**EECS 598-005: Theoretical Foundations of Machine Learning** **Fall 2015**

# Lecture 11: Margin Theory

*Lecturer: Jacob Abernethy* *Scribe: Daniel LeJeune, Editors: Pengyu Xiao*

## 11.1 Review: Lower Bounds

**Theorem 11.1.** *Given any class $\mathcal{C}$ with VC-dimension $d$, there exists a distribution $D$ such that, for a sample $S$ of size $m = \frac{d-1}{16\epsilon}$,*

$$\Pr(R(h_S) \leq \epsilon) \leq e^{-(d-1)/48}.$$

For example, if $d \geq 49$, the probability is less than or equal to $1/e$, which is less than $1/2$. In the case of the distribution used in the proof of this theorem, we have to see more than half of the rare samples to ensure the error rate. What we can take away from this theorem is that it is hard to learn a hypothesis on a set you can shatter.

**Conclusions**

1) We can guarantee that, for some constant $c_1$, if $m = c_1 \frac{d \log 1/\epsilon + \log 1/\delta}{\epsilon}$, then $R(h_S) \leq \epsilon$ with probability at least $1 - \delta$.

2) We can guarantee that, for some constant $c_2$, if $m = c_2 \frac{d}{\epsilon}$, then there exists a distribution $D$ such that $\Pr(R(h_S) \leq 1/2$.

These statements are about the distribution-agnostic VC-dimension, but real world distributions often make the problem nicer. Consider an alternate statement of conclusion 1: with $m$ examples, the error $\epsilon \approx \frac{d}{m}$, which is very bad for the case when $d > m$. However, large margins (see Figure 11.1) come to our rescue and drive the VC-dimension down.

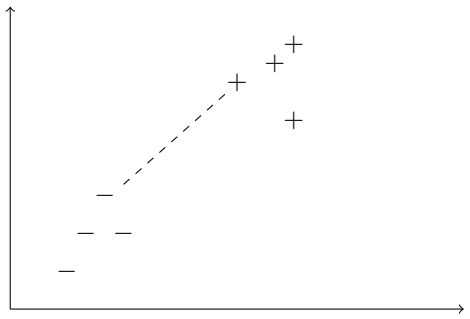

Figure 11.1: Separation margin (dashed) between classes.

## 11.2 Margin Theory

Let's define a hypothesis class. *Note: We present margin theory with linear threshold functions today, but it can be generalized. Also, the need for a bias term in the classifier can be omitted by appending a 1 to each* $\mathbf{x}$.

$$\mathcal{H} \triangleq \{h_{\mathbf{w}}(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x}) : \mathbf{w} \in \mathbb{R}^n\}$$

**Definition 11.2** (linearly separable)**.** *Examples* $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m) \in \mathbb{R}^n \times \{-1, 1\}$ *are* ***linearly separable*** *if there exists an* $h_{\mathbf{w}} \in \mathcal{H}$ *such that* $h_{\mathbf{w}}(\mathbf{x}_i) = y_i$ *for all* $i = 1, \ldots, m$*.*

Notice that this condition is equivalent to $y_i(\mathbf{w} \cdot \mathbf{x}_i) > 0$, which is, due to the degree of freedom in the magnitude of $\mathbf{w}$, equivalent to $y_i(\mathbf{w} \cdot \mathbf{x}_i) \geq 1$ if we rescale $\mathbf{w}$ so that $\min_i y_i(\mathbf{w} \cdot \mathbf{x}_i) = 1$.

**Definition 11.3** (margin)**.** *Given examples* $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m) \in \mathbb{R}^n \times \{-1, 1\}$*, the* ***margin*** *of* $\mathbf{w}$ *is*

$$\min_i \frac{y_i(\mathbf{w} \cdot \mathbf{x}_i)}{\|\mathbf{w}\|_2}$$

Using the rescaled version of $\mathbf{w}$, this is simply $1/\|\mathbf{w}\|_2$.

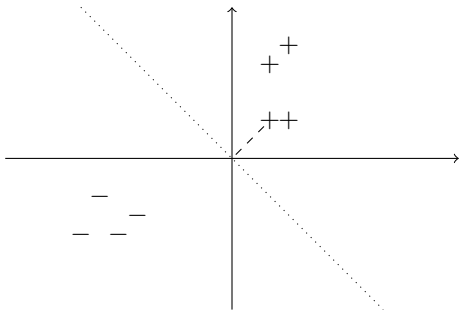

Figure 11.2: Margin (dashed) of a linear classifier (dotted).

## 11.3 Support Vector Machine Algorithm (separable case)

The rule the support vector machine takes is, "always make the margin as large as possible." Specifically, find the $\mathbf{w}$ that solves

$$\max_{\mathbf{w} \in \mathbb{R}^n} \frac{1}{\|\mathbf{w}\|_2} \text{ s.t. } y_i(\mathbf{w} \cdot \mathbf{x}_i) \geq 1 \; \forall i = 1, \ldots, m$$

To actually solve this, we can minimize $\frac{1}{2}\|\mathbf{w}\|_2^2$, which is convex. But why is a large margin what we want?

