9.520 – Math Camp 2009

Probability Theory

When we design a learning algorithm that maps training data S into a function f_S, we want f_S to be predictive at points that aren't in the dataset. We formalize predictivity as *generalization*. This is a statement about probabilities: for any $\varepsilon > 0$,

$$\lim_{n \to \infty} \mathbf{P}\{|I_S[f_S] - I[f_S]| \geq \varepsilon\} = 0$$

where n is the number of training samples. Remember that $I_S[f_S] = \frac{1}{n} \sum_i V(f_S(x_i), y_i)$ is the empirical error and $I[f_S] = \int V(f_S(x_i), y_i) d\mathbf{P}(x_i, y_i)$ is the expected/true error with respect to the (unknown!) distribution from which the points (x_i, y_i) are drawn. This basically says that, as the number of samples increases, the empirical error should get close to the true error.

Goal: We'll try here to make sense of this definition of generalization (and hopefully other statements that look like it).

First, some definitions.

A **random variable** X, for our purposes, is defined by its range of possible values (assume this is $\mathbb{R}$) and a **probability distribution** over *events*, a.k.a. sets of those values, written $\mathbf{P}(A)$ where $A \subset \mathbb{R}$. It's common think of probability distributions in terms of the associated **cumulative distribution function**, $\text{cdf}(x) = \mathbf{P}(X \leq x)$, where $x \in \mathbb{R}$, and/or the **probability density function** $\mathbf{p}(x) = \frac{d\text{cdf}(x)}{dx}$ (in other words, $\text{cdf}(x) = \int_{-\infty}^{x} \mathbf{p}(x) dx$).

A collection of random variables $\{X_n\}$ is **independent and identically distributed** if $\mathbf{p}_{X_1, X_2, \ldots}(X_1 = x_1, X_2 = x_2, \ldots) = \prod\limits_i \mathbf{p}_{X_1}(X_i = x_i) = \prod\limits_i \mathbf{p}_{X_2}(X_i = x_i) = \ldots$.

The **expectation** (mean) of a random variable is given by

$$\mathbf{E}X \triangleq \int x d\mathbf{P}(x)$$

where we will think of $d\mathbf{P}(x)$ as being $\mathbf{p}(x)dx$.

Now we'll get into the interesting stuff.

The problem: We want to prove things about the probability of $I_S[f_S]$ being close to $I[f_S]$. But *we don't know how $I_S[f_S]$ and $I[f_S]$ are distributed!* (We don't know the underlying distribution of the datapoints (x_i, y_i).) How do we bound the probability of certain events (like $|I_S[f_S] - I[f_S]| \geq \varepsilon$) if we don't know how they're distributed?

The solution: **Concentration inequalities**. These inequalities put bounds on the probability of an event (like $X \geq c$), in terms of only some limited information about the actual distribution involved (say, X's mean). We can prove that any distribution that is consistent with our limited information must concentrate its probability density around certain events.

Say we know the expectation of a random variable. Then we can apply **Markov's Inequality**: Let X be a non-negative-valued random variable. Then for any constant $c > 0$

$$\mathbf{P}(X \geq c) \leq \frac{\mathbf{E}X}{c}$$

More generally, if $f(x)$ is a non-negative function, then

$$\mathbf{P}(f(X) \geq c) \leq \frac{\mathbf{E}f(X)}{c}$$

Proof. We'll prove the former, although the proof for nonnegative $f(X)$ is essentially the same.

$$\begin{aligned}
\mathbf{E}X &= \int_0^{+\infty} x\mathbf{p}(x)dx \\
&\geq \int_c^{+\infty} x\mathbf{p}(x)dx \\
&\geq c\int_c^{+\infty} \mathbf{p}(x)dx \\
&= c[\mathbf{P}(x < +\infty) - \mathbf{P}(X < c)] \\
&= c\mathbf{P}(X \geq c)
\end{aligned}$$

Rearranging this gives the inequality. □

Now say we know both the expectation and the variance. We can use Markov's inequality to derive **Chebychev's Inequality**: Let X be a random variable with finite variance σ^2, and define $f(X) = |X - \mathbf{E}X|$. Then for any constant $c > 0$, Markov's inequality gives us

$$\mathbf{P}(|X - \mathbf{E}X| \geq c) = \mathbf{P}((X - \mathbf{E}X)^2 \geq c^2) \leq \frac{\mathbf{E}(X - \mathbf{E}X)^2}{c^2} = \frac{\sigma^2}{c^2}$$

Example: What's the probability of a 3σ event if all we know about the random variable X is its mean and variance? (Hint: the answer is that it's $\leq \frac{1}{9}$)

When we talk about generalization, we are talking about **convergence** of a sequence of random variables, $I_S[f_S]$, to a limit $I[f_S]$. Random variables are defined by probability distributions over their values, though, so we have to define what convergence means for sequences of distributions. There are several possibilities and we'll cover one.

First, a reminder: **plain old convergence** means that you have a sequence $\{x_n\}_{n=1}^{\infty}$ in some space with a distance $|y - z|$ and the values get arbitrarily close to a **limit** x. Formally, for any $\varepsilon > 0$, there exists some $N \in \mathbb{N}$ such that for all $n \geq N$,

$$|x_n - x| < \varepsilon$$

A sequence of random variables $\{X_n\}_{n=1}^{\infty}$ **converges in probability** to a random variable X if for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbf{P}(|X_n - X| \geq \varepsilon) = 0$$

In other words, in the limit the joint probability distribution of X_n and X gets concentrated arbitrarily tightly around the event $X_n = X$.

We can throw Markov's inequality together with convergence in probability to get the **weak law of large numbers**: let $\{X_n\}_{n=1}^{\infty}$ be a sequence of i.i.d. random variables with mean $\mu = \mathbf{E}X_i$ and finite variance σ^2. Define the "empirical mean" to be $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ (note that this is itself a random variable). Then for every $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbf{P}(|\bar{X}_n - \mu| \geq \varepsilon) = 0$$

Proof. This goes just like the derivation of Chebychev's inequality. We have

$$\begin{aligned}\mathbf{P}(|\bar{X}_n - \mathbf{E}X_i| \geq \varepsilon) &= \mathbf{P}((\bar{X}_n - \mu)^2 \geq \varepsilon^2) \\ &\leq \frac{\mathbf{E}(\bar{X}_n - \mu)^2}{\varepsilon^2} \\ &= \frac{\mathrm{Var}\bar{X}_n}{\varepsilon^2} \\ &= \frac{\sum_{i=1}^n \mathrm{Var}\frac{X_i}{n}}{\varepsilon^2} \\ &= \frac{\sigma^2}{n\varepsilon^2}\end{aligned}$$

where the second step follows from Markov's inequality. This goes to zero as $n \to \infty$. □

Now let's take another look at our definition of generalization:

$$\lim_{n\to\infty} \mathbf{P}\{|I_S[f_S] - I[f_S]| \geq \varepsilon\} = 0, \quad \forall\varepsilon$$

We are really saying that a learning algorithm that generalizes is one for which, as the number of training samples increases, the empirical loss *converges in probability* to the true loss, regardless of the underlying distribution of the data. Notice that this looks a lot like the weak law of large numbers. There's an important complication, though: even though we assume the training data (x_i, y_i) are i.i.d. samples from an unknown distribution, the random variables $V(f_S(x_i), y_i)$ are not i.i.d., because the function f_S depends on all of the training points simultaneously. We will talk about how to prove that learning algorithms generalize in class.