6.858 Machine Learning	Lecture 9 : October 12, 1994
Lecturer: Ron Rivest	*Scribe: Vladimir Roussakov*

9.1 Outline

- Estimating Error Rates
- Uniform Convergence and VC-dimension

The major result we have seen so far is the relationship between VC-dimension and PAC-learning. It turns out that *finite* VC-dimension enables PAC- learnability. Finite VC-dimension is used in other areas, such as computational geometry.

9.2 Estimating Error Rates

9.2.1 Introduction

We are moving away from PAC learning to learning in a more general sense. The objective is to find a relationship between *observed* error rate for a sample and the *true* error rate. We hope that picking a concept c with a small observed error rate gives us small true error rate.

Let X = domain,
C = concept class on X,
$$c(x) = \begin{cases} 0 - \text{negative example,} \\ 1 - \text{positive example} \end{cases}$$
D = some unknown distribution on X from which we draw labelled examples.

As in the PAC learning framework we attempt to learn the target concept c_*. However, we do not guarantee that c belongs to C and we may not be able to come up with a perfect target concept. Our goal is to minimize the error rate.

Notation:
Define $\tilde{c} = c \oplus c_*$ as the error region for c;
$\tilde{c} = 1$ if and only if c makes a mistake on x: $\tilde{c}(x) = \begin{cases} 0, c(x) = c_*(x) \\ 1, c(x) \neq c_*(x) \end{cases}$

Then $D\tilde{c}$ = true error rate of c with respect to D, and we are looking for a concept c with a small error rate.

Let $S = \{x_1, x_2, ..., x_m\}$ be a random sample of size m drawn according to D. Define D_m to be a uniform distribution on S. That is, each point in S has weight $\frac{1}{m}$ and the weight is zero outside S.

Then $D_m\tilde{c}$ = observed (empirical) error rate of concept c on sample S. $D_m\tilde{c} = \frac{\#errors}{m}$. Note that $D_m\tilde{c}$ is a random variable which depends on sample S.

Learning algorithms are often in the situation of estimating true error rates from empirical error rates, in order to attempt to find a hypothesis c with low true error rate. The questions that need to be answered are the following:

- Why should approximate minimization of observed error rate yield a concept c that approximately minimizes true error rate?
- How big a sample do we need?

9.3 Simple Error Rate Estimation (for a single concept)

For a fixed concept c, every example is classified as positive or negative. Let $p = D\tilde{c}$ be the true error rate. This case is equivalent to flipping a biased coin a certain number of times. We are trying to estimate the bias p, (i.e., probability of heads of the coin).

9.3.1 Review of the Law of Large Numbers

Theorem: (Strong Law of Large Numbers)

With probability 1, $lim_{m\rightarrow\infty} p_m = p$

In other words, the observed frequency of heads converges to the true frequency of heads. Here $p_m = D_m \tilde{c}$ is the observed error rate. This theorem does not tell us what happens for a particular number of samples. It is useless for estimating the sample size. For practical purposes we need the *rate* of convergence.

Note that expected error rate $E(p_m) = p$,
variance $Var(p_m) = \frac{pq}{m} \leq 1/4m, (q = (1-p))$,
standard deviation $\sigma_m = \sigma(p_m) = \sqrt{pq/m} \leq 1/2\sqrt{m}$

This is a characteristic result for estimating p with p_m - the quadratic dependency between σ_m and the sample size.

9.3.2 Review of the Law of Iterated Logarithm

Theorem: (Law of Iterated Logarithm)

$limsup_{m \to \infty} \frac{p_m - p}{\sigma_m \sqrt{2 \ln \ln m}} = 1$, this is the upper limit

$limsup_{m \to \infty} \frac{p - p_m}{\sigma_m \sqrt{2 \ln \ln m}} = 1$, this is the lower limit

In other words we expect p_m to stay close to p. Asymptotically p_m is never more than $\sqrt{2 \ln \ln m}$ standard deviations away, or at most $\sqrt{\frac{\ln \ln m}{2m}}$ away. This is still an asymptotic result, and we are more interested in finite bounds for m.

