A first look at generalization

In these notes, we will get a first look at the theory of generalization for the binary supervised classification problem.

We observe data $(\boldsymbol{x}_i, y_i)$ for $i = 1, \ldots, n$, where the $\boldsymbol{x}_i \in \mathbb{R}^d$ are the "feature vectors" and the $y_i \in \{0, 1\}$ are the "class labels". The data are assumed to be random, in that they are independent samples generated from some joint probability distribution on $\mathbb{R}^d \times \{0, 1\}$, but nothing is known about this probability distribution a priori.

We will use this data to train a classifier $h(\boldsymbol{x})$. A classifier is simply a function which takes a feature vector and returns a class label. We can think of it as a map from $\mathbb{R}^d \rightarrow \{0, 1\}$, or as a partition of $\mathbb{R}^d$ into two regions, one of which corresponds to $\boldsymbol{x}$ that map to label 0, the other with vectors that map to label 1.

To start, we will assume that we only have a finite number of choices for this classifier. We will use the data to decide on one of the classifiers in the set

$$\mathcal{H} = \{h_1, h_2, \ldots, h_m\}.$$

We would like to choose the classifier that minimizes the **risk**, which in this case is simply the probability of error:

$$R(h) = \mathrm{P}\left[h(X) \neq Y\right].$$

The risk tells us what the long-term performance of h will be. Without knowledge of the distribution, however, we cannot compute this risk, so instead we minimize the **empirical risk** ... in this setting, this simply means that we choose the $h \in \mathcal{H}$ that minimizes the number of misclassifications in the training data. The empirical risk of a candidate classifier h working from the n samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$

is

$$\widehat{R}_n(h) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{i:h(\boldsymbol{x}_i)\neq y_i\}}(i),$$

where $1_{\mathcal{A}}$ is our notation for the indicator function:

$$1_{\mathcal{A}}(t) = \begin{cases} 1, & t \in \mathcal{A}, \\ 0, & t \notin \mathcal{A}. \end{cases}$$

In short, $\widehat{R}_n(h)$ is the fraction of the n training samples that h misclassifies. The empirical risk $\widehat{R}_n(h)$ should be thought of as an estimate for the true risk $R(h)$; by the weak law of large numbers, we know that $\widehat{R}_n(h) \to R(h)$ as $n \to \infty$.

We are now faced with the central question: how well does the classifier chosen by empirical risk minimization compare with the best possible classifier in $\mathcal{H}$? Set

$$h^\sharp = \arg\min_{h\in\mathcal{H}} R(h) \quad \text{(best possible classifier)},$$

and

$$h^* = \arg\min_{h\in\mathcal{H}} \widehat{R}_n(h) \quad \text{(our classifier chosen by ERM)}.$$

We are interested in $R(h^*) - R(h^\sharp)$ — sometimes called the *excess risk*, this is the difference in the long-term performance of the classifier we have chosen and the best we could have chosen.

A direct analysis is a little tricky since h^* depends on the (random) data in a rather complicated way, but it turns out that we can analyze this in a fairly straightforward way by first considering the case of a single fixed h.

How close is the empirical risk to the true risk?

We will start by getting a feel for how well we can assess the risk for a particular classifier. With h fixed, we will be looking for a bound on $\widehat{R}_n(h) - R(h)$. We can compute $\widehat{R}_n(h)$ from the data, but $R(h)$ is unknown.

At this point, it is critical to realize that $\widehat{R}_n(h)$ is a random variable, as it depends on the data $(\boldsymbol{x}_i, y_i)$ which is random. So our bounds will be probabilistic; we want something of the form

$$\mathrm{P}\left[|\widehat{R}_n(h) - R(h)| \leq \epsilon\right] \;\geq\; ??,$$

or

$$\mathrm{P}\left[|\widehat{R}_n(h) - R(h)| \geq \epsilon\right] \;\leq\; ??.$$

In both cases, the bound will depend on ϵ (as well as the number of data points n); in the first case, we are looking for the right hand side to be close to 1, in the second case, we are looking for the right hand side to be close to 0.

