Chapter 1

Statistics of Binary Classification

Let $\mathcal{X}$ be an input space. Consider the following statistical model for labelling elements of $\mathcal{X}$: an underlying unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \{0, 1\}$. We are given a training set $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ drawn iid from $\mathcal{D}$ and wish to "learn" a classifier $g : \mathcal{X} \to \{0, 1\}$ that approximates the distribution $\mathcal{D}$ as best as possible. Consider the following loss of a classifier g on a sample (x, y):

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

The average loss, then, is given by

$$L(g) := \mathbb{E}_{(x,y)\sim\mathcal{D}} [L(g, (x, y))] = \Pr_{(x,y)\sim\mathcal{D}} [g(x) \neq y] .$$

We only have a finite sample to measure losses of potential classifiers: we define the empirical loss of a classifier on the sample by

$$L_n(g) := \frac{1}{n} \sum_{i=1}^{n} L(g, (x_i, y_i)).$$

Notice that it is a random variable since it depends on the random sample S.

Suppose that an algorithm given sample S returns a classifier $\widehat{g}_n$. We are interested in two questions about the classifier.

1. Is $L(\widehat{g}_n)$ close to $\inf\limits_{g\in\mathcal{C}} L(g)$? This is the accuracy question.
2. Is $L(\widehat{g}_n)$ close to $L_n(\widehat{g}_n)$? This is the generalization question.

Consider the natural algorithm for binary classification in Figure 1 that simply chooses the classifier minimizing *empirical loss* (what more do we have to go on?).

1: **Input:** Sample $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$.
2: **Output:** Classifier $\widehat{g}_n$.
3: $\widehat{g}_n = \arg\min_{g\in\mathcal{C}} L_n(g)$.
4: **return** $\widehat{g}_n$.

Figure 1.1: Just minimize Empirical Loss!

Let's estimate its accuracy. Assume that $g^* = \arg\min\limits_{g\in\mathcal{C}} L(g)$ is the best classifier in the class $\mathcal{C}$ (called the Naïve Bayes classifier). We have

$$\begin{aligned} L(\widehat{g}_n) - L(g^*) &= L(\widehat{g}_n) - L_n(\widehat{g}_n) + L_n(\widehat{g}_n) - L(g^*) \\ &\leq L(\widehat{g}_n) - L_n(\widehat{g}_n) + L_n(g^*) - L(g^*) \\ &\leq 2 \sup_{g\in\mathcal{C}} |L_n(g) - L(g)|. \end{aligned}$$

The goal now is to boudn the quantity $\sup_{g\in\mathcal{C}} |L_n(g) - L(g)|$, as n grows large. If it can be shown to go to 0, we have asymptotically perfect accuracy. Notice that for any particular g, the value $L_n(g)$ is a random variable with expectation (over S) being $L(g)$, i.e. $\mathbb{E}_S[L_n(g)] = L(g)$. As n grows large, one expects that $L_n(g)$ is close to $L(g)$. However, do we have a uniform upper bound for all classifiers in $\mathcal{C}$? If $\mathcal{C}$ is finite, this is easy to see; however, in the infinite case, it is not clear. We essentially need a *uniform* LLN.

Remark. For those familiar with some analysis, the situation here is similar to uniform convergence of a sequence of functions; we are looking to check if the sequence of functions $\{L_n\}$ converges uniformly to L.

§ 1.1 Uniform Law of Large Numbers

Suppose that $S = \{X_1, \cdots, X_n\}$ is a list of iid random objects (distributed according to $\mathcal{D}$, say) taking values in $\mathcal{X}$. Let $\mathcal{F}$ be a class of real-valued functions $\mathcal{X} \to \mathbb{R}$. What can we say about the random variable Z over $\mathcal{D}^S$ defined by

$$Z = \sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}_{\mathcal{D}}[f(X)] \right| ?$$

[Of course, this only makes sense when $\mathbb{E}_X[f(X)]$ is well-defined for all $f \in \mathcal{F}$.]

In particular, we are interested in the following three questions.

1. Whether $Z \to 0$ when $n \to \infty$ (asymptotic question).
2. Can we obtain non-asymptotic guarantees, at large n?
3. Can we provide conditions such that Z converges to 0?

We have already seen that the ULLN can analyse binary classification. Another application is in M-estimation.