## 11.4 Why a Large Margin?

Let $S$ be a sample from $\mathcal{X}$, such that $\|\mathbf{x}_i\|_2 \leq r$ for all $\mathbf{x}_i$ in $S$. Then let us define a hypothesis class

$$\mathcal{H}_{S,\Lambda} \triangleq \{h_{\mathbf{w}}(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x}) : \min_{\mathbf{x}_i \in S} |\mathbf{w} \cdot \mathbf{x}_i| = 1 \; \wedge \|\mathbf{w}\|_2 \leq \Lambda\}.$$

**Theorem 11.4.** *The VC-dimension of* $\mathcal{H}_{S,\Lambda}$ *is less than or equal to* $r^2\Lambda^2$*.*

**Proof:** Let $d$ be the VC-dimension of $\mathcal{H}_{S,\Lambda}$. Then there exists $\{\mathbf{x}_1, \ldots, \mathbf{x}_d\} \subseteq S$ that are shattered by $\mathcal{H}_{S,\Lambda}$. Now, we know that for any labeling $(y_1, \ldots, y_d)$ there exists $h_{\mathbf{w}} \in \mathcal{H}_{S,\Lambda}$ such that $y_i(\mathbf{w} \cdot \mathbf{x}_i) \geq 1$ for all

$i = 1, \ldots, d$. Summing across $i$, we have

$$\begin{aligned} d &\leq \mathbf{w} \cdot \left( \sum_{i=1}^{d} y_i \mathbf{x}_i \right) \\ &\leq \|\mathbf{w}\|_2 \left\| \sum_{i=1}^{d} y_i \mathbf{x}_i \right\|_2 \quad \text{(Cauchy-Schwarz inequality)} \\ &\leq \Lambda \left\| \sum_{i=1}^{d} y_i \mathbf{x}_i \right\|_2 \quad \text{(definition of } \mathcal{H}_{S,\Lambda}\text{)} \end{aligned}$$

Now, let $y_1, \ldots, y_d$ be independent and random with equal probability of 1 or $-1$, so that $\mathbb{E}[y_i] = 0$. Now, if we take the expectation of both sides of the bound, the inequality will still hold.

$$\begin{aligned} d &\leq \Lambda \mathbb{E} \left( \left\| \sum_{i=1}^{d} y_i \mathbf{x}_i \right\|_2^2 \right)^{1/2} \\ &\leq \Lambda \left( \mathbb{E} \left\| \sum_{i=1}^{d} y_i \mathbf{x}_i \right\|_2^2 \right)^{1/2} \quad \text{(Jensen's inequality)} \\ &= \Lambda \left( \mathbb{E} \sum_{i,j=1}^{d} y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \right)^{1/2} \\ &= \Lambda \left( \sum_{i,j=1}^{d} \mathbb{E}[y_i y_j] (\mathbf{x}_i \cdot \mathbf{x}_j) \right)^{1/2} \\ &= \Lambda \left( \sum_{i}^{d} \mathbf{x}_i \cdot \mathbf{x}_i \right)^{1/2} \quad (\mathbb{E}[y_i y_i] = 1, \mathbb{E}[y_i y_{j \neq i}] = 0) \\ &\leq \Lambda r d^{1/2} \quad (\|\mathbf{x}_i\|_2 \leq r) \end{aligned}$$

So $d \leq r^2 \Lambda^2$. ■

So, smaller $\Lambda$ forces us to use classifiers with larger margins, but in return the VC-dimension is decreased significantly.

## 11.5 Noisy Probabilistic Setting

Typically, the "noise-free" property doesn't hold. That is, there does not exist a function $f$ such that $\Pr(f(\mathbf{x}) = y) = 0$.

**Definition 11.5** (Bayes risk)**.** *Given a distribution $D$ on $\mathcal{X} \times \mathcal{Y}$, the **Bayes risk** $R^*$ is defined as*

$$R^* \triangleq \inf_f R(f).$$

Now, we want to bound $R(h_S) - R^*$. A little algebra lets us rewrite this as

$$\underbrace{R(h_S) - \inf_{h \in \mathcal{H}} R(h)}_{\text{estimation error}} + \underbrace{\inf_{h \in \mathcal{H}} R(h) - R^*}_{\text{approximation error}}.$$

The estimation error measures how poorly we learn the correct classifier of our hypothesis class, and the approximation error measures how poorly our hypothesis class fits the data. Simpler functions (e.g., linear classifiers) often have low estimation error and high approximation error, and more complex functions (e.g., neural networks) often have high estimation error and low approximation error.

In future lectures, we will see how we can bound the estimation error using

$$\sup_{h \in \mathcal{H}} |\hat{R}_S(h) - R(h)|.$$