Computational Learning Theory
Lecture 10: Margin Theory

Lecturer: James Worrell

1 Introduction

Coupled with the kernel trick, SVMs allow us to learn linear classifiers in high and even infinite dimensional spaces. The error bounds based on VC dimension that we proved in Lectures 3 and 4, which are independent of the distribution over examples, are too general to provide guarantees in this case. Indeed the class of linear classifiers over an infinite-dimensional vector space has infinite VC dimension. In this lecture we will prove margin-based error bounds for linear classifers that do not explicitly depend on the dimension of the underlying space. These allow us to give good generalisation error bounds in those favourable situations in which we able to derive a large-margin classifier. For simplicity, except in Section 2, we work in the realisable setting: we assume that the distribution on labelled examples admits a linear classifier with zero error and we analyse the output of the hard-margin SVM algorithm.[1]

2 Leave-One-Out Analysis

We first give a relatively straightforward learning guarantee for SVMs based on the notion of leave-one-out error.

Suppose that a learning algorithm $\mathcal{A}$ returns hypothesis h_S given training set S. The *leave-one-out error* of $\mathcal{A}$ on a training set S of size m is defined to be

$$\mathrm{LOO}(S) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}\{h_{S\setminus\{(x_i,y_i)\}}(x_i) \neq y_i\}\,.$$

Proposition 1. *For all $m \geq 1$ and any distribution D on labelled examples we have*

$$\mathop{\mathrm{E}}_{S\sim D^{m+1}}[\mathrm{LOO}(S)] = \mathop{\mathrm{E}}_{S\sim D^{m}}[\mathrm{err}(h_S)]\,.$$

Proof. We have

$$\begin{aligned}
\mathop{\mathrm{E}}_{S\sim D^{m+1}}[\mathrm{LOO}(S)] &= \frac{1}{m+1}\sum_{i=1}^{m+1} \mathop{\mathrm{E}}_{S\sim D^{m+1}}\left[\mathbb{I}\{h_{S\setminus\{(x_i,y_i)\}}(x_i) \neq y_i\}\right] \\
&= \mathop{\mathrm{E}}_{S\sim D^{m+1}}\left[\mathbb{I}\{h_{S\setminus\{(x_1,y_1)\}}(x_1) \neq y_1\}\right] \\
&= \mathop{\mathrm{E}}_{\substack{(x,y)\sim D\\ S\sim D^m}}[\mathbb{I}\{h_S(x) \neq y\}] \\
&= \mathop{\mathrm{E}}_{S\sim D^m}[\mathrm{err}(h_S)]\,.
\end{aligned}$$

□

[1] Another simplification is that we work with linear classifiers $\vec{x} \mapsto \vec{w}\cdot\vec{x}$ with no constant term. This is no loss of generality since if we add an extra dimension to the input space and make each example have coordinate 1 in the extra dimension then an arbitrary linear classifier can equivalently be written as one with no constant term.

Theorem 1. *Let h_S denote the hypothesis returned by the (soft-margin) SVM algorithm on sample S. Let $N(S)$ be the number of support vectors defining h_S. Then*

$$\mathop{\mathrm{E}}_{S\sim D^m}[\mathrm{err}(h_S)] \leq \mathop{\mathrm{E}}_{S\sim D^{m+1}}\left[\frac{N(S)}{m+1}\right].$$

Proof. Suppose that $x \in S$ is not a support vector. Then $h_{S-\{x\}} = h_S$ (see Exercise Sheet 4) and thus h_S correctly classifies x. Thus the number of leave-one-out errors is at most $N(S)$, i.e., $\mathrm{LOO}(S) \leq \frac{N(S)}{m+1}$. It follows from Proposition 1 that

$$\mathop{\mathrm{E}}_{S\sim D^m}[\mathrm{err}(h_S)] = \mathop{\mathrm{E}}_{S\sim D^{m+1}}[\mathrm{LOO}(S)] \leq \mathop{\mathrm{E}}_{S\sim D^{m+1}}\left[\frac{N(S)}{m+1}\right]$$

□

Theorem 1 tells us that a distribution that leads to classifiers with relatively few support vectors on average will also lead to classifiers that generalise well on average. However this result has limited usefulness. For example, it does not entail any correlation between the number of support vectors and the generalisation error of a particular classifier.

