Selecting a Classifier in the Simple, Optimistic Case

Dany Haddad

August 3, 2020

Why does empirical risk minimization work? Here's a simple and optimistic case that is specific to binary classification:

Say we have a finite set of classifiers $\mathcal{C}$ that we're choosing from. Assume that one of those classifiers achieves an error probability of 0. To run empirical risk minimization, we'll choose a classifier based on its performance on a training set of size n; let's call this classifier ϕ_n. A pretty simple proof shows that the expected probability of error $\mathbb{E}[L(\phi_n)]$ is bounded above by $(1 + \log |\mathcal{C}|)/n$. So if the number of samples is far larger than the log of the number of classifiers we're choosing from, then the model we choose by running ERM will do well.

Of course, the assumptions are pretty strong, we can rarely (if ever) guarantee that if we're choosing from a finite set of classifiers that one of them will have an error probability of 0. This example sets up the intution for the more complex and typical cases where we are choosing from an infinite set of classifiers and cannot guarantee that one of them achieves an error probability of 0 (in that case we'll have to introduce ideas such as VC-dimension of a class of functions and uniform convergence).

Proof sketch

First we give a bound on the probability of the risk exceeding some fixed ϵ given n samples, then we integrate this bound to arrive at the bound on the expected risk. Let's call the true risk of a classifier $L(\phi)$ and the empirical risk using a dataset of n samples $L_n(\phi)$.

Since we know that one of our classifiers (call it ϕ^*) has a true risk $L(\phi^*) = 0$, at least one of the classifiers in our set will have an empirical risk of $L_n(\phi) = 0$. Therefor, the classifier we choose from our set, ϕ_n, must have an empirical risk of 0 (since that is obviously the minimum empirical risk). So, we just need to bound the probability that we choose a classifier with an empirical risk of 0 that actually has a true risk exceeding ϵ.

We'll bound this probability by the probability that ANY of the classifiers have $L_n(\phi) = 0$ but $L(\phi) > \epsilon$ (a union bound). Now the probability that a given classifier ϕ has $L_n(\phi) = 0$ but $L(\phi) > \epsilon$ is at most $(1 - \epsilon)^n$ since the set of points that ϕ makes mistakes on has probability at least ϵ, so the probability that we draw a sample of size n without seeing any of these points is bounded above by $(1 - \epsilon)^n$. So now putting the pieces together we have:

$$\mathbb{P}(L(\phi_n) > \epsilon) \leq |\mathcal{C}|(1 - \epsilon)^n$$

To bound $\mathbb{E}[L(\phi_n)]$ we just need a bound on the integral of $\mathbb{P}(L(\phi_n) > \epsilon)$.