To get the bound, we will show that $\widehat{R}_n(h)$ is a sum of independent random variables (this is easy), then show that $\mathrm{E}[\widehat{R}_n(h)] = R(h)$ (also easy), and then develop a general-purpose probabilistic tail bound that quantifies how such a sum concentrates around its mean (this is hard).

We start by re-writing the empirical risk as a sum of independent random variables. Let

$$S_i = \begin{cases} 1, & h(\boldsymbol{x}_i) \neq y_i, \\ 0, & h(\boldsymbol{x}_i) = y_i. \end{cases}$$

Since the $(\boldsymbol{x}_i, y_i)$ are independent and identically distributed, the S_i are independent Bernoulli random variables with

$$\mathrm{P}\left[S_i = 1\right] = \mathrm{P}\left[h(\boldsymbol{x}_i) \neq y_i\right], \quad \mathrm{P}\left[S_i = 0\right] = 1 - \mathrm{P}\left[h(\boldsymbol{x}_i) \neq y_i\right].$$

A simple calculation reveals that

$$\mathrm{E}[S_i] = R(h).$$

By construction,

$$\widehat{R}_n(h) = \frac{1}{n}\sum_{i=1}^{n} S_i, \tag{1}$$

and so

$$\mathrm{E}[\widehat{R}_n(h)] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[S_i] = R(h).$$

We are left with the question: How close is the sum of independent random variables $\frac{1}{n}\sum_i S_i$ to its mean?

An answer to this question is given by the Hoeffding inequality:

Hoeffding Inequality. Let $X_1, \ldots, X_n$ be independent random variables that are bounded, meaning $a \leq X_i \leq b$ with probability 1. Let $Z_n = \sum_{i=1}^{n} X_i$. Then for any $\epsilon \geq 0$,

$$\mathrm{P}\left[|Z_n - \mathrm{E}[Z_n]| \geq \epsilon\right] \ \leq \ 2e^{-2\epsilon^2/n(b-a)^2}. \tag{2}$$

Applying this to $\widehat{R}_n(h)$ in (1), with $a = 0, b = 1$, we have

$$\mathrm{P}\left[\left|n\widehat{R}_n(h) - nR(h)\right| \geq n\epsilon\right] \ \leq \ 2e^{-2n\epsilon^2},$$

and so

$$\mathrm{P}\left[|\widehat{R}_n(h) - R(h)| \geq \epsilon\right] \;\leq\; 2e^{-2n\epsilon^2}. \tag{3}$$

This gives us insight into how the performance of **one single** classification rule on the training set generalizes. What we want is some assurance that the one we judge to be the best, by performing ERM on the data, will be close to the best choice we could have made. We will get this assurance by developing a similar probability bound that holds **uniformly** over all classifiers in $\mathcal{H}$.

How close is the empirical minimizer to the true minimizer?

We have linked the performance of a single, fixed classifier to the amount of data n that we have seen. By re-arranging the main result (3) from the previous section, we see that with probability at least $1-\delta$,

$$|\widehat{R}_n(h) - R(h)| \leq \sqrt{\frac{1}{2n}\log(2/\delta)}.$$

But since our decision on which classifier was the best depended on the empirical risk of all of the classifiers in $\mathcal{H}$, we would like to make sure that there empirical performance was somewhat near their ideal performance. That is, we want to show that

$$\max_{h\in\mathcal{H}} |\widehat{R}_n(h) - R(h)| \;\leq\; \epsilon, \tag{4}$$

with probability at least $1-\delta$ for some appropriate choice of ϵ and δ. We want to fill in the right hand side of

$$\mathrm{P}\left[\max_{h\in\mathcal{H}} |\widehat{R}_n(h) - R(h)| > \epsilon\right] \;\leq\; ???.$$

We do this by applying the **union bound** to our expression for a single classifier. Recall the following fact from basic probability theory. If $\mathcal{A}_1, \ldots, \mathcal{A}_m$ are arbitrary events, then the probability of at least one of them occurring is less than the sum of their individual probabilities:

$$\mathrm{P}\left[\mathcal{A}_1 \text{ or } \mathcal{A}_2 \text{ or} \cdots \ \mathcal{A}_m\right] \ \leq \ \mathrm{P}\left[\mathcal{A}_1\right] + \mathrm{P}\left[\mathcal{A}_2\right] + \cdots + \mathrm{P}\left[\mathcal{A}_m\right].$$

As you know, the bound above holds with equality when the sets $\mathcal{A}_i$ are disjoint.

