Study Guide/ Summary for parts V-VII

Updated May 9, 2017

V. Stability, Smoothness, and Stochastic Gradient Descent

The first four parts of the course concentrated mainly on learnability results based on the mismatched minimization lemma, and uniform approximation of the general distribution P used to generate samples by the empirical distribution P_n. Part V begins with an example where consistency is proved without using uniform approximation. Specifically, it is for a setup with loss function such that $f \mapsto \ell(f, z)$ is strongly convex and Lipschitz continuous function defined on a closed convex set $\mathcal{F}$.

The key of the proof is that the ERM algorithm is stable with respect to replacing one sample. That stability property in this case can be traced to the strong convexity assumption.

The next section adopts specific definitions for stability and ability to generalize, and shows they are equivalent. If an algorithm has such properties and is asymptotically ERM it is consistent.

It is shown that stochastic gradient descent (SGD) is stable. In the final section, a result is given implying asymptotic ERM for SGD under fairly generation conditions. Together these results imply consistency of SGD under the technical assumptions.

For background in Section V, we considered most of the commonly used inequalities for σ-strongly convex functions and β smooth functions (such as, $\varphi(f') \geq \varphi(f) + \langle \nabla\varphi(f), f' - f\rangle + \frac{\sigma^2}{2}\|f - f'\|^2$ for a σ strongly convex function and $\varphi(f') \leq \varphi(f) + \langle \nabla\varphi(f), f' - f\rangle + \frac{\beta^2}{2}\|f - f'\|^2$ for a convex, β smooth function.

VI. More on SGD, Online Convex Function Minimization

The PAC learning model is adversarial to some degree. We'd like a given learning algorithm to be probably almost correct for any choice of probability distribution P that can be used to generate samples. In this section, following the work of Zinkevich, we consider an even more arbitrary model. The samples $z_1, z_2, \ldots$ can be selected arbitrarily. It is no longer possible to expect an algorithm to find classifiers that have loss nearly as small as what could be done with advance knowledge of the samples.

So instead the focus is on *regret* – the algorithm should do nearly as well as any single classifier that is the same for all samples. Two theorems were given, showing that sequential gradient descent has good maximum regret bounds. The regret grows no faster than $O(\sqrt{T})$ in case of convex functions and no faster than $O(\log T)$ in the case of strongly convex functions. In both cases, the regret per sample converges to zero.

It is pointed out that good regret bounds don't mean good performance in cases that there is strong correlation under consecutive samples. That is, if none of the constant classifiers perform well then it is not especially significant to do nearly as well as any constant classifier. In some such cases, SGD can significantly outperform any constant classifier.

The formalism was used to show a classical result about the classic perceptron algorithm, namely, for binary classification with a realizable model and strict separation margin, the number of incorrectly classified samples in a streaming setup is bounded.

We then saw that online learning algorithms naturally have a generalization ability. The idea is that each subsequent sample is fresh for the classifier applied to the sample. The generalization ability is a corollary of the Azuma-Hoeffding inequality. In case the loss function is convex in f, a convex combination of the classifiers found by a stochastic gradient algorithm gives a consistent estimator.

In problem set VI we saw a connection between the adversarial model and worst case distribution in the statistical learning framework.

Minimax lower bounds

Two quantitative versions of the no-free lunch theorem were proved in this section. The first gives a lower bound on the min max excess risk of any algorithm for a concept learning problem with a given VC dimension and for a class of probability distributions with a certain margin constraint. The second gives a stronger lower bound under stronger assumptions, namely, related to (N, D) richness. The proofs of the theorems flow from the assumptions. The supremum over all distributions in a certain class is replaced by an average over a finite collection of distributions, and the average excess loss for those distributions is related to bit error probability in a Bayesian inference problem. The inference problem involves a single distribution that is a mixture of other distributions. Then the bit error probability for that statistical inference problem is bounded below.

Two tools from statistical estimation theory were applied. First, a bound on the average error probability for binary hypothesis testing with Bernoulli observations. Secondly, information theoretic methods, involving mutual information. On one hand, for an estimator $\widehat{B}(X^n, Y^n)$ to estimate $\widehat{B}$ accurately, the mutual information $I(B, \widehat{B})$ must satisfy a lower bound. Since $I(B, \widehat{B}) = H(B) - H(B|\widehat{B})$ it amounts to bounding $H(B|\widehat{B})$ from above. On the other hand, by the data processing theorem, $I(B; \widehat{B}) \leq I(B; X^n, Y^n)$, and the mutual information $I(B; X^n, Y^n)$ is bounded by the fact there are only n observations and they are noisy. Specifically, he bound on mutual information based on a geometric property of KL divergence is used for this step,