Optimal Hyperplanes and Support Vector Machines

CS4780/5780 – Machine Learning
Fall 2019

Nika Haghtalab & Thorsten Joachims
Cornell University

Reading: UML 15.1, 15.2
Example: Reuters Text Classification

![Graph showing percent training/testing errors over iterations with different datasets. The graph indicates a decrease in errors as iterations increase, with a label pointing to an "optimal hyperplane."
Optimal Hyperplanes

- Assumption:
  - Training examples are linearly separable.
Margin of a Linear Classifier

• Definition: For a linear classifier $h_w$, the margin $\gamma$ of an example $(x, y)$ with $x \in \mathbb{R}^N$ and $y \in \{-1, +1\}$ is

\[
\gamma = y(w \cdot x + b)
\]

• Definition: The margin is called geometric margin, if $||w|| = 1$. For general $w$, the term functional margin is used to indicate that the norm of $w$ is not necessarily 1.

• Definition: The (hard) margin of a homogeneous linear classifier $h_w$ on sample $S$ is

\[
\gamma = \min_{(x,y) \in S} y(w \cdot x + b)
\]
Hard-Margin Separation

• Goal:
  – Find hyperplane with the largest distance to the closest training examples.

  Optimization Problem (Primal):

  \[
  \begin{align*}
  \min_{\vec{w}, b} & \quad \frac{1}{2} \vec{w} \cdot \vec{w} \\
  \text{s.t.} & \quad y_1(\vec{w} \cdot \vec{x}_1 + b) \geq 1 \\
                  & \quad \ldots \\
                  & \quad y_n(\vec{w} \cdot \vec{x}_n + b) \geq 1
  \end{align*}
  \]

• Support Vectors:
  – Examples with minimal distance (i.e. margin).
Non-Separable Training Data

• Limitations of hard-margin formulation
  – For some training data, there is no separating hyperplane.
  – Complete separation (i.e. zero training error) can lead to suboptimal prediction error.
Soft-Margin Separation

Idea: Maximize margin and minimize training error.

|-------------------------|--------------------------|
| \[
\min_{\vec{w}, b} \frac{1}{2} \vec{w} \cdot \vec{w}
\] | \[
\min_{\vec{w}, \xi, b} \frac{1}{2} \vec{w} \cdot \vec{w} + C \sum_{i=1}^{n} \xi_i
\] |
| s.t. \[
y_1(\vec{w} \cdot \vec{x}_1 + b) \geq 1
\] | s.t. \[
y_1(\vec{w} \cdot \vec{x}_1 + b) \geq 1 - \xi_1 \wedge \xi_1 \geq 0
\] |
| ... | ... |
| \[
y_n(\vec{w} \cdot \vec{x}_n + b) \geq 1
\] | \[
y_n(\vec{w} \cdot \vec{x}_n + b) \geq 1 - \xi_n \wedge \xi_n \geq 0
\] |

- Slack variable \( \xi_i \) measures by how much \((x_i, y_i)\) fails to achieve margin \( \gamma \)
- \( \Sigma \xi_i \) is upper bound on number of training errors
- \( C \) is a parameter that controls trade-off between margin and training error.
Controlling Soft-Margin Separation

- $\sum \xi_i$ is an upper bound on the number of training errors.
- $C$ is a parameter that controls the trade-off between margin and training error.

**Soft-Margin OP (Primal):**

$$\min_{\vec{w}, \xi, b} \frac{1}{2} \vec{w} \cdot \vec{w} + C \sum_{i=1}^{n} \xi_i$$

subject to:

- $y_1(\vec{w} \cdot \vec{x}_1 + b) \geq 1 - \xi_1 \land \xi_1 \geq 0$
- ...
- $y_n(\vec{w} \cdot \vec{x}_n + b) \geq 1 - \xi_n \land \xi_n \geq 0$

---

**Diagram:**

- Large $C$: Supports lie closer to the decision boundary.
- Small $C$: Supports lie further away from the decision boundary.
Example Reuters “acq”: Varying C

![Graph showing percent training/testing errors vs. C for a hard-margin SVM. The graph includes two data sets: "svm_trainerror.dat" and "svm_testerror.dat". The red line represents the training error, and the green dashed line represents the testing error.]