CS489 Winter07 Lectures 2 and 3 (the technical content)

Shai Ben-David

January 11, 2007

1 Preliminary Notation and Basic Observations

1.1 Learners Input

- X is the *Domain Set*
- L is the *Label Set*
  For our discussion, we will restrict L to be $\{0, 1\}$.
- The *Training data*, $S = ((x_1, l_1) \dots (x_m, l_m))$ is a finite sequence of pairs in $X \times L$.

1.2 Learners Output - "hypothesis"

- $h : X \to \{0, 1\}$ is the *hypothesis* or *prediction funtion.* (note: this is deterministic, for simplicity)

Our goal is to find some h with small error. Our next step is to define precisely what we mean by "small error"'.

1.2.1 Measure of hypothesis error

- Let P be a probability distribution (over $X \times L$). Namely, P assigns probability to pairs (x, l), where $x \in X$ and l is a label for the point x. Having such a P, one can measure how likely is h to make an error when labeled points are randomly drawn according to P:
  $E^P(h) = P(\{(x, l) : h(x) \neq l\})$.

When P is the data-generating distribution, $E^P(h)$ is called the *test error* or *true error* of h.

We would like to find a predictor, h, for which that error will be minimized. However, the learner does not know the data generating P. What the learner does have access to is the training data, S.

- $E^S(h) = \frac{|\{(x_i,l_i)\in S: h(x_i)\neq l_i\}|}{|S|}$ is called the *empirical error* or the *training error* of h.
  Given S, a learner can compute $E^S(h)$ for any function $h : X \to \{0,1\}$.
  Note that $E^S(h) = E^{P(\text{uniform over } S)}(h)$.

1.3 Basic Assumption

To have any chance of success we must make some assumptions about the relationship between S and P. We shall assume that the training data S is independent and identically distributed (iid) by P -the probability distribution that determines the true error. In other words, we assume that the data given to the learner for training, S, is generated by the same procedure that generates the data on which the learner's conclusion, h, will be evaluated.

1.4 Optimal Predictor

Given any probability distribution P over $X \times \{0,1\}$, the best label predicting function from X to $\{0,1\}$ will be

$$f(x) = \begin{cases} 1 & \text{if } P(y=1|x) \geq 1/2 \\ 0 & \text{otherwise} \end{cases}$$

Unfortunately, since we do not know P, we cannot utilize this optimal predictor f.

1.5 Law of Large Numbers (LLN)

Theorem 1.1 *Let $\{X_i\}$ be random variable that are identically and independently distributed (i.i.d) over the real line. Assume these random variables have a finite variance, and let μ denote their mean. Then, for every epsilon > 0,*

$$\lim_{m\to\infty} Pr\left(|\frac{1}{m}\Sigma_{i=1}^{m} X_i - \mu| > \epsilon\right) = 0$$

Now, given a probability distribution, P, over $X \times \{0,1\}$, we can define, for every predictor $h : X \to \{0,1\}$, a random variable, X_h^P by drawing a pair (x,l) according to P and setting

$$X^h(x,l) = \begin{cases} 1 & \text{if } h(x) \neq l \\ 0 & \text{if } h(x) = l \end{cases}$$

Note that, the mean of this random variable is just $E^P(h)$, and for a random sample S of size m, drawn i.i.d. by P, we get that $E^S(h) = \frac{1}{m}\Sigma_{(x,l)\in S}X^h(x,l)$. Applying the above Law of Large numbers to these variables, we get:

Theorem 1.2 *For any function $h : X \to \{0,1\}$, for every probability distribution P, over $X \times \{0,1\}$, if S are i.i.d. random samples drawn according to P,*

$$\lim_{|S|\to\infty} Pr\left(|E^S(h) - E^P(h)| > \epsilon\right) = 0$$

In other words, if one fixes a hypothesis h, then for every data generating distribution, the empirical error of h will converge to its true error, as sample sizes grow to infinity. This is a good start, however, it has two significant drawbacks that make it quite useless for our purposes. First, this is only an asymptotic result. It provides no information about the gap between the empirically estimated error and its =true value for any given, finite, sample size. The second issue with this result is that it holds only if h is chosen independently of S. This is not the case we are interested in - we would like to analyze a scenario in which h is chosen as a result of viewing the training sample, S. In the following, we address both these issues.

