Computational Learning Theory
Lecture 5: Rademacher Complexity

Lecturer: James Worrell

In the previous lecture we introduced the Growth Function and VC dimension as measures of the complexity of (possibly infinite) hypothesis sets. We will shortly give error bounds for learning based on these measures. These results will be established as a consequence of error bounds that involve yet another complexity measure—Rademacher complexity. The latter notion will come into its own in our analysis of SVMs and Boosting. Significantly, unlike the Growth Function and VC dimension, Rademacher complexity depends on the distribution over examples as well as the expressiveness of the class of hypotheses.

1 Loss Functions

A natural approach to selecting a classifier is to choose one with the smallest empirical error on a given sample. We can reformulate and generalise this idea in terms of *loss functions*.

Given a hypothesis $h : X \to Y$, we define an associated *0-1 loss function* $g : X \times Y \to \{0, 1\}$ by

$$g(x, y) = \begin{cases} 1 & \text{if } h(x) \neq y \\ 0 & \text{otherwise.} \end{cases}$$

In other words, any labelled example for which the hypothesis does not predict the correct label incurs a loss of one, while a correct prediction incurs no loss. In these terms the empirical error of a hypothesis is the average of the associated loss function over the sample, while the true error of the hypothesis is the expected value of the loss function with respect to the distribution over labelled examples.

It will be useful to consider more general loss functions than the family of 0-1 loss functions described above. For example, consider a linear classifier $h(\vec{x}) = \text{sign}(f(\vec{x}))$ for some linear function $f(\vec{x}) = \vec{w} \cdot \vec{x} + b$, where $\text{sign}(u) = +1$ if $u \geq 0$ and -1 otherwise. The *hinge loss* function associated to such a classifier is defined by

$$g(\vec{x}, y) = \max(0, 1 - yf(\vec{x})) \,.$$

Here the loss depends on the magnitude of $f(\vec{x})$ as well as its sign, that is, we distinguish between a near miss and a bad error. We also charge a loss for correctly classified examples with insufficient margin, i.e., when $0 < yf(\vec{x}) < 1$; see Figure 1.

There are several advantages to considering the hinge loss function instead of the 0-1-loss function. In particular, finding a linear classifier with minimal 0-1 loss is **NP**-hard, whereas minimising the hinge loss can be done efficiently using linear programming.

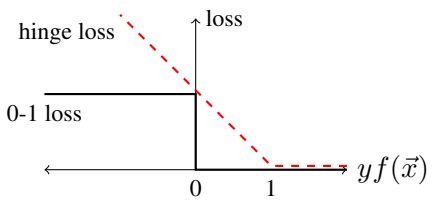

Figure 1: loss functions

2 McDiarmid's Inequality

The results of this lecture will use McDiarmid's inequality, which is a concentration bound for independent random variables.

Theorem 1 (McDiarmid's Inequality)**.** *Let V be a set and $f : V^m \to \mathbb{R}$ a function such that for some $c > 0$ and all $x_1, \ldots, x_m, x_i' \in V$,*

$$|f(x_1, \ldots, x_i, \ldots, x_m) - f(x_1, \ldots, x_i', \ldots, x_m)| \leq c\,.$$

Let $X_1, \ldots, X_m$ be independent random variables taking values in V. Then for all $\varepsilon > 0$,

$$Pr\left(f(X_1, \ldots, X_m) \geq E[f(X_1, \ldots, X_m)] + \varepsilon\right) \quad \leq \quad e^{-2\varepsilon^2/mc^2}\,.$$

We can recover Hoeffding's inequality from McDiarmid's Inequality by taking f to be the averaging function: $f(x_1, \ldots, x_m) = \frac{1}{m}\sum_{i=1}^m x_i$, with $c = 1/m$. More details about McDiarmid's Inequality can be found in [1].

