The Empirical Risk: Have a dataset

$$\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

where $x_i \in \mathcal{X}$ is an example and $y_i \in \mathcal{Y}$ is a label.

Let $h : \mathcal{X} \to \mathcal{Y}$ be a hypothesized model we are trying to evaluate.
Let $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a loss function which measures how different two labels are.

The empirical risk is

$$R(h) = \frac{1}{n} \sum_{i=1}^{n} L(h(x_i), y_i)$$

We need to compute the empirical risk a lot during training, both during validation and testing,

But the cost will be proportional to n, the number of training examples.

We can approximate the empirical risk using subsampling

Idea: Let Z be a random variable that takes on value $L(h(x_i), y_i)$ w.p. $1/n$ $\forall i \in \{1, \ldots, n\}$

Equivalently, Z is the result of sampling a single element from the sum in the formula for empirical risk.

By construction of Z, have

$$\mathbb{E}[Z] = \frac{1}{n} \sum_{i=1}^{n} L(h(x_i), y_i) = R(h)$$

If we sample a bunch of iid $Z_i \sim Z$, then their average will be a good approximation of the empirical risk!

That is,

$$S_K = \frac{1}{K} \sum_{i=1}^{K} Z_i \approx R(h)$$

Approximation does **NOT** depend on n!

This is an instance of the statistical principle that the average of a collection of independent random variables tends to cluster around their mean.

We formalize this with the **strong law of large numbers** which says that

$$\Pr\left(\lim_{K\to\infty} \sum_{k=1}^{K} Z_k = \mathbb{E}[Z] = R(h)\right) = 1$$

i.e as #samples $\to \infty$ their average $\to$ mean almost surely,

This gives estimate of average but NOT what the distribution looks like.

If our random variables Z have bounded mean and variance, then

$$\sqrt{K}\left(\frac{1}{K}\sum_{k=1}^{K} Z_k - \mathbb{E}[Z]\right) \xrightarrow{K} N(0, \mathrm{Var}(Z))$$

Problem: LLN and CLT tell us that our approximations will be asymptotically accurate, but NOT how long we need to average to get approximations we can be confident in.

To address this, need a **concentration inequality**: a formula which lets us bound what the probability a finite sum will diverge from its expected value will be.

Markov Inequality

If S is a rv with finite expected value, then $\forall a > 0$

$$Pr(S \geq a) \leq \frac{\mathbb{E}[S]}{a}$$

This gives the following bound for our empirical risk:

Assume $L(y, \vec{y}) \geq 0$, $L(\vec{y}, y) \leq 1$

Want $Pr(S \geq a) \leq \frac{R(h)}{a} \rightarrow Pr(S \geq a) \leq \varepsilon$

$$\Rightarrow a = \frac{R(h)}{\varepsilon}$$

$$Pr(S_K \geq a) = Pr\left(\frac{1}{K}\sum_{k=1}^{K} z_k \geq \frac{R(h)}{\varepsilon}\right) \leq \varepsilon$$

$$\downarrow$$

$$Pr\left(\underbrace{S_K - R(h)}_{\text{estimate} - \text{actual avg}} \geq R(h)\left(\frac{1}{\varepsilon} - 1\right)\right) \leq \varepsilon$$

Another concentration inequality is Chebyshev's inequality. This inequality uses the variance of the random variable, in addition to its expected value, to bound its distance from its expected value.

If S is a rv w/ finite expected value and variance, then $\forall a > 0$

$$Pr(|S - \mathbb{E}[S]| \geq a) \leq \frac{Var(z)}{a^2}$$

Suppose a 0-1 loss.