We can rewrite the event of interest as

$$\begin{aligned}\left\{\max_{h\in\mathcal{H}} |\widehat{R}_n(h) - R(h)| > \epsilon\right\} = &\left\{|\widehat{R}_n(h_1) - R(h_1)| > \epsilon\right\} \text{ or} \\ &\left\{|\widehat{R}_n(h_2) - R(h_2)| > \epsilon\right\} \text{ or} \\ &\qquad\vdots \qquad\qquad\qquad\quad \vdots \\ &\left\{|\widehat{R}_n(h_m) - R(h_m)| > \epsilon\right\}.\end{aligned}$$

Thus

$$\begin{aligned}\mathrm{P}\left[\max_{h\in\mathcal{H}} |\widehat{R}_n(h) - R(h)| > \epsilon\right] &\leq \sum_{j=1}^{m} \mathrm{P}\left[|\widehat{R}_n(h_j) - R(h_j)| > \epsilon\right] \\ &\leq 2me^{-2n\epsilon^2}.\end{aligned}$$

When the bound (4) holds, we can relate the generalization performance of the empirical risk minimizer h^* to the performance of the best possible choice $h^\sharp$. We have[1]

$$\begin{aligned}R(h^*) - R(h^\sharp) &= R(h^*) - \widehat{R}_n(h^*) + \widehat{R}_n(h^*) - R(h^\sharp) \\ &\leq |R(h^*) - \widehat{R}_n(h^*)| + |\widehat{R}_n(h^*) - R(h^\sharp)|\end{aligned}$$

[1] Note that $R(h^*) - R(h^\sharp)$ will always be positive.

The first term above is immediately controlled by (4). For the second term, we combine (4) with optimality of $h^\sharp$ and h^* in two different ways. Since $h^\sharp$ is the minimizer of the true risk,

$$R(h^\sharp) \;\leq\; R(h^*) \;\leq\; \widehat{R}_n(h^*) + \epsilon,$$

and since h^* is the minimizer of the empirical risk,

$$\widehat{R}_n(h^*) \;\leq\; \widehat{R}_n(h^\sharp) \;\leq\; R(h^\sharp) + \epsilon.$$

Combining the two statements above gives us $|\widehat{R}_n(h^*) - R(h^\sharp)| \leq \epsilon$, and so

$$\max_{h\in\mathcal{H}} |\widehat{R}_n(h) - R(h)| \;\leq\; \epsilon \quad \Rightarrow \quad R(h^*) - R(h^\sharp) \;\leq\; 2\epsilon.$$

Putting it all together gives us our main result:

$$\mathrm{P}\left[R(h^*) - R(h^\sharp) > \epsilon\right] \;\leq\; 2me^{-n\epsilon^2/2}.$$

ERM with finite $\mathcal{H}$. Let $\mathcal{H}$ be a set of classifiers with finite size $|\mathcal{H}| = m$. We are presented with n iid labeled data points $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$. Let h^* be the empirical risk minimizer,

$$h^* = \arg\min_{h\in\mathcal{H}} \widehat{R}_n(h) = \arg\min_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n 1_{\{i:h(\boldsymbol{x}_i)\neq y_i\}}(i),$$

and $h^\sharp$ be the true risk minimizer

$$h^\sharp = \arg\min_{h\in\mathcal{H}} R(h) = \arg\min_{h\in\mathcal{H}} \mathrm{P}\left[h(X) \neq Y\right].$$

Then with probability exceeding $1-\delta$

$$R(h^*) - R(h^\sharp) \;\leq\; \sqrt{\frac{2}{n}\left(\log m + \log(2/\delta)\right)}.$$

Technical Details: Proof of the Hoeffding Ineq.

We start with a basic question: how close is a single random variable X to its mean? This question is answered by applying the following basic result from probability theory.