We shall impose restrictions on the possible h's the learner (or learning algorithm) can choose from and develop results of the form:

$\forall \epsilon, \delta$ there exists an m such that for all P and h (subject to some restrictions)

$$P_S(|E^S(h) - E^P(h)| > \epsilon) < \delta$$

We will call ϵ the measure of *accuracy* and δ the measure of *confidence*

2 First Statistical Learning Bounds - Finite Choice for h

Let H be a finite set of prediction functions. Given a sample S, we will bound the probability that there exists an $h \in H$ that looks perfect on the

training data but has true error above ϵ.

We will start by taking any fixed h and bounding the probability (over the samples) that $E^S(h) = 0$ assuming that $E^P(h) > \epsilon$.

$$P(E^S(h) = 0) \leq (1 - \epsilon)^m < e^{-m\epsilon}$$

Note, that this probability is taken over the i.i.d. random samples of size m.

2.1 The fortunate case- we find a perfect looking predictor

Next, we apply the union bound to establish to bound the probability that for *some* $h \in H$, $E^S(h) = 0$ in spite of having $E^P(h) > \epsilon$.

$$P[\exists h \in H(E^S(h) = 0 \text{ and } E^P(h) > \epsilon)] < |H|e^{-m\epsilon}$$

Therefore, if δ is such that $\delta \geq |H|e^{-m\epsilon}$, or, equivalently,

$$\ln(\delta) \geq \ln|H| - m\epsilon$$

or if

$$m \geq \frac{\ln|H| + \ln(1/\delta)}{\epsilon}$$

we get

$$P[\exists h \in H(E^S(h) = 0 \text{ and } E^P(h) > \epsilon)] < \delta$$

2.2 The general bound for finite H

Next we address the case where H is still finite but one does not assume that $E^S(h) = 0$. We will therefore have to apply a slightly stronger argument.

Chernoff bound

Let X_i be independent binary valued random variable. Let the expectation of each X_i be some fixed μ. We view empirical values of the X_i's and wish to estimate μ.

$$\forall\ \epsilon\ P(\mu - \frac{1}{m}\sum_{i=1}^{m} X_i > \epsilon) < e^{-2m\epsilon^2}$$

For a sample $S = ((x_i y_i) \ldots (x_m y_m))$, let $X_i = \begin{cases} 1 & \text{if } h(x_i) \neq y_i \\ 0 & \text{if } h(x_i) = y_i \end{cases}$

Note that under this formulation $\mu = E^P(h)$ *and* $\frac{1}{m}\sum_{i=1}^{m} X_i = E^S(h)$.

Applying Chernoff bound, for any fixed h, we have:

$$P[E^P(h) > E^S(h) + \epsilon] \leq e^{-2m\epsilon^2}$$

It follows that the probability that for some $h \in H$;

$$P(E^P(h) - E^S(h) > \epsilon) < |H|e^{-2m\epsilon^2}$$

Thus, repeating the same argumentation we had for the previous bound (for the case $E^S(h) = 0$), we get, $\forall\ H$ (finite), $\forall\ m\ \forall\ \delta$, for every probability distribution P, with probability $> (1 - \delta)$,

$$\forall\ h \in H,\ \ E^P(h) \leq E^S(h) + \sqrt{\frac{\ln|H| + \ln\frac{1}{\delta}}{2m}}$$

2.3 Empirical Risk Minimization

The above discussion motivates the following learning paradigm, which is called *Empirical Risk Minimization (ERM)*:

1. Fix a finite set H of predictors (that is, functions from X to $\{0, 1\}$).
2. Upon viewing the training sample, S, find an h^{ERM} in H that minimizes $E^S(h)$ (over all h's in H).
3. Use that h^{ERM} to predict the labels of test points.

For such a learning paradigm we can now derive a *relative* error bound. Namely, we can bound by how much may our chosen predictor h^{ERM} be worse than the best possible predictor in H.

Theorem 2.1 *For every finite set of functions,* H*, and any probability distribution* P *(over* $X \times \{0, 1\}$*), let* h^* *be an element of* H *that minimizes the true error,* $E^P(h)$*. For every* $\epsilon, \delta > 0$*, let* S *is an i.i.d sample randomly generated according to* P*, so that* $|S| \geq \frac{\ln(|H|)+\ln(1/\delta)}{2\epsilon^2}$*. Let* h^{ERM} *is a minimizer of* $E^S(h)$ *over the functions* $h \in H$*, then, with probability exceeding* $(1 - \delta)$*,*

$$E^P(h^{ERM}) \leq E^P(h^*) + 2\epsilon$$