3 Rademacher Complexity

Let G be a family of functions mapping a set Z into $\mathbb{R}$. Given a probability distribution D over Z, write $L(g)$ for the expected loss $\mathrm{E}_{z\sim D}[g(z)]$ of $g \in G$. Similarly, given a list $S = (z_1, \ldots, z_m)$ of elements from Z, write $L_S(g)$ for the average loss $\frac{1}{m}\sum_{i=1}^m g(z_i)$ of $g \in G$ over S. In this section we prove a result of the following form: *if a sufficiently large sample is drawn from distribution D, then with high probability $L(g)$ and $L_S(g)$ are not too far apart for all functions $g \in G$.*

In order to formulate such a result, we introduce a complexity measure for the class of functions G. To this end, let $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_m)$ be a list of independent random variables, where, for each $i \in \{1, \ldots, m\}$, σ_i takes value $+1$ with probability $1/2$ and takes value -1 with probability $1/2$. Then the *empirical Rademacher complexity* of G with respect to S is defined to be

$$R_S(G) = \mathrm{E}_{\boldsymbol{\sigma}}\Big[\sup_{g\in G}\frac{1}{m}\sum_{i=1}^m \sigma_i g(z_i)\Big]\,. \tag{1}$$

For any integer $m \geq 1$ the *Rademacher complexity* of G with respect to samples of size m drawn according to D is

$$R_m(G) = \mathop{\mathrm{E}}_{S\sim D^m}\left[R_S(G)\right]\,.$$

Intuitively, the empirical Rademacher complexity $R_S(G)$ measures how well the class of functions G correlates with randomly generated labels on the set S. The richer the class of functions G the better the chance of finding $g \in G$ that correlates with a given $\boldsymbol{\sigma}$, and hence the larger $R_S(G)$.

The following result is a bound on the difference between the true loss and the empirical loss, uniform over $g \in G$. Comparing with the error bound in Section 2.1 in Lecture 3, we see a similar dependence on the sample size m and the confidence δ. The essential difference is that the term $\sqrt{\frac{\log|H|}{m}}$ is replaced by the Rademacher complexity $R_m(G)$ below.

Theorem 2. *Let G be a family of functions mapping a set Z to the unit interval $[0, 1]$. Suppose that a sample S of size m is drawn according to distribution D on Z. Then for any $\delta > 0$, with probability at least $1 - \delta$ the following holds for all functions $g \in G$:*

$$L(g) \quad \leq \quad L_S(g) + 2R_m(G) + O\left(\sqrt{\frac{\log\frac{1}{\delta}}{m}}\right)\,. \tag{2}$$

Proof. We are interested in obtaining an upper bound for $L(g) - L_S(g)$, valid for all $g \in G$, that holds with high probability with respect to the sample S. To this end, we consider the random variable

$$\Phi(S) = \sup_{g \in G} \left(L(g) - L_S(g) \right) .$$

Our strategy is to give an upper bound of the expected value $\mathrm{E}_S[\Phi(S)]$ and then to use a concentration bound to argue that $\Phi(S)$ is close to $\mathrm{E}_S[\Phi(S)]$ with high probability.

Write $S = \{z_1, \ldots, z_m\}$ and consider Φ as a function of the independent random variables $z_1, \ldots, z_m$. Notice that Φ satisfies the hypothesis of McDiarmid's inequality. Specifically, if we change the value of the i-th argument from z_i to z_i' then the value of Φ changes by at most $1/m$:

$$\begin{aligned} |\Phi(z_1, \ldots, z_i, \ldots, z_m) - \Phi(z_1, \ldots, z_i', \ldots, z_m)| &\leq \sup_{g \in G} \frac{1}{m} \left| g(z_i) - g(z_i') \right| \\ &\leq \frac{1}{m} . \end{aligned}$$

Applying McDiarmid's inequality with $c = 1/m$ we have that for any $\varepsilon > 0$,

$$\Pr\big(\Phi(S) \geq \mathrm{E}_S[\Phi(S)] + \varepsilon\big) \leq e^{-2m\varepsilon^2} .$$

Taking $\varepsilon = \sqrt{\frac{\log 1/\delta}{2m}}$ we have that with probability at least $1 - \delta$,

$$\Phi(S) \leq \mathrm{E}_S[\Phi(S)] + \sqrt{\frac{\log 1/\delta}{2m}} . \tag{3}$$

