9.520 – Math Camp 2011

Probability Theory

Say we have some training data $S^{(n)}$, comprising n input points $\{x_i\}_{i=1}^n$ and the corresponding labels $\{y_i\}_{i=1}^n$:

$$S^{(n)} = \{(x_1, y_1), \dots, (x_n, y_n)\}$$

We want to design a learning algorithm that maps the training data $S^{(n)}$ into a function $f_S^{(n)}$ that will convert any new input x into a prediction $f_S^{(n)}(x)$ of the corresponding label y.

The ability of the learning algorithm to find a function that is predictive at points not in the training set is called *generalization*. There's a wrinkle, though, in that we aren't saying that the algorithm should find a function that predicts *well* at new points, but rather that the algorithm should consistently find a function that performs about as well on any new points as it does on the training set.

We formalize generalization by saying that, as the number n of points in the training set gets large, the error of our learned function (which can change with n) on the training set should converge to the expected error of that same learned function over all possible inputs. We'll denote the error of a function f on the training set by $I_S^{(n)}$:

$$I_S^{(n)}[f] = \frac{1}{n} \sum_i V(f(x_i), y_i)$$

V is the *loss function*, e.g. the squared error: $V(f(x_i), y_i) = (y_i - f(x_i))^2$. The expected error of f over the whole input space is I:

$$I[f] = \int V(f(x_i), y_i) d\mu(x_i, y_i)$$

where μ is the probability distribution (unknown to us!) from which the points (x_i, y_i) are drawn. Using this notation, the formal condition for generalization of a learning algorithm is:

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

for all $\varepsilon > 0$, where n is the number of training samples and $\mathbf{P}\{\cdot\}$ denotes the probability. So the probability of the training set error being different from the expected error should go to zero, as we increase the number of training samples.

Goal: We'll try here to make sense of this definition of generalization and show, in the most basic cases, how to prove statements like it.

First, some definitions.

A **random variable** X, for our purposes, is a variable that randomly assume a value in some range (assume this is $\mathbb{R}$) according to a **probability distribution**. X's probability distribution (a.k.a. probability measure) is a function that assigns probabilities to subsets of X's range, written $\mathbf{P}(A)$ where $A \subset \mathbb{R}$. (Worth repeating: $\mathbf{P}$ maps subsets of $\mathbb{R}$ to probabilities, rather than elements of $\mathbb{R}$ to probabilities.)

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

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

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

You can think of $d\mathbf{P}(x)$ analogously to the usual dx: dx is the area of an infinitesimal chunk of the domain of integration and $d\mathbf{P}(x)$ is the probability of an infinitesimal chunk of the domain of integration.

Now we'll get into the interesting stuff.

The problem: We want to prove things about the probability of $I_S^{(n)}[f_S]$ being close to $I[f_S^{(n)}]$. In what sense is there a probability distribution over the values of $I[f_S^{(n)}]$ and $I_S^{(n)}[f_S^{(n)}]$? It derives from the fact that the function $f_S^{(n)}$ depends on the training set (via the learning algorithm), and the training set is drawn from a probability distribution. The key challenge here is that *we don't know this underlying distribution of the datapoints* (x_i, y_i)! So the problem is to bound the probability of certain events (like $|I_S^{(n)}[f_S^{(n)}] - I[f_S^{(n)}]| \geq \varepsilon$), without knowing much about 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 (i.e. on certain sets).

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.

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

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 put 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:

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

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.