (3.2) and (3.6), an optimum point has the following properties:

1. \(\theta^* = \sum_{i=1}^{n} \alpha_i^* Y^{(i)} X^{(i)}\).
2. If \(\alpha_i > 0\), then \(1 - Y^{(i)} \langle \theta^*, X^{(i)} \rangle = 0\).

The second condition is known as complementary slackness. Together, these two conditions tell us that an optimum \(\theta^*\) only depends on the data \((X^{(i)}, Y^{(i)})\) where the corresponding \(\alpha_i^*\) are nonzero. The points \(X^{(i)}\) where \(\alpha_i^* > 0\) are called support vectors, and they are contained in the hyperplanes \(\{X \in \mathbb{R}^d | \langle \theta^*, X \rangle = 1\}\) and \(\{X \in \mathbb{R}^d | \langle \theta^*, X \rangle = -1\}\). This method of finding the maximum-margin classifier is called the support vector machine.

3.4 Extension

What if the data \(D_n\) are not linearly separable? Observe that for an optimum \((\theta^*, \alpha^*)\),

\[\langle \theta^*, X \rangle = \sum_{i=1}^{n} \alpha_i^* Y^{(i)} \langle X^{(i)}, X \rangle.\]

Hence, we are mainly concerned with the inner product on \(\mathcal{X}\). If we extend our notion of inner product, we can use the support vector machine ideas described above to classify data sets that are not linearly separable.

Consider the data set shown in Figure 3.1. Clearly, these data are not linearly separable. However, if we apply the map

\[(X_1, X_2) \mapsto (1, X_1, X_2, X_1^2, X_2^2),\]

and take our inner product in \(\mathbb{R}^6\), rather than \(\mathbb{R}^2\), we can use the techniques described above to determine \(\theta^* \in \mathbb{R}^6\) that correctly classifies the data (see Figure 3.2).

This intuition motivates the use of kernel methods, to which we turn in the next lectures.
Figure 3.1. A set of data that is not linearly separable.

Figure 3.2. Classifier that separates the data.