◊ M-estimation (intro)

Many problems in statistics are concerned with estimators of the form

$$\widehat{\theta}_n = \arg\max_{\theta\in\Theta} \frac{1}{n} \sum_{i=1}^{n} m_\theta(x_i),$$

where $X_1, \cdots, X_n$ are iid observations. Here, Θ is the parameter space, and $m_\theta : \mathcal{X} \to \mathbb{R}$ is a real-valued function parameterized by θ.

Example 1 (familiar M-estimators).

1. $m_\theta(x) = \log p_\theta(x)$ where p is a family of distributions parameterized by θ. This is the maximum likelihood estimator, i.e. $\widehat{\theta}_n$ is the MLE estimate.
2. $m_\theta(x) = -(x-\theta)^2$. This estimator is the sample mean, i.e. $\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$.
3. $m_\theta(x) = -|x-\theta|$. This estimator is the sample median.

In M-estimation, the target quantity for $\widehat{\theta}_n$ (what we want $\widehat{\theta}_n$ to approach) is

$$\theta^* = \arg\max_{\theta\in\Theta} \mathbb{E}[m_\theta(X)].$$

Similar to binary classification accuracy, we want $d(\widehat{\theta}_n, \theta^*) = |\widehat{\theta}_n - \theta^*|$ to be small. It turns out (we will see this later) that

$$d(\widehat{\theta}_n, \theta^*) \leq 2 \sup_{\theta\in\Theta} \left| \frac{1}{n} \sum_{i=1}^{n} m_\theta(X_i) - \mathbb{E}[m_\theta(X)] \right|,$$

which is another instance of the use of the uniform LLN.

$\lozenge$ Statement and Proof

Back to the uniform LLN. To show some results, we first need the following observation.

Key observation. Z concentrates around

$$\mathbb{E}_S\left[Z\right] = \mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}_{\mathcal{D}}\left[f(X)\right]\right|\right].$$

We can then control $\mathbb{E}\left[Z\right]$ through techniques like symmetrization (leading to Rademacher complexity) and chaining (leading to VC dimension).

Remark. A family of functions $\mathcal{F}$ is called Glivenko-Cantelli if $Z \to 0$ a.s. as $n \to \infty$ (more on this later).

Chapter 2

Concentration

§ 2.1 Teaser

check: A term like Z is important in understanding the *generalization* error of ML algorithms.

Recall that we wish to analyze the concentration properties of the random variable

$$Z \equiv Z(S) := \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{x \in S} f(x) - \mathbb{E}\left[f(X)\right] \right|.$$

Assumption. We assume that the functions in $\mathcal{F}$ are uniformly bounded, that is, there exists a $B > 0$ such that $\sup_x |f(x)| \leq B$ for all $f \in \mathcal{F}$.

Theorem 1 (McDiarmid's Inequality). *Suppose* $X_1, \cdots, X_n$ *and* $g : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$ *satisfy bounded difference:*

$$|g(x_1, \cdots, x_i, \cdots, x_n) - g(x_1, \cdots, x_i', \cdots, x_n)| \leq c_i$$

for every choice of variables and index i. *Then, for any* $t > 0$, *we have*

$$\Pr\left[g(X_1, \cdots, X_n) - \mathbb{E}\left[g(X_1, \cdots, X_n)\right] \geq t\right] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

Here, the probability is over the joint distribution of $X_1, \cdots, X_n$.

The inequality essentially says that a function g of n random variables that has bounded difference in each variable is a random variable that concentrates around its expectation.

◊ What makes concentration natural?

The bounded difference condition essentially says that a particular variable can only change the function by a certain amount. A random choice of variables X_i is expected to, *on average*, roughly cancel out increases and decreases in the function: that is, every choice of X_i gives a similar value of the function, i.e. the value of the function stays near some value (its expectation). For the value to deviate significantly from the expectation, a lot of events must conspire to change the function in the same direction, and this is unlikely: remarkably, *exponentially* unlikely.

Of course, if the constants c_i are larger, it's easier to deviate by the same amount t (less things need to go right) and we get a weaker bound.

◊ Does Z concentrate?

We pick g such that McDiarmid gives us concentration for Z.