3 Rademacher Complexity and Ramp Loss

Given $\vec{w} \in \mathbb{R}^n$, let $f_{\vec{w}} : \mathbb{R}^n \to \mathbb{R}$ denote the linear map $f_{\vec{w}}(\vec{x}) = \vec{w} \cdot \vec{x}$. Fix $\Lambda > 0$ and consider the family of functions $F = \{f_{\vec{w}} : \|\vec{w}\| \leq \Lambda\}$.

Proposition 2. *Given $S = \{\vec{x}_1, \ldots, \vec{x}_m\} \subseteq \mathbb{R}^n$ and $r > 0$ such that $\|\vec{x}_i\| \leq r$, $i = 1, \ldots, m$, we have $R_S(F) \leq \frac{r\Lambda}{\sqrt{m}}$.*

Proof. We have

$$\begin{aligned}
R_S(F) &= \frac{1}{m}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\vec{w}\|\leq\Lambda}\sum_{i=1}^m \sigma_i \vec{w}\cdot\vec{x}_i\right] \\
&= \frac{1}{m}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\vec{w}\|\leq\Lambda}\vec{w}\cdot\sum_{i=1}^m \sigma_i \vec{x}_i\right] \\
&\leq \frac{\Lambda}{m}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left[\Big\|\sum_{i=1}^m \sigma_i \vec{x}_i\Big\|\right] \qquad \text{(Cauchy-Schwarz inequality)} \\
&\leq \frac{\Lambda}{m}\left(\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left[\Big\|\sum_{i=1}^m \sigma_i \vec{x}_i\Big\|^2\right]\right)^{\frac{1}{2}} \qquad \text{(Jensen's inequality)} \\
&= \frac{\Lambda}{m}\left(\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left[\sum_{i,j=1}^m \sigma_i\sigma_j \vec{x}_i\cdot\vec{x}_j\right]\right)^{\frac{1}{2}} \\
&\leq \frac{\Lambda}{m}\Big[\sum_{i=1}^m \|\vec{x}_i\|^2\Big]^{\frac{1}{2}} \\
&\leq \frac{r\Lambda}{\sqrt{m}}.
\end{aligned}$$

□

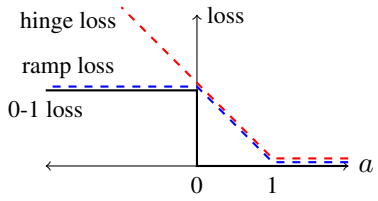

Figure 1: loss functions

Now consider the so-called *ramp loss* function $\Phi : \mathbb{R} \to [0,1]$, given by

$$\Phi(a) = \begin{cases} 0 & \text{if } x \geq 1 \\ 1-x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{otherwise.} \end{cases}$$

Notice that the ramp loss is an upper bound of the 0-1 loss and is it itself bounded by the hinge loss, see Figure 1. In proving our generalisation bounds we will exploit the fact that the ramp loss takes values in the interval $[0,1]$, which allows us to apply Theorem 2 of Lecture 5.[2]

With each map $f \in F$ we associate a loss function $g : \mathbb{R}^n \times \{-1,+1\} \to [0,1]$ defined by $g(\vec{x}, y) \to \Phi(yf(\vec{x}))$. Let G be collection of such loss functions as f ranges over F. We can use the following result (proof omitted) to bound the Rademacher complexity of G:

Lemma 1 (Talagrand's Lemma)**.** *Let $\Phi : \mathbb{R} \to \mathbb{R}$ be k-Lipshitz. Then for any set H of real-valued functions the following inequality holds*

$$R_m(\{\Phi \circ h : h \in H\}) \leq kR_m(H)\,.$$

Now consider a distribution D on $\{(\vec{x}, y) \in \mathbb{R}^n \times \{-1,+1\} : \|\vec{x}\| \leq r\}$.

Proposition 3. *With respect to the distribution D we have $R_m(G) \leq \frac{r\Lambda}{\sqrt{m}}$*

Proof. Let $S = \{(\vec{x}_1, y_1), \ldots, (\vec{x}_m, y_m\}$ be a sample chosen from D and write $S' = \{\vec{x}_1, \ldots, \vec{x}_m\}$ for the corresponding set of unlabelled points. Then

$$\begin{aligned} R_S(G) &= R_S(\{(\vec{x}, y) \mapsto \Phi(yf(x)) : f \in F\}) \\ &\leq R_S(\{(\vec{x}, y) \mapsto yf(x) : f \in F\}) \qquad \text{(Talagrand's Lemma)} \\ &= R_{S'}(F) \qquad \text{(direct calculation)} \end{aligned}$$

□

4 Generalisation Bounds for SVMs

Let the distribution D and family of loss functions G be as in the previous section. Assume that there is some linear classifier h such that $\mathrm{err}(h) = 0$ with respect to the distribution D, i.e., we're in the realisable setting.

