

Lecture 2: Estimating the empirical risk with samples

CS4787 — Principles of Large-Scale Machine Learning Systems

Review: The empirical risk. Suppose we have a dataset $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, where $x_i \in \mathcal{X}$ is an example and $y_i \in \mathcal{Y}$ is a label. Let $h : \mathcal{X} \rightarrow \mathcal{Y}$ be a hypothesized model (mapping from examples to labels) we are trying to evaluate. Let $L : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ be a *loss function* which measures how different two labels are. The *empirical risk* is

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

Most notions of error or accuracy in machine learning can be captured with an empirical risk. A simple example: measure the error rate on the dataset with the *0-1 Loss Function*

$$L(\hat{y}, y) = 0 \text{ if } \hat{y} = y \text{ and } 1 \text{ if } \hat{y} \neq y.$$

Other examples of empirical risk?

We need to compute the empirical risk a lot during training, both during validation (and hyperparameter optimization) and testing, so it's nice if we can do it fast.

Question: how does computing the empirical risk scale? Three things affect the cost of computation.

- The number of training examples n . The cost will certainly be proportional to n .
- The cost to compute the loss function L .
- The cost to evaluate the hypothesis h .

Question: what if n is very large? Must we spend a large amount of time computing the empirical risk?

Answer: We can approximate the empirical risk using subsampling.

Idea: let Z be a random variable that takes on the value $L(h(x_i), y_i)$ with probability $1/n$ for each $i \in \{1, \dots, n\}$. Equivalently, Z is the result of sampling a single element from the sum in the formula for the empirical risk. By construction we will have

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

If we sample a bunch of independent identically distributed random variables Z_1, Z_2, \dots, Z_K identical to Z , then their average will be a good approximation of the empirical risk. That is,

$$S_K = \frac{1}{K} \sum_{k=1}^K Z_k \approx R(h).$$

Question: what is the cost of computing this approximation to the empirical risk? Does it 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 can formalize this asymptotically with the *strong law of large numbers*, which says that

$$\mathbf{P} \left(\lim_{K \rightarrow \infty} \frac{1}{K} \sum_{k=1}^K Z_k = \mathbf{E}[Z] = R(h) \right) = 1,$$

i.e. that as the number of samples approaches infinity, their average converges to their mean almost surely.

Problem: the law of large numbers tells us that our approximations will be asymptotically accurate, but not what the distribution of the average will look like. To get this, we use the *central limit theorem*, which characterizes the behavior of large sums of independent variables. If our random variables Z have bounded mean and variance, then

$$\sqrt{K} \left(\frac{1}{K} \sum_{k=1}^K Z_k - \mathbf{E}[Z] \right) \text{ converges in distribution to } \mathcal{N}(0, \text{Var}(Z)) \text{ as } K \rightarrow \infty.$$

Here, $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2 . That is, pretty much any large enough average is going to look like a bell curve.

Problem: the law of large numbers and the central limit theorem 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 problem, we need something called a *concentration inequality*: a formula that lets us bound for a finite sum what the probability it will diverge from its expected value will be.

The granddaddy of all concentration inequalities is **Markov's inequality**, which states that if S is a non-negative random variable with finite expected value, then for any constant $a > 0$,

$$\mathbf{P}(S \geq a) \leq \frac{\mathbf{E}[S]}{a}.$$

What bound do we get for our empirical risk sum using Markov's inequality? Is this a useful bound?

$$\mathbf{P}\left(\frac{1}{K} \sum_{k=1}^K Z_k \geq a\right) \leq$$

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 non-negative random variable with finite expected value and variance, then for any constant $a > 0$,

$$\mathbf{P}(|S - \mathbf{E}[S]| \geq a) \leq \frac{\mathbf{Var}(Z)}{a^2}.$$

What bound can we get using Chebyshev's inequality if we are using a 0-1 loss? Is this a useful bound?

$$\mathbf{P}\left(\left|\frac{1}{K} \sum_{k=1}^K Z_k - R(h)\right| \geq a\right) \leq$$

Important consequence of Chebyshev's inequality: if we want to get error ϵ with probability δ , it suffices to choose a number of examples such that...

One problem with Chebyshev's inequality: the probabilities we are getting are not that small. Since we know that the sums are approaching something like a Gaussian distribution, we'd expect the probability of diverging some amount from the expected value to decrease exponentially as a increases, since this is what happens for a Gaussian. But Chebyshev's inequality only gives us a polynomial decrease. In this case, we can use *Hoeffding's inequality*, which gives us a much tighter bound on the tail probabilities of a sum. Hoeffding's inequality states that if Z_1, \dots, Z_K are independent random variables, and

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

then if those variables are bound absolutely by $z_{\min} \leq Z_k \leq z_{\max}$, then

$$\mathbf{P}(|S - \mathbf{E}[S]| \geq a) \leq 2 \exp\left(-\frac{2Ka^2}{(z_{\max} - z_{\min})^2}\right).$$

Important consequence of Hoeffding's inequality: for 0-1 loss, if we want to get error ϵ with probability δ , it suffices to choose a number of examples such that...

Many other concentration inequalities exist.

- *Azuma's inequality* for when the components of your sum Z_k are not independent.
- *Bennett's inequality* for when you want to take the variance into account in addition to the absolute bounds.

Takeaway point: concentration inequalities

1. let you be confident that your subsampled estimate of the empirical risk is a "good" estimate, and
2. tell you how many samples you need for to have a certain level of confidence.