§ 2.2 Hoeffding

Theorem 2 (Hoeffding's inequality). *Suppose $X_1, \cdots, X_n$ are independent random variables, with $a_i \le X_i \le b_i$ a.s. for each i. Let $S = (X_1 - \mathbb{E}[X_1]) + \cdots + (X_n - \mathbb{E}[X_n])$. Then, for any $t > 0$, we have*

$$\Pr[S \ge t] \le \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

and

$$\Pr[S \le -t] \le \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

That is, S concentrates around 0 (its expectation).

Moment-generating functions are a powerful tool in proving concentration inequalities, especially because of the much tighter Markov's inequality on exponential tails.

The first key trick is to notice that for any $\lambda > 0$, $\Pr[S \ge t] = \Pr\left[e^{\lambda S} \ge e^{\lambda t}\right]$. Markov on the latter variable $e^{\lambda S}$ gives

$$\Pr[S \ge t] \le \frac{\mathbb{E}\left[e^{\lambda S}\right]}{e^{\lambda t}}.$$

Next, we exploit independence and the multiplicativity of independent expectations: $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$ for independent X, Y. This simplifies the numerator to $\prod_{i=1}^n \mathbb{E}\left[e^{\lambda(X_i - \mathbb{E}[X_i])}\right]$. The right-hand side simplifies to

$$\prod_{i=1}^n \mathbb{E}\left[e^{\lambda X_i}\right] = \exp\left(-\lambda t + \psi(\lambda)\right),$$

where $\psi(\lambda) = \sum_{i=1}^n \log \mathbb{E}_{X_i}\left[e^{\lambda(X_i - \mathbb{E}[X_i])}\right]$. To bound the log terms in terms of λ, we look at the Taylor series expansion of

$$\psi_{\mathcal{D}}(\lambda) := \log \mathbb{E}_{x\sim\mathcal{D}}\left[\exp(\lambda x)\right].$$

When studying a weird function, it is often useful to study a partial Taylor series for the function. In this case, we look at the degree-2 Taylor expansion of $\psi_{\mathcal{D}}(\lambda)$ around $\lambda = 0$:

$$\psi_{\mathcal{D}}(\lambda) = \psi_{\mathcal{D}}(0) + \lambda \psi'_{\mathcal{D}}(0) + \frac{\lambda^2}{2}\psi''_{\mathcal{D}}(\lambda')$$

for some $\lambda' \in (0, \lambda)$. In our case, we remark x is the variable $X_i - \mathbb{E}[X_i]$: it has expectation 0 and is between $a_i - \mathbb{E}[X_i]$ and $b_i - \mathbb{E}[X_i]$ a.s.

The first term is $\log \mathbb{E}[1] = 0$; the second term is

$$\frac{d}{d\lambda} \log \mathbb{E}_{x\sim\mathcal{D}}\left[\exp(\lambda x)\right]\bigg|_{\lambda=0} = \frac{1}{\mathbb{E}\left[\exp(\lambda x)\right]} \cdot \mathbb{E}\left[x \exp(\lambda x)\right]\bigg|_{\lambda=0} = \mathbb{E}[x] = 0.$$

The third term is

$$\frac{d}{d\lambda}\frac{\mathbb{E}\left[x\exp(\lambda x)\right]}{\mathbb{E}\left[\exp(\lambda x)\right]} = \frac{\mathbb{E}\left[x^2\exp(\lambda x)\right]\mathbb{E}\left[\exp(\lambda x)\right] - \mathbb{E}\left[x\exp(\lambda x)\right]^2}{\mathbb{E}\left[\exp(\lambda x)\right]^2} = \mathbb{E}\left[x^2\frac{\exp(\lambda x)}{\mathbb{E}\left[\exp \lambda x\right]}\right] - \left(\mathbb{E}\left[x\frac{\exp(\lambda x)}{\mathbb{E}\left[\exp \lambda x\right]}\right]\right)^2.$$

This is certainly not zero, and is fairly complicated to upper-bound. We use a neat trick here to

Chapter 3

Central Limit Theorem

Hoeffding from last time captured the concentration of the sum of iid random variables: the sum concentrates around its expectation. The central limit theorem goes further to characterize its distribution in the limit as the number of variables $N \to \infty$.