Next we give an upper bound for $\mathrm{E}_S[\Phi(S)]$. To this end, suppose that we draw a second sample $S' = \{z_1', \ldots, z_m'\}$ according to distribution D. Then

$$\begin{aligned} \mathrm{E}_S[\Phi(S)] &= \mathrm{E}_S\Big[\sup_{g \in G} \left(L(g) - L_S(g) \right) \Big] \\ &= \mathrm{E}_S\Big[\sup_{g \in G} \mathrm{E}_{S'} \left[L_{S'}(g) - L_S(g) \right] \Big] && (4) \\ &\leq \mathrm{E}_{S,S'}\Big[\sup_{g \in G} \left(L_{S'}(g) - L_S(g) \right) \Big] && (5) \\ &= \mathrm{E}_{S,S'}\Big[\sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} (g(z_i') - g(z_i)) \Big] && (6) \\ &= \mathrm{E}_{S,S',\boldsymbol{\sigma}}\Big[\sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} \sigma_i (g(z_i') - g(z_i)) \Big] && (7) \\ &\leq \mathrm{E}_{S',\boldsymbol{\sigma}}\Big[\sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(z_i') \Big] + \mathrm{E}_{S,\boldsymbol{\sigma}}\Big[\sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} -\sigma_i g(z_i) \Big] && (8) \\ &= 2R_m(G) . && (9) \end{aligned}$$

Line (4) follows from the easily established fact that $L(g) = \mathrm{E}_{S'}[L_{S'}(g)]$. Line (5) follows from the fact that for any family $\{X_i : i \in I\}$ of random variables on a finite probability space, $\sup_{i \in I} \mathrm{E}[X_i] \leq \mathrm{E}[\sup_{i \in I} X_i]$.

The magic is in Line (7), where we introduce the Rademacher random variables. Notice here that setting $\sigma_i = -1$ has the same effect as swapping z_i and z_i'. To see why this step is okay, imagine that after choosing S and S' we update both sets by swapping z_i and z_i' with probability $1/2$ for $i = 1, \ldots, m$. Now this swapping operation leaves the distribution over pairs of sets S and S' unaffected: each possible outcome of S and S' has the same probability before and after the swapping procedure. Thus the expected values in (6) and (7) are equal.

Line (8) follows from the fact that $\sup_{i \in I}(a_i + b_i) \leq \sup_{i \in I} a_i + \sup_{i \in I} b_i$ for real numbers $\{a_i, b_i : i \in I\}$, while Line (9) follows from the fact that σ_i and $-\sigma_i$ are identically distributed.

The above chain of inequalities shows that $\mathrm{E}_S[\Phi(S)] \leq 2R_m(G)$. The statement of the theorem follows from this and (3). □

Next we give an error bound in terms of the empirical Rademacher complexity rather than the expected Rademacher complexity.

Corollary 1. *Suppose that a sample S of size m is drawn according to distribution D. Then for any $\delta > 0$, with probability at least $1 - \delta$ the following holds for all $g \in G$:*

$$L(g) \quad \leq \quad L_S(g) + 2R_S(G) + O\left(\sqrt{\frac{\log \frac{1}{\delta}}{m}}\right) \tag{10}$$

Proof. We may consider the empirical Rademacher complexity $R_S(G) := \mathrm{E}_{\boldsymbol{\sigma}}\left[\sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} z_i\right]$ as a function of the points $z_1, \ldots, z_m$ that comprise the sample S. Changing one of the z_i to a new value z_i' changes $R_S(G)$ by at most $1/m$. Applying McDiarmid's inequality with $c = 1/m$ and $\varepsilon = \sqrt{\frac{\log 2/\delta}{2m}}$, we have that with probability at least $1 - \delta/2$

$$R_S(G) \leq R_m(G) + \sqrt{\frac{\log 2/\delta}{2m}} \tag{11}$$

By a union bound, with probability at least $1 - 2\delta$ the inequalities (2) and (11) both hold. But these two inequalities together imply that (10) holds for all $g \in G$ with probability at least $1 - 2\delta$. Replacing δ by $\delta/2$ gives the required result.

□

References

[1] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. MIT Press, 2012.