10-601 Machine Learning

Maria-Florina Balcan Spring 2015

Generalization Abilities: Sample Complexity Results.

The ability to *generalize* beyond what we have seen in the training phase is the *essence* of machine learning, essentially what makes machine learning, machine learning. In these notes we describe some basic concepts and the classic formalization that allows us to talk about these important concepts in a precise way.

Distributional Learning

The basic idea of the distributional learning setting is to assume that examples are being provided from a fixed (but perhaps unknown) distribution over the instance space. The assumption of a fixed distribution gives us hope that what we learn based on some training data will carry over to new test data we haven't seen yet. A nice feature of this assumption is that it provides us a well-defined notion of the error of a hypothesis with respect to target concept.

Specifically, in the distributional learning setting (captured by the PAC model of Valiant and Statistical Learning Theory framework of Vapnik) we assume that the input to the learning algorithm is a set of labeled examples

S : $\quad (x_1, y_1), \ldots, (x_m, y_m)$

where x_i are drawn i.i.d. from some fixed but unknown distribution D over the the instance space X and that they are labeled by some target concept c^*. So $y_i = c^*(x_i)$. Here the goal is to do optimization over the given sample S in order to find a hypothesis $h : X \to \{0, 1\}$, that has small error over whole distribution D. The true error of h with respect to a target concept c^* and the underlying distribution D is defined as

$$err(h) = \Pr_{x \sim D}(h(x) \neq c^*(x)).$$

($\Pr_{x \sim \mathcal{D}}(A)$ means the probability of event A given that x is selected according to distribution $\mathcal{D}$.)

We denote by

$$err_S(h) = \Pr_{x \sim S}(h(x) \neq c^*(x)) = \frac{1}{m} \sum_{i=1}^{m} I[h(x_i) \neq c^*(x_i)]$$

the empirical error of h over the sample S (that is the fraction of examples in S misclassified by h).

What kind of guarantee could we hope to make?

- We converge quickly to the target concept (or equivalent). But, what if our distribution places low weight on some part of X?

- We converge quickly to an approximation of the target concept. But, what if the examples we see don't correctly reflect the distribution?
- With high probability we converge to an approximation of the target concept. This is the idea of **P**robably **A**pproximately **C**orrect learning.

Distributional Learning. Realizable case

Here is a basic result that is meaningful in the realizable case (when the target function belongs to an a-priori known finite hypothesis space H.)

Theorem 1 *Let H be a finite hypothesis space. Let D be an arbitrary, fixed unknown probability distribution over X and let c^* be an arbitrary unknown target function. For any ϵ, $\delta > 0$, if we draw a sample from D of size*

$$m = \frac{1}{\epsilon}\left[\ln(|H|) + \ln\left(\frac{1}{\delta}\right)\right],$$

then with probability at least $1-\delta$, all hypotheses/concepts in H with error $\geq \epsilon$ are inconsistent with the data (or alternatively, with probability at least $1-\delta$ any hypothesis consistent with the data will have error at most ϵ.)

Proof: The proof involves the following steps:

1. Consider some specific "bad" hypothesis h whose error is at least ϵ. The probability that this bad hypothesis h is consistent with m examples drawn from $\mathcal{D}$ is at most $(1-\epsilon)^m$.
2. Notice that there are (only) at most $|H|$ possible bad hypotheses.
3. (1) and (2) imply that given m examples drawn from $\mathcal{D}$, the probability *there exists* a bad hypothesis consistent with all of them is at most $|H|(1-\epsilon)^m$. Suppose that m is sufficiently large so that this quantity is at most δ. That means that with probability $(1-\delta)$ there are *no* consistent hypothesis whose error is more than ϵ.
4. The final step is to calculate the value m needed to satisfy

$$|H|(1-\epsilon)^m \leq \delta. \tag{1}$$

   Using the inequality $1 - x \leq e^{-x}$, it is simple to verify that (1) is true as long as:

$$m \geq \frac{1}{\epsilon}\left[\ln(|H|) + \ln\left(\frac{1}{\delta}\right)\right].$$

■

Note: Another way to write the bound in Theorem 1 is as follows:

For any $\delta > 0$, if we draw a sample from D of size m then with probability at least $1 - \delta$, any hypothesis in H consistent with the data will have error at most

$$\frac{1}{m}\left[\ln(|H|) + \ln\left(\frac{1}{\delta}\right)\right].$$

This is the more "statistical learning theory style" way of writing the same bound.

