CMSC35000-1 Introduction to Artificial Intelligence **Winter 2007**

Lecture 6: 1/25/2007

Lecturer: Partha Niyogi *Scribe: Peter Brune*

6.1 Generalization Error

We have described classes that can be built in which no classification errors are present on the training data; however this can be a bad idea. Define the class of functions $\mathcal{H} : X \to \{0,1\} = Y$. Assume our training data expressed as $z_i = (x_i, y_i)$ is independently, identically drawn. We can define the error rate and the empirical error rate as follows.

Definition 6.1 (Error Rate) $\forall h \in \mathcal{H},\ I(h) = \mathbb{E}[|y - h(x)|] = P_e(h)$

Definition 6.2 (Empirical Error Rate) $\forall h \in \mathcal{H},\ I_{emp}(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - h(x_i))$

The empirical error rate being directly calculable given the training data z.

6.1.1 Performance Inequalities

Recall the statement and proof of the minimum error classifier, $P(y = 1|x) > \frac{1}{2} \Rightarrow g(x) = 1$ and that $\forall f : X \to Y,\ I(g) \leq I(f)$. However we have minimized the risk over some training data, creating the classifier $\hat{h_n}$ as a random function depending on $z_1 ... z_n$. This implies that $I_{emp}(\hat{h_n}) \leq I_{emp}(f)\ \forall f \in \mathcal{H}$. We define the future performance of $\hat{h_n}$ as I($\hat{h_n}$), and hope that the gap converges as $n \to \infty$. Note that for a fixed $h,\ I_{emp}(h) \to I(h)$. We also see:

$$\forall \epsilon > 0,\ P(I_{emp}(h) - I(h) > \epsilon) \leq 2e^{-\epsilon^2 n}$$

By standard bounds. Suppose we fix $\delta > 0 :\ 2e^{-\epsilon^2 n} \leq \delta \forall n \geq \frac{1}{\epsilon^2} log(\frac{2}{\delta})$. This implies that $\forall n \geq log(\frac{2}{\delta}$, $P(|I_{emp}(h) - I(h)| > \epsilon) \leq \delta$.

Suppose we have a family of classifiers consisting solely of h; $\mathcal{H} = \{h\}$. This would imply that $I_{emp}(\hat{h_n}) = I_{emp}(h) \to I_{emp}(h) = I_{emp}(\hat{h_n})$, and tells us nothing about the case in which $|\mathcal{H}| > 1$. The ideal is for the empirical risk and the true risk to be closed $\forall h \in \mathcal{H}$. We want their minimums to correspond directly. This leads to the following uniform statement:

$$\forall \epsilon > 0,\ P[\sup_{h \in \mathcal{H}} |I_{emp}(h) - I(h)| > \epsilon] \to 0 : n \to \infty$$

The opposite statement may also have explanatory power. This statement leads from the uniform law of large numbers and applies if and only if $\mathcal{H}$ is of finite **VC dimension** (Vapnik and Chervonenkis - 1971).

$$P[\forall h\ |I_{emp}(h) - I(h)| \leq \epsilon] \to 1 : n \to \infty$$

A very different statement leading from the simple law of large numbers is:

$$\forall h,\ P[|I_{emp}(h) - I(h)| \leq \epsilon] \to 1 : n \to \infty$$

6.1.2 Convergence of $\hat{h_n} \to h^*$

So, in a nutshell, we want to find $h^* \in \mathcal{H}$, but we don't have access to $I(h)$, so we must make do with $I_{emp}(h)$, and the ideal classifier derived from minimizing it, $\hat{h_n}$. We would rather it be the case that $\hat{h_n} \to h^*$, but to do this we need uniform convergence. Otherwise there is no guarantee of a minimized error. However we can make the following statements regarding the various rates of error:

$$I_{emp}(\hat{h_n}) \leq I_{emp}(h^*) \leq_{prob(1-\delta)} I(h^*) + \epsilon$$

Additionally $I_{emp}(\hat{h_n}) \geq I(\hat{h_n}) - \epsilon$ with high probability. Together this leads to:

$$I(h^*) \leq I(\hat{h_n}) \leq I(h^*) + 2\epsilon$$

Next we investigate some consequences of the uniform law of large numbers for finite $|\mathcal{H}|$. Suppose $\mathcal{H} = \{h_1, ..., h_N\}$ and that A_i is the event that $|I_{emp}(h_i) - I(h_i)| > \epsilon$. We can see that $\forall i P(A_i) \leq 2e^{-\epsilon^2 n}$ by the simple law of large numbers.

