Lecture 3: Exponential Concentration Inequalities and ERM

CS4787 — Principles of Large-Scale Machine Learning Systems

Review: Chebyshev's inequality. Recall: the *0-1 empirical risk* (i.e. the error rate) is

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

where δ here is the Kronecker delta function $\delta(\hat{y}, y) = 1$ if $\hat{y} = y$ and 0 otherwise. Let Z be a random variable that takes on the value $L(h(x_i), y_i)$ with probability $1/n$ for each $i \in \{1, \ldots, n\}$. 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,

$$S_K = \frac{1}{K} \sum_{k=1}^{K} Z_k \approx R(h) \quad \text{and} \quad \mathbf{E}\left[S_K\right] = \mathbf{E}\left[Z\right] = R(h).$$

For this sum, Chebyshev's inequality says that

$$\mathbf{P}\left(\left|S_K - \mathbf{E}\left[S_K\right]\right| \geq a\right) \leq \frac{\mathbf{Var}\left(S_K\right)}{a^2} = \frac{\mathbf{Var}\left(Z\right)}{a^2 K} \leq \frac{1}{4a^2 K}$$

where the last inequality holds specifically for the 0-1 empirical risk, since in this case Z is a Bernoulli random variable (that is, it is supported on $Z \in \{0, 1\}$) and the largest variance a Bernoulli random variable can have is $1/4$.

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 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 1% 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.

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.

Empirical Risk Minimization.

We don't just want to estimate the empirical risk: we also want to minimize it. To do so, we *parameterize* the hypotheses using some parameters $w \in \mathbb{R}^d$. That is, we assign each hypothesis a d-dimensional vector of parameters and vice versa and solve the optimization problem

$$\text{minimize: } R(h_w) = \frac{1}{n} \sum_{i=1}^{n} L(h_w(x_i), y_i) \text{ over } w \in \mathbb{R}^d$$

where h_w denotes the hypothesis associated with the parameter vector w. Often, we denote this more explicitly as a function of w, the weights, as

$$f(w) = R(h_w) = \tfrac{1}{n} \textstyle\sum_{i=1}^{n} L(h_w(x_i), y_i) = \tfrac{1}{n} \textstyle\sum_{i=1}^{n} f_i(w).$$

This is an instance of a principle of scalable ML: **Write your learning task as an optimization problem, then solve it with an optimization algorithm.**

Gradient descent (GD). Decrease the value of the empirical risk iteratively by running

$$w_{t+1} = w_t - \alpha_t \cdot \nabla f(w_t) = w_t - \alpha_t \cdot \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(w)$$

where w_t is the value of the parameter vector at time t, α_t is a parameter called the *learning rate* or *step size*, and ∇f denotes the gradient (vector of partial derivatives) of f. **What does this cost to compute?**

Stochastic gradient descent (SGD). Apply the subsampling principle to gradient descent:

$$\text{pick } i \text{ uniformly at random from } \{1, \ldots, n\}, \text{ then update } w_{t+1} = w_t - \alpha_t \cdot \nabla f_i(w_t).$$

At each step, we pick a new random example from the dataset and update the parameters based only on that example. **How does this affect the computational cost of the update?**