Lecture 6: Estimating the empirical risk with samples

CS4787 — Principles of Large-Scale Machine Learning Systems

Recall: The empirical risk. Suppose we have a dataset $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, 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 (mapping from examples to labels) 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 hyperparameter optimization) and testing, so it's nice if we can do it fast. But the cost will certainly be proportional to n, the number of training examples.

Question: what if n is very large? Must we spend a long 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 $i \in \{1, \ldots, n\}$. 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

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

If we sample a bunch of independent identically distributed random variables $Z_1, Z_2, \ldots, Z_K$ 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? 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

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

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

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

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 $a > 0$,

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

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 $a > 0$,

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

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

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

Activity: if we want to estimate the empirical risk with 0-1 loss to within 10% error (i.e. $|S_K - R(h)| \leq 10\%$) with probability 99%, how many samples K do we need to average up if we use this Chebyshev's inequality bound?

$K \geq$

Problem: this is just the number of samples we need to evaluate the empirical risk of a single model. But we may want to approximate the empirical risk many times during training, either to validate a model or to 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 $(S_K^{(1)}, \ldots, S_K^{(M)}$, each of size $K)$ to approximate the empirical risk for each of them. What bound can we get using Chebyshev's inequality on the probability that **all** T of our independent approximations are within a distance a of their true empirical risk?

$\mathbf{P}\left(\left|S_K^{(m)} - R(h^{(m)})\right| \leq a \text{ for all } m \in \{1, \ldots, M\}\right) \geq$

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

$K \geq$

- We needed a lot more than we did for the one-hypothesis case.
- This seems to be a problem for training where we want to validate potentially thousands of models across potentially hundreds of epochs.
- The 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.

A better bound. 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 $Z_1, \ldots, Z_K$ are independent random variables, and

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

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

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

Activity: if we want to estimate the empirical risk with 0-1 loss to within 10% error (i.e. $|S_K - R(h)| \leq 10\%$) with probability 99%, how many samples K do we need to average up if we use this Hoeffding's inequality bound?

$K \geq$

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

$K \geq$

Takeaway: the Hoeffding's inequality bound is much tighter, and scales better with the number of times we want to estimate using subsampling. We can use this sort of bound to estimate the number of samples we need to use to estimate a sum like the empirical risk to within some level of accuracy with high probability.

General 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.**

3. **This is useful not just for subsampling for efficiency, but also for bounding the errors that result from using a sample of test/validation data rather than the exact statistics on the true "real-world" test distribution.**

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.