§ 3.1 Teaser: CLT for Bernoulli variables, aka Random Walks

§ 3.2 The Central Limit Theorem

Theorem 3 (CLT). *Let $X_1, \cdots, X_n$ be iid random variables with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}_{X_i}[=] \sigma^2$. Let $S_n = X_1 + \cdots + X_n$. Then, for any $t \in \mathbb{R}$,*

$$\Pr\left[\frac{S_n - n\mu}{\sigma\sqrt{n}} \geq t\right] \to \Pr_{Z \sim \mathcal{N}(0,1)}[Z \geq t] \quad \textit{as } n \to \infty.$$

That is, the standardized sum $\frac{S_n - n\mu}{\sigma\sqrt{n}}$ converges in distribution to a standard normal as $n \to \infty$.

One may also re-write it as

$$\sqrt{n}(S_n - n\mu) \overset{\text{dist}}{\to} \mathcal{N}(0, \sigma^2).$$

Let's upper bound $\psi(t) := \Pr[Z \geq t]$ for $Z \sim \mathcal{N}(0, \sigma^2)$.

Exercise. Let $Z \sim \mathcal{N}(0, \sigma^2)$. Show that its EGF is

$$\mathbb{E}\left[e^{\lambda Z}\right] = \exp\left(\frac{\lambda^2\sigma^2}{2}\right).$$

Now we use a similar MGF-idea to the proof of Hoeffding to bound $\psi(t)$. That is, we let $\lambda > 0$ and note

$$\psi(t) = \Pr\left[e^{\lambda Z} \geq e^{\lambda t}\right] \leq \frac{\mathbb{E}\left[e^{\lambda Z}\right]}{e^{\lambda t}} = \exp\left(-\lambda t + \frac{t^2\sigma^2}{2}\right).$$

Minimize the right-hand side over λ to get

$$\psi(t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Suppose now for a moment that n is very large, so we may reasonably assume that

$$\Pr\left[\sqrt{n}(S_n - n\mu) \geq t\right] \approx \Pr[Z \geq t] \leq \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Let's compare this bound to that from Hoeffding. Suppose the variables X_i are bounded a.s. between a and b and have variance σ^2. Hoeffding gives for $t' > 0$

$$\Pr\left[S_n - n\mu \geq t'\right] \leq \exp\left(-\frac{2t'^2}{n(b-a)^2}\right).$$

Let $t = t'/\sqrt{n}$. This yields

$$\Pr\left[\sqrt{n}(S_n - n\mu) \geq t\right] \leq \exp\left(-\frac{2t^2}{(b-a)^2}\right).$$

How do we relate σ^2 to $(b-a)^2$?

Exercise. Show that for any random variable X with $\mathrm{Var}_X\left[=\right] \sigma^2$ and $a \leq X \leq b$ a.s., we have

$$\sigma^2 \leq \frac{(b-a)^2}{4}.$$

(Hint: What is the worst case?)

This means

$$\frac{-t^2}{2\sigma^2} \leq \frac{-2t^2}{(b-a)^2} \implies \exp\left(-\frac{t^2}{2\sigma^2}\right) \leq \exp\left(-\frac{2t^2}{(b-a)^2}\right).$$

The bound from CLT is tighter than Hoeffding, sometimes substantially, since σ^2 could be much smaller than $(b-a)^2/4$.

However, the obvious strength of Hoeffding and drawback of CLT is that Hoeffding is exact, while CLT is an asymptotic result. It does not mention how bad the approximation is at a particular n and value of $t > 0$.

Remark. The Berry-Esseen theorem is a quantitative form of the CLT that gives a bound on the error in the approximation. In particular, one form states the following. Suppose $X_1, \cdots, X_n$ have mean μ, variance σ^2 and absolute third moment $\rho := \mathbb{E}_X\left[|X-\mu|^3\right] < \infty$. Then, for any $t > 0$,

$$\left|\Pr\left[\frac{S_n - n\mu}{\sigma\sqrt{n}} \geq t\right] - \Pr_{Z\sim\mathcal{N}(0,1)}\left[Z \geq t\right]\right| \leq \frac{C\rho}{\sigma^3\sqrt{n}}.$$