Distributional Learning. The Non-realizable case

In the general case, the target function might not be in the class of functions we consider. Formally, in the non-realizable or agnostic passive supervised learning setting, we assume assume that the input to a learning algorithm is a set S of labeled examples $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$. We assume that these examples are drawn i.i.d. from some fixed but unknown distribution D over the the instance space X and that they are labeled by some target concept c^*. So $y_i = c^*(x_i)$. The goal is just as in the realizable case to do optimization over the given sample S in order to find a hypothesis $h : X \to \{0, 1\}$ of small error over whole distribution D. Our goal is to *compete* with the best function (the function of smallest true error rate) in some concept class H.

A natural hope is that picking a concept c with a small observed error rate gives us small true error rate. It is therefore useful to find a relationship between *observed* error rate for a sample and the *true* error rate.

Concentration Inequalities. Hoeffding Bound

Consider a hypothesis with true error rate p (or a coin of bias p) observed on m examples (the coin is flipped m times). Let S be the number of observed errors (the number of heads seen) so S/m is the observed error rate.

Hoeffding bounds state that for any $\epsilon \in [0, 1]$,

1. $\Pr[\frac{S}{m} > p + \epsilon] \leq e^{-2m\epsilon^2}$, and
2. $\Pr[\frac{S}{m} < p - \epsilon] \leq e^{-2m\epsilon^2}$.

Simple sample complexity results for finite hypotheses spaces

We can use the Hoeffding bounds to show the following:

Theorem 2 *Let H be a finite hypothesis space. Let D be an arbitrary, fixed unknown probability distribution over X and let c^* be an arbitrary unknown target function. For any ϵ, $\delta > 0$, if we draw a sample S from D of size*

$$m \geq \frac{1}{2\epsilon^2}\left(\ln(2|H|) + \ln\left(\frac{1}{\delta}\right)\right),$$

then probability at least $(1 - \delta)$, all hypotheses h in H have

$$|err(h) - err_S(h)| \leq \epsilon. \tag{2}$$

Proof: Let us fix a hypothesis h. By Hoeffding, we get that the probability that its observed error within ϵ of its true error is at most $2e^{-2m\epsilon^2} \leq \delta/|H|$. By union bound over all all h in H, we then get the desired result. ■

Note: A statement of type one is called a *uniform convergence* result. It implies that the hypothesis that minimizes the empirical error rate will be very close in generalization error to the best hypothesis in the class. In particular if $\widehat{h} = argmin_{h \in H} err_S(h)$ we have $err(\widehat{h}) \leq err(h^*) + 2\epsilon$, where h^* is a hypothesis of smallest true error rate.

Note: The sample size grows quadratically with $1/\epsilon$. Recall that the learning sample size in the realizable (PAC) case grew only linearly with $1/\epsilon$.

Note: Another way to write the bound in Theorem 2 is as follows:

For any $\delta > 0$, if we draw a sample from D of size m then with probability at least $1 - \delta$, all hypotheses h in H have

$$err(h) \leq err_S(h) + \sqrt{\frac{\ln(2|H|) + \ln\left(\frac{1}{\delta}\right)}{2m}}$$

This is the more "statistical learning theory style" way of writing the same bound.

Sample complexity results for infinite hypothesis spaces

In the case where H is not finite, we will replace $|H|$ with other measures of complexity of H (shattering coefficient, VC-dimension, Rademacher complexity).