$$L(y, \hat{y}) = \begin{cases} 1 & \text{if } y \neq \hat{y} \\ 0 & \text{if } y = \hat{y} \end{cases}$$

$$\text{Var}(Z_k) = \mathbb{E}\left[(Z_k - \mathbb{E}[Z_k])^2\right] \leq 1$$

(after stuff) $\leq \frac{1}{4}$

From Chebyshev's inequality, can get

$$\Pr\left(\left|\frac{1}{K}\sum_{k=1}^{K} Z_k - R(h)\right| \geq a\right) \leq \frac{\text{Var}(S_K)}{a^2}$$

$$\leq \frac{1}{a^2}\text{Var}\left(\frac{1}{K}\sum_{k=1}^{K} Z_k\right)$$

$$= \frac{1}{a^2K^2}\text{Var}\left(\sum_{k=1}^{K} Z_k\right)$$

$$= \frac{1}{a^2K^2}\sum_{k=1}^{K}\text{Var}(Z_k)$$

$$= \frac{1}{a^2K}\text{Var}(Z)$$

$$\leq \frac{1}{4a^2K}$$

A useful bound!

Example

Estimate empirical risk with 0-1 loss to within 10% error w.p. 99%. How many samples K do we need to average up if we use this Chebyshev's inequality bound?

Set $a = 0.1$, $\frac{1}{4a^2K} \leq 1 - 0.99 = 0.01$

$$K = \frac{1}{4(0.1)^2(0.01)} = 2500$$

$\rightarrow K \geq 2500$

Problem: this is just the number of samples we need to evaluate the empirical risk of a single model.

We may want to approximate the empirical risk MANY times during training, either to validate a model or monitor convergence of training loss.

For example, suppose we have M hypotheses we want to validate $(h^{(1)}, \ldots, h^{(M)})$, and we use independent subsamples $(S_K^{(1)}, \ldots, S_K^{(M)})$ to approximate the empirical risk for each of them.

What bound can we get using Chebyshev's inequality on the probability that <u>all</u> T of our independent approximations are within a distance a of their true empirical risk?

$$\Pr\left(\left|S_K^{(m)} - R(h^{(m)})\right| \leq a \quad \forall m \in \{1,\ldots,M\}\right) \geq$$

Now if we want to estimate the empirical risk w/ 0-1 loss to within the same 10% error rate w/ the same probability of 99%, but for all $M=100$ hypotheses, how many samples do we need according to this Chebyshev bound?

$K \geq$

- We need a lot more than we did for one-hypothesis case
- Problem!

A better bound. Use <u>Hoeffding's inequality</u> which gives us a MUCH tighter bound on the tail probabilities of a sum.

Hoeffding's Inequality

If $Z_1, \ldots, Z_K$ are iid and

$$S_K = \frac{1}{K} \sum_{k=1}^{K} Z_k$$

then if $z_{min} \leq Z_k \leq z_{max}$, then

$$P_r\left(|S_k - \mathbb{E}[S_k]| \geq a\right) \leq 2 \exp\left(-\frac{2Ka^2}{(z_{max} - z_{min})^2}\right)$$

Estimate empirical risk with 0-1 loss to within 10% error w.p. 99%. How many samples K do we need to average up if we use this Hoeffding's inequality bound?

For 0-1 loss, want

$$\Pr\left(\left|S_K - \mathbb{E}[S_K]\right| \geq a\right) \leq \varepsilon$$

know

$$Z_{min} = 0, \quad Z_{max} = 1$$

have

$$2\exp\left(\frac{-2Ka^2}{1}\right) \leq \varepsilon$$

$$\rightarrow K \geq \frac{1}{2a^2} \log\left(\frac{2}{\varepsilon}\right)$$

$$a = 0.1, \quad \varepsilon = 0.01 \rightarrow \underline{K = 265} \quad \text{a LOT less}$$

Now if we want to estimate the empirical risk w/ 0-1 loss to within the same 10% error rate w/ the same probability of 99%, but for all $M = 100$ hypotheses, how many samples do we need according to this Hoeffding's bound?

$$\Pr\left(\left|S_K^{(m)} - \mathbb{E}[S_K^{(m)}]\right| \geq a \neq m \in \{1, \ldots, M\}\right) = \prod_{i=1}^{M} \Pr\left(\left|S_K - \mathbb{E}[S_K]\right| \geq a\right)$$

$$\geq \prod_{i=1}^{m} (1-\varepsilon) = (1-\varepsilon)^M$$

$$\simeq e^{-\varepsilon M}$$