9.3.3 Review of Chernoff Bounds

Notation:

$$GE(p, m, r) = Prob[p_m \geq r]$$
$$LE(p, m, r) = Prob[p_m \leq r]$$

That is, this is useful when we are trying to relate the empirical bias to the true bias.

Theorem: [Hoeffding] (Additive form of Chernoff Bounds)

$$GE(p, m, p + \epsilon) \leq e^{-2m\epsilon^2}$$
$$LE(p, m, p - \epsilon) \leq e^{-2m\epsilon^2}$$

This result has a practical use. We take a coin with true bias p, flip it m times, and now we can bound the probability p_m differs too much from p.

Corollary:

We have error $|p_m - p| \leq \epsilon$ with probability $\geq 1 - \delta$ if

$$m \geq \frac{1}{2\epsilon^2} ln(\frac{2}{\delta})$$

Note: The sample size grows quadratically with $1/\epsilon$. Recall that in PAC learning sample sizes grew only linearly with $1/\epsilon$. The discrepancy is due to the fact that

$\sigma_m = \Theta(1/\sqrt{m})$, if $p = 1/2$, but $\sigma_m = \Theta(1/m)$, if $p \leq 1/m$ (or $p \to 0$) .

The region we are working with in PAC-learning is small, and we are dealing with small probabilities and small deviations. Here we are near the middle of the region and the deviation is a lot larger. In practice, we should use the fact that p is small and stay within a linear dependency on $1/\epsilon$.

This problem can be fixed using relative bounds rather than absolute bounds.

Theorem: [Angluin and Valiant] (Multiplicative form of Chernoff Bounds)

$$GE(p, m, (1+\alpha)p) \leq e^{mp\alpha^2/3}$$

$$LE(p, m, (1-\alpha)p) \leq e^{mp\alpha^2/2}$$

This bounds the probability that p_m is off by a multiplicative factor from p. As before the dependency is exponential. In addition the bound also depends on p.

Note: If $p \to 0$, (i.e., the true probability is small), then the required sample size to get a good estimate for p_m grows.

Corollary:
If we keep p from being too small, i.e., $p \geq \epsilon$, then the number of points that suffice to achieve $(1-\alpha) \leq \frac{p_m}{p} \leq (1+\alpha)$ with probability $\geq (1-\delta)$ is

$$m = \frac{2}{\epsilon\alpha^2} ln(\frac{2}{\delta})$$

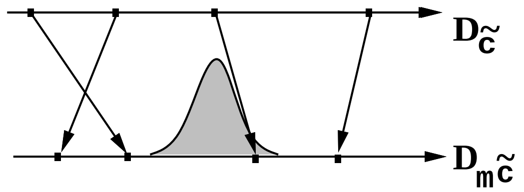

Figure 9.1: True and estimated error rates in case of multiple concepts

9.4 Uniform Convergence. Estimating Error in Case of Many Concepts

Now the objective is to estimate many biased coins simultaneously. The question is whether we can estimate all $D_m\tilde{c}$ simultaneously.

Example:

$X = [0, 1]$ is a unit interval
$D =$ uniform distribution
$C =$ set of all concepts on [0, 1]

When many concepts converge simultaneously the estimate for one concept does not give the estimate for all (see figure 9.1). For any sample S there exists concept c whose observed error rate $D_m\tilde{c} = 0$, and the true error rate $D\tilde{c} = 1$ for any m. In other words concept c agrees with the target everywhere on S and does not agree anywhere outside S for any m. Such a concept class is too rich. The VC-dimension of this concept class is infinite.

We are interested in a bound on $sup_{c\in C}|D_m\tilde{c} - D\tilde{c}|$, so that even for the worst-case $c \in C$ the observed error rate is close to the true error rate. If $sup_{c\in C}|D_m\tilde{c} - D\tilde{c}| \to 0$ as $m \to \infty$ we have *uniform convergence* of observed error rates to true error rates.

The case of finite concept class C is easy, the result is given by the additive form of Chernoff bounds:

$$2|C|e^{2m\epsilon^2} \leq \delta$$

So,

$$m \geq \frac{1}{2\epsilon^2}(ln|C| + ln(\frac{2}{\delta}))$$

and the observed error rate converges.