Shattering, VC dimension

Let H be a concept class over an instance space X, i.e. a set of functions functions from X to $\{0, 1\}$ (where both H and X may be infinite). For any $S \subseteq X$, let's denote by $H(S)$ the set of all behaviors or dichotomies on S that are induced or realized by H, i.e. if $S = \{x_1, \cdots, x_m\}$, then $H(S) \subseteq \{0, 1\}^m$ and

$$H(S) = \{(c(x_1), \cdots, c(x_m)); c \in \mathcal{H}\}.$$

Also, for any natural number m, we consider $H[m]$ to be the maximum number of ways to split m points using concepts in H, that is

$$H[m] = \max\{|H(S)|; |S| = m, S \subseteq X\}.$$

To instantiate this, to get a feel of what this result means imagine that H is the class of thresholds on the line, then $H[m] = m + 1$, or that H is the class of intervals, then $H[m] = O(m^2)$, or for linear separators in R^d, $H[m] = m^{d+1}$.

Definition 1 *If $|H(S)| = 2^{|S|}$ then S is* **shattered** *by H.*

Definition 2 *The* **Vapnik-Chervonenkis dimension** *of H, denoted as $VCdim(H)$, is the cardinality of the largest set S shattered by H. If arbitrarily large finite sets can be shattered by H, then $VCdim(H) = \infty$.*

Note 1 *In order to show that the VC dimension of a class is at least d we must simply find some shattered set of size d. In order to show that the VC dimension is at most d we must show that no set of size $d+1$ is shattered.*

Examples

1. Let H be the concept class of thresholds on the real number line. Clearly samples of size 1 can be shattered by this class. However, no sample of size 2 can be shattered since it is impossible to choose threshold such that x_1 is labeled positive and x_2 is labeled negative for $x_1 \leq x_2$. Hence the $VCdim(H) = 1$.

2. Let H be the concept class intervals on the real line. Here a sample of size 2 is shattered, but no sample of size 3 is shattered, since no concept can satisfy a sample whose middle point is negative and outer points are positive. Hence, $VCdim(H) = 2$.

3. Let H be the concept class of k non-intersecting intervals on the real line. A sample of size $2k$ shatters (just treat each pair of points as a separate case of example 2) but no sample of size $2k+1$ shatters, since if the sample points are alternated positive/negative, starting with a positive point, the positive points can't be covered by only k intervals. Hence $VCdim(H) = 2k$.

4. Let H the class of linear separators in $\mathbf{R^2}$. Three points can be shattered, but four cannot; hence $VCdim(H) = 3$. To see why four points can never be shattered, consider two cases. The trivial case is when one point can be placed within a triangle formed by the other three; then if the middle point is positive and the others are negative, no half space can contain only the positive points. If however the points cannot be arranged in that pattern, then label two points diagonally across from each other as positive, and the other two as negative In general, one can show that the VCdimension of the class of linear separators in $\mathbf{R^n}$ is $n+1$.

5. The class of axis-aligned rectangles in the plane has $VC_{DIM} = 4$. The trick here is to note that for any collection of five points, at least one of them must be interior to or on the boundary of any rectangle bounded by the other four; hence if the bounding points are positive, the interior point cannot be made negative.

Sauer's Lemma

Lemma 1 *If $d = VCdim(H)$, then for all m, $H[m] \leq \Phi_d(m)$, where $\Phi_d(m) = \sum_{i=0}^{d} \binom{m}{i}$.*
For $m > d$ we have:

$$\Phi_d(m) \leq \left(\frac{em}{d}\right)^d.$$

Note that for H the class of intervals we achieve $H[m] = \Phi_d(m)$, where $d = VCdim(H)$, so the bound in the Sauer's lemma is tight.

Sample Complexity Results based on Shattering and VCdim

Interestingly, we can roughly replace $\ln(|H|)$ from the case where H is finite with the shattering coefficient $H[2m]$ when H is infinite. Specifically:

Theorem 3 *Let H be an arbitrary hypothesis space. Let D be an arbitrary, fixed unknown probability distribution over X and let c^* be an arbitrary unknown target function. For any ϵ, $\delta > 0$, if we draw a sample S from D of size*

$$m > \frac{2}{\epsilon} \cdot \left[\log_2 \left(2 \cdot H[2m]\right) + \log_2 \left(\frac{1}{\delta}\right)\right] \tag{3}$$

then with probability $(1 - \delta)$, all bad hypothesis in H (with error $> \epsilon$ with respect to c and D) are inconsistent with the data.