For any $\delta > 0$, if we choose a sample $S = \{z_1, \ldots, z_m\}$ according to distribution D then with probability at least $1 - \delta$, for all $g \in G$ we have

$$\begin{aligned} E_{z\sim D}[g(z)] &\leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2R_m(G) + \sqrt{\frac{\log(1/\delta)}{m}} \qquad \text{(Theorem 2 in Lecture 5)} \\ &\leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + \frac{2r\Lambda}{\sqrt{m}} + \sqrt{\frac{\log(1/\delta)}{m}} \qquad \text{(Proposition 3)} \end{aligned}$$

[2]However for training it is computationally easier to optimise hinge loss, which is a convex function.

We would like to apply the above inequality to obtain error bounds for the output of the SVM algorithm. A natural idea is to set Λ to be the length of the vector $\vec{w}^*$ that is output by the algorithm. The problem with this is that in order for the probabilistic guarantee to make sense Λ must be fixed before the sample is drawn. The way round the problem is to simultaneously establish error bounds for an infinite family of Λ's and apply the appropriate bound to the obtained vector $\vec{w}^*$.

Theorem 2. *Let $\vec{w}^*$ be the output of the hard-margin SVM algorithm on a labelled sample of size m drawn from distribution D. Let $h^*(\vec{x}) = \mathrm{sign}(\vec{w}^* \cdot \vec{x})$ be the associated linear classifier. Then with probability at least $1 - \delta$ we have*

$$\mathrm{err}(h^*) \leq \frac{4r\|\vec{w}^*\|}{\sqrt{m}} + \sqrt{\frac{\log(4/\delta)\log_2 \|\vec{w}^*\|}{m}}\,.$$

Proof. For each integer $k \geq 1$, let $\Lambda_k = 2^k$ and define $F_k = \{f_{\vec{w}} : \|\vec{w}\| \leq \Lambda_k\}$ with $G_k = \{(\vec{x}, y) \mapsto \Phi(yf(\vec{x}) : f \in F_k\}$ the corresponding family of ramp-loss functions. Writing $\delta_k = \frac{\delta}{2k^2}$ we have that with probability at least $1 - \delta_k$,

$$\forall g \in G_k,\ \mathop{\mathrm{E}}_{z\sim D}[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + \frac{2r\Lambda_k}{\sqrt{m}} + \sqrt{\frac{\log(1/\delta_k)}{m}}\,. \tag{1}$$

Applying the union bound, using the fact that $\sum_{k=1}^{\infty} \delta_k \leq \delta$, we obtain that with probability at least $1 - \delta$ the inequality (1) holds for all $k \geq 1$.

Now suppose $\vec{w}^*$ is the output of the SVM algorithm and let $g^* \in G$ be the corresponding loss function. Writing $k = \lceil \log_2 \|\vec{w}^*\| \rceil$, we have $g^* \in G_k$ and $\frac{2}{\delta_k} \leq \frac{(2k)^2}{\delta} \leq \frac{4\log_2(\|\vec{w}\|)^2}{\delta}$. Thus

$$\begin{aligned} E_{z\sim D}[g^*(z)] &\leq \frac{1}{m}\sum_{i=1}^{m} g^*(z_i) + + \frac{2r\Lambda_k}{\sqrt{m}} + \sqrt{\frac{\log(1/\delta_k)}{m}} \\ &\leq \frac{1}{m}\sum_{i=1}^{m} g^*(z_i) + \frac{4r\|\vec{w}^*\|}{\sqrt{m}} + \sqrt{\frac{\log(4/\delta)\log_2 \|\vec{w}^*\|}{m}} \end{aligned}$$

To obtain the statement of the theorem it remains to observe that $\mathrm{err}(h^*) \leq E_{z\sim D}[g^*(z)]$ (the ramp loss is an upper bound for the 0-1 loss) and $\sum_{i=1}^{m} g^*(z_i) = 0$ (the constraints of the hard-margin SVM algorithm ensure that $\vec{w}^*$ has empirical ramp loss 0). □

Informally, Theorem 2 says that with high probability a sample that leads to a large margin (and hence small $\|\vec{w}^*\|$) has small generalisation error.