Support Vector Machines and Optimal Hyperplanes

CS6780 – Advanced Machine Learning
Spring 2019

Thorsten Joachims
Cornell University

Reading: Murphy 14.5
Schoelkopf/Smola Chapter 5 (rest), Chapter 7.1-7.3, 7.5
Example: Reuters Text Classification

"optimal hyperplane"
Optimal Hyperplanes

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

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

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

Definition: The (hard) margin of an unbiased linear classifier $h_{\tilde{w}}$ on a sample $S$ is $\delta = \min_{(\tilde{x}, y) \in S} y(\tilde{w} \cdot \tilde{x})$.

Definition: The (hard) margin of an unbiased linear classifier $h_{\tilde{w}}$ on a task $P(X, Y)$ is

$$\delta = \inf_{S \sim P(X, Y)} \min_{(\tilde{x}, y) \in S} y(\tilde{w} \cdot \tilde{x}).$$
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).
Vapnik Chervonenkis Dimension

• Definition: The VC-Dimension of $H$ is equal to the maximum number $d$ of examples that can be split into two sets in all $2^d$ ways using functions from $H$ (shattering).
Generalization Error Bound: Infinite H, Non-Zero Error

• Setting
  – Sample of \( n \) labeled instances \( S \)
  – Learning Algorithm \( L \) using a hypothesis space \( H \) with \( VCDim(H)=d \)
  – ERM learner \( L \) returns hypothesis \( \hat{h}=L(S) \) with lowest training error

• Given hypothesis space \( H \) with \( VCDim(H) \) equal to \( d \) and an i.i.d. sample \( S \) of size \( n \), with probability \( (1-\delta) \) it holds that

\[
Err_P(h_L(S)) \leq Err_S(h_L(S)) + \sqrt{\frac{d \left( \ln \left( \frac{2n}{d} \right) + 1 \right) - \ln \left( \frac{\delta}{4} \right)}{n}}
\]
VC Dimension of Hyperplanes

• Theorem: The VC Dimension of unbiased hyperplanes over N features is N.
• Theorem: The VC Dimension of biased hyperplanes over N features is N+1.
VC Dimension of Margin Hyperplanes

Theorem: Unbiased linear classifiers $H_X$ with $\|w\| = 1/\delta$ and $\max_i \|x_i\| \leq R$ and margin

$$\min_i |w \cdot x_i| = 1$$

for a given set of instances $X = \{x_1, \ldots, x_k\}$, have VC Dimension

$$VCDim(H_X) \leq \frac{R^2}{\delta^2}$$