Theorem 4 *Let H be an arbitrary hypothesis space. Let D be an arbitrary, fixed unknown probability distribution over X and let c^* be an arbitrary unknown target function. For any ϵ, $\delta > 0$, if we draw a sample S from D of size*

$$m > (8/\epsilon^2)[\ln(2H[2m]) + \ln(1/\delta)]$$

then with probability $1 - \delta$, all h in H have

$$|err_D(h) - err_S(h)| < \epsilon.$$

We can now use Sauer's lemma to get a nice closed form expression on sample complexity (an upper bound on the number of samples needed to learn concepts from the class) based on the VC-dimension of a concept class. The following is the VC dimension based sample complexity bound for the realizable case:

Theorem 5 *Let H be an arbitrary hypothesis space of VC-dimension d. Let D be an arbitrary unknown probability distribution over the instance space and let c^* be an arbitrary unknown target function. For any ϵ, $\delta > 0$, if we draw a sample S from D of size m satisfying*

$$m \geq \frac{8}{\epsilon}\left[d \ln\left(\frac{16}{\epsilon}\right) + \ln\left(\frac{2}{\delta}\right)\right].$$

then with probability at least $1 - \delta$, all the hypotheses in H with $err_D(h) > \epsilon$ are inconsistent with the data, i.e., $err_S(h) \neq 0$.

So it is possible to learn a class C of VC-dimension d with parameters δ and ϵ given that the number of samples m is at least $m \geq c\left(\frac{d}{\epsilon}\log\frac{1}{\epsilon} + \frac{1}{\epsilon}\log\frac{1}{\delta}\right)$ where c is a fixed constant. So, as long as $VCdim(H)$ is finite, it is possible to learn concepts from H even though H might be infinite!

One can also show that this sample complexity result is tight within a factor of $O(\log(1/\epsilon))$. Here is a simplified version of the lower bound:

Theorem 6 *Any algorithm for learning a concept class of VC dimension d with parameters ϵ and $\delta \leq 1/15$ must use more than $(d-1)/(64\epsilon)$ examples in the worst case.*

The following is the VC dimension based sample complexity bound for the non-realizable case:

Theorem 7 *Let H be an arbitrary hypothesis space of VC-dimension d. Let D be an arbitrary, fixed unknown probability distribution over X and let c^* be an arbitrary unknown target function. For any ϵ, $\delta > 0$, if we draw a sample S from D of size*

$$m = O\left(\frac{1}{\epsilon^2}\left(d + \ln\left(\frac{1}{\delta}\right)\right)\right),$$

then probability at least $(1-\delta)$, all hypotheses h in H have

$$|err(h) - err_S(h)| \leq \epsilon. \tag{4}$$

Note: As in the finite case, we can rewrite the bounds in Theorems 5 and 7 in the "statistical learning theory style" as follows:

Let H be an arbitrary hypothesis space of VC-dimension d. For any $\delta > 0$, if we draw a sample from D of size m then with probability at least $1-\delta$, any hypothesis in H consistent with the data will have error at most

$$O\left(\frac{1}{m}\left[d\ln(m/d) + \ln\left(\frac{1}{\delta}\right)\right]\right).$$

For any $\delta > 0$ if we draw a sample from D of size m then with probability at least $1-\delta$, all hypotheses h in H have

$$err(h) \leq err_S(h) + O\left(\sqrt{\frac{d + \ln(1/\delta)}{m}}\right).$$

We can see from these bounds that the gap between true error and empirical error in the realizable case is $O(\ln(m)/m)$, whereas in the general (non-realizable) case this is (larger) $O(1/\sqrt{m})$.