Back to linear classification

• Last time: we’ve seen that kernels can help capture non-linear patterns in data while keeping the advantages of a linear classifier

• Today: Support Vector Machines
  – A hyperplane-based classification algorithm
  – Highly influential
  – Backed by solid theoretical grounding (Vapnik & Cortes, 1995)
  – Easy to kernelize
The Maximum Margin Principle

• Find the hyperplane with maximum separation margin on the training data
Margin of a data set $D$

$$margin(D, w, b) = \begin{cases} 
\min_{(x,y) \in D} y(w \cdot x + b) & \text{if } w \text{ separates } D \\
-\infty & \text{otherwise}
\end{cases} \quad (3.8)$$

Distance between the hyperplane $(w,b)$ and the nearest point in $D$

$$margin(D) = \sup_{w,b} margin(D, w, b) \quad (3.9)$$

Largest attainable margin on $D$
Support Vector Machine (SVM)

A hyperplane based linear classifier defined by $\mathbf{w}$ and $b$

Prediction rule: $y = \text{sign}(\mathbf{w}^T \mathbf{x} + b)$

**Given:** Training data $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$

**Goal:** Learn $\mathbf{w}$ and $b$ that achieve the maximum margin
Characterizing the margin

Let’s assume the entire training data is correctly classified by \((w, b)\) that achieve the maximum margin.

- Assume the hyperplane is such that
  - \(w^T x_n + b \geq 1\) for \(y_n = +1\)
  - \(w^T x_n + b \leq -1\) for \(y_n = -1\)
  - Equivalently, \(y_n(w^T x_n + b) \geq 1\)
    \[\Rightarrow \min_{1 \leq n \leq N} |w^T x_n + b| = 1\]

- The hyperplane’s margin:
  \[\gamma = \min_{1 \leq n \leq N} \frac{|w^T x_n + b|}{||w||} = \frac{1}{||w||}\]
The Optimization Problem

We want to maximize the margin $\gamma = \frac{1}{\|w\|}$

Maximizing the margin $\gamma = \text{minimizing } \|w\|$ (the norm)

Our optimization problem would be:

Minimize $f(w, b) = \frac{\|w\|^2}{2}$

subject to $y_n(w^T x_n + b) \geq 1, \quad n = 1, \ldots, N$
Large Margin = Good Generalization

• Intuitively, large margins mean good generalization
  – Large margin $\Rightarrow$ small $||w||$
  – small $||w||$ $\Rightarrow$ regularized/simple solutions

• (Learning theory gives a more formal justification)
Solving the SVM Optimization Problem

Our optimization problem is:

\[
\begin{align*}
\text{Minimize } & f(\mathbf{w}, b) = \frac{\|\mathbf{w}\|^2}{2} \\
\text{subject to } & 1 \leq y_n(\mathbf{w}^T \mathbf{x}_n + b), \quad n = 1, \ldots, N
\end{align*}
\]

Introducing Lagrange Multipliers \( \alpha_n \) \((n = \{1, \ldots, N\})\), one for each constraint, leads to the Lagrangian:

\[
\begin{align*}
\text{Minimize } & L(\mathbf{w}, b, \alpha) = \frac{\|\mathbf{w}\|^2}{2} + \sum_{n=1}^{N} \alpha_n \{1 - y_n(\mathbf{w}^T \mathbf{x}_n + b)\} \\
\text{subject to } & \alpha_n \geq 0; \quad n = 1, \ldots, N
\end{align*}
\]
Solving the SVM Optimization Problem

Our optimization problem is:

\[
\begin{align*}
\text{Minimize} & \quad f(w, b) = \frac{\|w\|^2}{2} \\
\text{subject to} & \quad 1 \leq y_n(w^T x_n + b), \quad n = 1, \ldots, N
\end{align*}
\]

Introducing Lagrange Multipliers \( \alpha_n \ (n = \{1, \ldots, N\}) \), one for each constraint, leads to the Lagrangian:

\[
\begin{align*}
\text{Minimize} & \quad L(w, b, \alpha) = \frac{\|w\|^2}{2} + \sum_{n=1}^{N} \alpha_n \{1 - y_n(w^T x_n + b)\} \\
\text{subject to} & \quad \alpha_n \geq 0; \quad n = 1, \ldots, N
\end{align*}
\]
Solving the SVM Optimization Problem

Take (partial) derivatives of $L_P$ w.r.t. $\mathbf{w}$, $b$ and set them to zero

$$\frac{\partial L_P}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n, \quad \frac{\partial L_P}{\partial b} = 0 \Rightarrow \sum_{n=1}^{N} \alpha_n y_n = 0$$

Substituting these in the Primal Lagrangian $L_P$ gives the Dual Lagrangian

Maximize $L_D(\mathbf{w}, b, \alpha) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{m,n=1}^{N} \alpha_m \alpha_n y_m y_n (\mathbf{x}_m^T \mathbf{x}_n)$

subject to $\sum_{n=1}^{N} \alpha_n y_n = 0$, $\alpha_n \geq 0$; $n = 1, \ldots, N$
Solving the SVM Optimization Problem

Take (partial) derivatives of $L_P$ w.r.t. $w$, $b$ and set them to zero

$$
\frac{\partial L_P}{\partial b} = 0 \Rightarrow \sum_{n=1}^{N} \alpha_n y_n = 0
$$

A Quadratic Program for which many off-the-shelf solvers exist

Substituting these into the Primal Lagrangian $L_P$ gives the Dual Lagrangian

Maximize

$$
L_D(w, b, \alpha) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{m,n=1}^{N} \alpha_m \alpha_n y_m y_n (x_m^T x_n)
$$

subject to

$$
\sum_{n=1}^{N} \alpha_n y_n = 0, \quad \alpha_n \geq 0; \quad n = 1, \ldots, N
$$
SVM: the solution!

Once we have the $\alpha_n$’s, $\mathbf{w}$ and $b$ can be computed as:

$$
\mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n
$$

$$
b = -\frac{1}{2} \left( \min_{n:y_n=+1} \mathbf{w}^T \mathbf{x}_n + \max_{n:y_n=-1} \mathbf{w}^T \mathbf{x}_n \right)
$$

**Note:** Most $\alpha_n$’s in the solution are zero (sparse solution)

- Reason: Karush-Kuhn-Tucker (KKT) conditions
- For the optimal $\alpha_n$’s
  $$
  \alpha_n \{1 - y_n (\mathbf{w}^T \mathbf{x}_n + b)\} = 0
  $$

- $\alpha_n$ is non-zero only if $\mathbf{x}_n$ lies on one of the two margin boundaries, i.e., for which $y_n (\mathbf{w}^T \mathbf{x}_n + b) = 1$
- These examples are called support vectors
- Support vectors “support” the margin boundaries
What if the data is not separable?

Non-separable case: We will allow misclassified training examples
  but we want their number to be minimized
⇒ by minimizing the sum of slack variables ($\sum_{n=1}^{N} \xi_n$)

The optimization problem for the non-separable case

Minimize $f(w, b) = \frac{||w||^2}{2} + C \sum_{n=1}^{N} \xi_n$

subject to $y_n(w^T x_n + b) \geq 1 - \xi_n, \quad \xi_n \geq 0 \quad n = 1, \ldots, N$
Support Vector Machines

• Find the max margin linear classifier for a dataset

• Discovers “support vectors”, the training examples that “support” the margin boundaries

• Allows misclassified training examples

• Today: we’ve seen how to learn an SVM if the data is separable

• Next time: we’ll solve the more general case