Before moving on, let's simplify notation in the CLT to make it more memorable. Suppose $\mu = 0$, and let $\overline{X}_n = S_n/n$ be the sample mean. Let $F_n(t)$ denote the cumulative distribution function (CDF) of $\overline{X}_n/\sigma\sqrt{n}$, and $\Phi(t)$ denote the CDF of $Z \sim \mathcal{N}(0,1)$. Then, the CLT states

$$F_n(t) \to \Phi(t) \quad \text{as } n \to \infty$$

and the Berry-Esseen theorem states that

$$|F_n(t) - \Phi(t)| \leq \frac{C\mathbb{E}\left[|X_i|^3\right]}{\sigma^3\sqrt{n}}.$$

◊ Symmetric and Assymetric bounds

Certain bounds on tail probabilities (expressions of the form $\Pr\left[X \geq t\right]$ or $\Pr\left[X \leq -t\right]$) are *symmetric*. That is, turning each variable $X_i \mapsto -X_i$ does not change the bound but turns an upper tail probability (e.g. $\Pr\left[X \geq t\right]$) into the corresponding lower tail probability (e.g. $\Pr\left[X \leq -t\right]$).

We saw that Hoeffding's inequality is symmetric. The variables $-X_i$ satisfy the constraints $-b \leq -X_i \leq -a$ a.s., and the (upper-tail) bound becomes

$$\Pr\left[-S_n - (-\mu) \geq t\right] \leq \exp\left(-\frac{2t^2}{n(b-a)^2}\right),$$

or equivalently

$$\Pr\left[S_n - \mu \leq -t\right] \leq \exp\left(-\frac{2t^2}{n(b-a)^2}\right),$$

which is the corresponding lower-tail bound. Symmetric bounds are useful since they concentrate the random variable around its expectation from both sides. For example, adding both tails, Hoeffding can be written

$$\Pr\left[|S_n - n\mu| \geq t\right] \leq 2\exp\left(-\frac{2t^2}{n(b-a)^2}\right).$$

The CLT is also symmetric.

Exercise. Show that the CLT is symmetric:

$$\Pr\left[\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq -t\right] \overset{\text{dist}}{\to} \Pr_{Z\sim\mathcal{N}(0,1)}\left[Z \geq t\right].$$

Show that this means

$$\Pr\left[\left|\frac{\sqrt{n}(\overline{X}_n - \mu)}{}\right|\right]$$

§ 3.3 Confidence Intervals

Suppose $X_1, \cdots, X_n$ iid $\sim X$ have $\mathbb{E}[X] = \mu$ and $\mathrm{Var}_X[=]\sigma^2$. Also suppose $a \leq X \leq b$ a.s. Here is the problem.

We knoe (a, b, σ^2) and want to estimate μ. We obtain a set (interval) known as a confidence interval (CI) in which μ lies with high probability. A CI is of the form

$$u \leq \mu \leq v \quad \text{with probability} \geq 1-\alpha.$$

Let's use (approximate) CLT. Let $t = z_{\alpha/2}$ be the α-quantile of the Gaussian. That is, $\Pr\left[|Z| \leq z_{\alpha/2}\right] = 1-\alpha$. (Approximate) CLT gives

$$\Pr\left[\left|\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}\right| \leq z_{\alpha/2}\right] \approx 1-\alpha,$$

or with probability at least $1-\alpha$,

$$-z_{\alpha/2} \leq \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \leq z_{\alpha/2} \implies \overline{X}_n - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}_n + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}.$$

Now, let's use Hoeffding. What do we expect? We have for $t > 0$,

$$\Pr\left[|\sqrt{n}(\overline{X}_n - \mu)| \geq t\right] \leq 2\exp\left(-\frac{2t^2}{n(b-a)^2}\right).$$

Let $t = (b-a)\sqrt{\frac{1}{2}\log\frac{2}{\alpha}}$. This gives that w.p. $\geq 1-\alpha$, we have

$$\mu \in \left[\overline{X}_n - (b-a)\sqrt{\frac{1}{2}\log\frac{2}{\alpha}}, \overline{X}_n + (b-a)\sqrt{\frac{1}{2}\log\frac{2}{\alpha}}\right].$$

Exercise. Show that the interval from CLT is smaller than that from Hoeffding.

§ 3.4 Sub-Gaussian random variables

We visit

Precisely what did we need in the proof of Hoeffding.

Reading: Subexponential random variables. Martingales. AKA chapter 2 of HDS (Martin).