$$P(\bigcup_i A_i) \leq \sum_{i=1}^{N} P(A_i) \leq \sum_{i=1}^{N} 2e^{-\epsilon^2 n} = 2Ne^{-\epsilon^2 n}$$

By the fact that $\bigcup_i A_i = \sup |I_{emp}(h_i) - I(h_i)| > \epsilon$, we get that

$$P[\sup_{h \in \mathcal{H}} |I_{emp}(h_i) - I(h_i)| > \epsilon] \leq 2Ne^{-\epsilon^2 n}$$

Combining this with the previous arguments gives us the following:

Theorem 6.3 *Fix $\delta > 0$, and $2Ne^{-\epsilon^2 n} \leq \delta$. $e^{-\epsilon^2 n} \geq (\frac{2N}{\delta})$. If we set $n \geq \frac{log(\frac{2N}{\delta})}{\epsilon^2}$ and draw n examples, $I(\hat{h_n}) \leq I(h^*) + 2\epsilon$.*

The proof comes directly from the previous statements. The meaning of this is that we now have quantified the necessary number of examples. However, another way to write this is as:

Corollary 6.4 *If we draw n examples then with probability $> (1 - \delta)$, $I(\hat{h_n}) \leq I(h^*) + \frac{1}{\sqrt{n}} log(\frac{2N}{\delta})$*

This brings us to a paradox of sorts. On one hand to improve $I(h^*)$, $|\mathcal{H}|$ must be large, however the second half of this term implies that we want to make δ large or $N = |\mathcal{H}|$ small.

6.2 No Free Lunch

The holy grail here would of course be an oracle. Ideally we could say that this oracle, g_p, is in $\mathcal{H}$, however there's no way that we can arrive at this conclusion from Learning Theory. We can say that $|I(g_p) - I(h^*)|$ would imply that a large $\mathcal{H}$ is needed, however $|I(h^*) - I(\hat{h_n})|$ implies that we want a small $\mathcal{H}$. In some respects this form of the theorem can be seen as closer to reality.

Another problem that can be considered is universal or nonparametric learning. Here we consider nested and increasing families of functions; say the polynomials of increasing degree. Call the polynomials $\mathcal{H} = \cup \mathcal{H}_i$. This is dense in $\mathcal{L}_2$. Suppose we choose $\mathcal{H}$ as a function of n, and we give a recipe from $n \rightarrow i(n)$, where i is monotonically increasing. What this says is that if one has n examples, one would like to consider $\mathcal{H}_{i(n)}$, or make $|\mathcal{H}| = N(n) \sim e^{\sqrt{n}}$ in order to make the difference in error between the classifier and the ideal disappear. In these cases we never do empirical risk minimization over all things at once. However, as $n \rightarrow \infty$, $I(h^*) \rightarrow I(g_p)$.

Consider $\mathcal{H}$ now as the polynomials with rational coefficients (a countably infinite class). It should be possible to discover the Bayes optimal classifier given enough data; however we have no guarantee involving the rate at which we approach it. This is a strong and useless statement.

Now consider $\mathcal{H}$ as the linear halfspaces over $\mathbb{R}^s$. Here are argument breaks down for technical reasons. We have one-parameter functions with infinite VC dimension, so learning is not possible given this class of functions.

6.2.1 VC Dimension

We see that a lot comes down to the complexity of $\mathcal{H}$, however $|\mathcal{H}|$ is a weak notion. The proper notion is that of VC dimension, which is often related to the number of free parameters. For example $w.x \geq 0$, $x \in \mathbb{R}^d$ has d free parameters. Functions such as the sign of $sin(a(x))$, however, have infinite VC dimension.

6.2.2 Other Issues

This all only really covers one half of the problem. The other part is computational. As always, we would like to say that the empirical minimum is discoverable. A procedure would most likely look at N samples to evaluate the empirical risk. This puts the problem into the realm of algorithmics, where the outlook becomes more bleak. Consider the family of neural networks; We can't prove anything about the closeness of the empirical minimum. For decision trees using linear halfspaces we have no guarantee of a polynomial time procedure.

An important philosophical point is that to deal with these complexities there are varying areas of emphasis. In economics one assumes that the agents are omniscient or have bounded rationality (They cannot compute the utility functions). These two approaches give quite different results. There is also the issue of one-shot learning, which humans appear to be capable of. Here a small, finite number of data points appear to be enough.