Markov inequality. Let X be any non-negative random variable. Then for any $t \geq 0$,

$$\mathrm{P}\left[X \geq t\right] \;\leq\; \frac{\mathrm{E}[X]}{t}.$$

Proof of this statement is straightforward. For convenience, we assume here that X has a probability density function $f_X(x)$, but the result holds regardless:

$$\begin{aligned}
\mathrm{E}[X] &= \int_0^\infty x\, f_X(x)\; dx \\
&\geq \int_t^\infty x\, f_X(x)\; dx \quad (\text{for } t \geq 0) \\
&\geq t \int_t^\infty f_X(x)\; dx \\
&= t \cdot \mathrm{P}\left[X \geq t\right].
\end{aligned}$$

The Markov inequality actually tells us much more than what is in the box above. It is easily extended by realizing that for any function $\phi(x)$ which is non-negative and strictly monotonically increasing,

$$\mathrm{P}\left[X \geq t\right] = \mathrm{P}\left[\phi(X) \geq \phi(t)\right].$$

We now have any number of ways to modify the bound, as

$$\mathrm{P}\left[X \geq t\right] \;\leq\; \frac{E[\phi(X)]}{\phi(t)},$$

for any such ϕ. Moreover, the above holds for general random variables X, as we only need $\phi(X) \geq 0$ to apply Markov.

A **Chernoff bound** is simply an application of Markov with $\phi(t) = e^{\lambda t}$ for some $\lambda > 0$:

$$\mathrm{P}\left[X \geq t\right] \;\leq\; e^{-\lambda t}\,\mathrm{E}[e^{\lambda X}].$$

This is particularly useful when X is a sum of independent random variables. For instance, suppose that $Z_1, Z_2, \ldots, Z_n$ are iid random variables. Then the Chernoff bound on their sum is

$$\begin{aligned}\mathrm{P}\left[Z_1 + \cdots + Z_n \geq t\right] &\leq e^{-\lambda t}\,\mathrm{E}[e^{\lambda(Z_1+\cdots+Z_n)}] \\ &= e^{-\lambda t}\,\mathrm{E}[e^{\lambda Z_1} e^{\lambda Z_2} \cdots e^{\lambda Z_n}] \\ &= e^{-\lambda t}\,\mathrm{E}[e^{\lambda Z_1}]\,\mathrm{E}[e^{\lambda Z_2}] \cdots \mathrm{E}[e^{\lambda Z_n}] \quad \text{(independence)} \\ &= e^{-\lambda t}\left(\mathrm{E}[e^{\lambda Z_1}]\right)^n \quad \text{(identically dist.)}.\end{aligned}$$

Thus we can get a tail bound on the sum by looking at moment generating function (mgf) of one of the terms. Recall that the mgf is the Laplace transform of the density:

$$\mathrm{mgf}_Z(\lambda) = \mathrm{E}[e^{\lambda Z}] = \int e^{\lambda z} f_Z(z)\; dz.$$

To get (2), Hoeffding proved the following lemma:

Let Z be a random variable that falls in the interval $[a, b]$ with probability 1. Then

$$\mathrm{E}\left[e^{\lambda(Z-\mathrm{E}[Z])}\right] \;\leq\; e^{-\lambda^2(b-a)^2/8},$$

for all $\lambda > 0$.

Proof of this is not so straightforward, but in the end it just relies on the convexity of the function $e^{\lambda t}$ combined with the Taylor theorem. The proof is done nicely on Wikipedia[2].

Now if $Z_1, Z_2, \ldots, Z_n$ are iid and fall in $[a, b]$, we have

$$\mathrm{P}\left[\sum_{i=1}^{n} Z_i - \mathrm{E}[Z_i] > t\right] \;\leq\; e^{-\lambda t} e^{n\lambda^2(b-a)^2/8}, \quad \text{for all } \lambda > 0.$$

The value of λ that minimizes the right hand side above is

$$\lambda = \frac{4t}{n(b-a)^2},$$

and so plugging this in a simplifying gives us

$$\mathrm{P}\left[\sum_{i=1}^{n} Z_i - \mathrm{E}[Z_i] > t\right] \;\leq\; e^{-2t^2/n(b-a)^2}.$$

[2] https://en.wikipedia.org/wiki/Hoeffding%27s_inequality