What if C is infinite? We want to use finite VC-dimension (assuming C has finite VC-dimension) to replace $|C|$ and prove uniform convergence.

Theorem: [Vapnik and Chervonenkis; Improved by Devroye, J. *Multivariate Analysis* 12, 1 (1982), 72-79]

If class C has VC-dimension d, then empirical error rates converge uniformly to true error rates:

$$Pr\{sup_{c\in C}|D_m\tilde{c} - D\tilde{c}| \geq \epsilon\} \leq 4e^{(4\epsilon+4\epsilon^2)}\Pi_c(m^2)e^{-2m\epsilon^2}$$

Qualitatively, the proof is similar to the proof of the VC-dimension theorem for PAC learning. Recall that

$$|\Pi_c(m)| \leq \left(\frac{em}{d}\right)^d$$

That is, it grows polynomially with m. Similar analysis yields that

$$m \geq \Omega(\frac{d}{\epsilon^2}log\frac{1}{\epsilon} + \frac{1}{\epsilon^2}log\frac{1}{\delta})$$

This gives the lower bound of the number of examples that suffice to ensure that, with probability $\geq (1-\delta)$ all empirical error rates are within ϵ of their true error rates.

Note: Once again there is a quadratic dependency on $1/\epsilon$, if we ignore the logarithm factors.

The uniform convergence is a powerful notion, and it gives a good learning algorithm. Consider the following procedure.

Procedure:

- Draw a sample of size $m \geq \Omega(\frac{d}{\epsilon^2}log\frac{1}{\epsilon} + \frac{1}{\epsilon^2}log\frac{1}{\delta})$
- Return concept c' with the smallest empirical $D_m\tilde{c}$.

Theorem:
Suppose we choose a sample big enough so that

$$(\forall c)|D_m\tilde{c} - D\tilde{c}| \leq \epsilon$$

then the distance between the error rate of the target c_* and c' is

$$|D_m\tilde{c_*} - D\tilde{c'}| \leq 2\epsilon$$

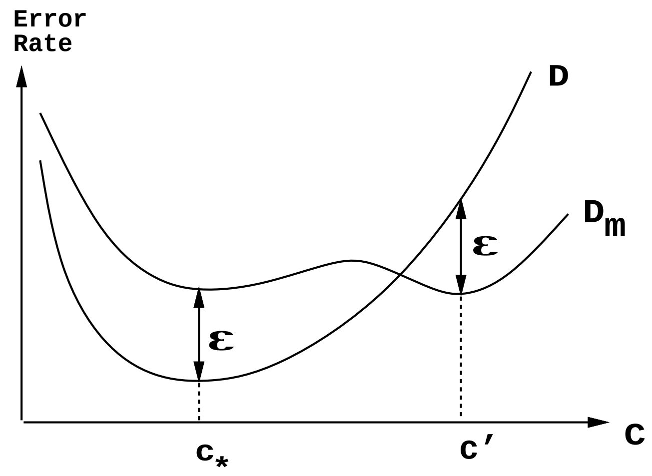

Figure 9.2: Error rates for c_* and c'. D is the true error rate, $D = m$ is the observed error rate.

Here we assume that c_* belongs to the concept class C, so the condition of the theorem holds for c_*.

Proof:
For the target concept c_* the distance between observed and true error rates is

$$|D_m \tilde{c_*} - D\tilde{c_*}| \le \epsilon$$

Because c' has the minimal observed error rate

$$D_m \tilde{c'} \le D_m \tilde{c_*}$$

Consequently (see figure 9.2),

$$|D_m \tilde{c'} - D\tilde{c'}| \le \epsilon$$

■

This result means that in practice we just need to minimize on $D_m \tilde{c}$.

Note: We assume that we have infinite computation power and ignore how hard it is to compute c'. In some cases it is better not to find c', but only to approximate it.

References

[1] Devroye, J. *Multivariate Analysis* 12, 1:72-79 (1982), .

[2] Vapnik, V.N. and Chervonenkis, A.Y. On the uniform covergence of relative frequences of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264-280 (1971).