PAC Learning

September 29, 2015

1 PAC learning

Notation:

1. X the set of examples (training and test);
2. $\{-1, +1\}$ is the set of labels (binary classification);
3. $X \times \{-1, +1\}$ the set of input/output pairings;
4. P (unknown) probability distribution on X such that $x \in X$ is distributed according to P (we say that the examples are iid - identically independent distributed according to P which means that "the probability of selecting an example (x, y), with $y \in \{-1, +1\}$ is the same for any (x, y) and is independent on the selection of any other such pair").

Let $f_\alpha(x)$ denote a family of functions (that is, a learning machine) and let $err_P(f_\alpha)$ denote the following quantity:

$$err_P(f_\alpha) = P(\{(x, y);\ f_\alpha(x) \neq y\}) = \int I_{\{(x,y);\ f_\alpha(x)\neq y\}} dP(x, y) = E_P(I_{\{(x,y);\ f_\alpha(x)\neq y\}}) \tag{1}$$

In equation (1) I_A denotes the indicator function (event) defined as $I_A(x) = 1$ if $x \in A$, 0 otherwise.

We show now that $err_P(f_\alpha)$ as is the same as $R(\alpha)$ defined as

$$R(\alpha) = \int \frac{1}{2}|f_\alpha(x) - y| dP(x, y)$$

Indeed,

$$\begin{aligned} R(\alpha) &= \int \frac{1}{2}|f_\alpha(x) - y| dP(x, y) \\ &= \int_{\{(x,y);\ f_\alpha(x)\neq y\}} \frac{1}{2}|f_\alpha(x) - y| dP(x, y) + \int_{\{(x,y);\ f_\alpha(x)=y\}} \frac{1}{2}|f_\alpha(x) - y| dP(x, y) \end{aligned}$$

The second integral is 0 because on the domain of integration, $\{(x, y);\ f_\alpha(x) = y\}$, the integrand, i.e.the function being integrated, is 0.

For the first integral, we note that on its domain of integration, $\{(x, y);\ f_\alpha(x) = y\}$, its integrand is always 1. Therefore,

$$\int_{\{(x,y);\ f_\alpha(x)\neq y\}} \frac{1}{2}|f_\alpha(x)-y|dP(x,y) = \int_{\{(x,y);\ f_\alpha(x)\neq y\}} 1dP(x,y) = P(\{(x,y);\ f_\alpha(x)\neq y\})$$

which, according to equation 1, is exactly $err_P(f_\alpha)$.

In the PAC approach to learning the selection of a particular hypothesis, f_α is done such that with *high probability the risk is very small*, and this makes this learning *probably* (high probability) *almost correct* (small risk).

Furthermore, it has been customary to fix the bounds on the error and probability of error and to calculate the sample size, i.e. the size of the training set, **needed** to reach these bounds.

This is usually unrealistic because:

- (1) In a learning problem we have a training set, '**"AS IS"** and most often cannot get more examples,
- (2) it turns out that in order to reach good results (tight bounds) the needed sample size needed is so large that for reasonable sample sizes (which would be available) the errors and their probabilities are too big to be useful.

More precisely, given S a set of training examples,

$$P(S; R(\alpha) = err_P(f_\alpha) \leq \epsilon) \geq 1-\delta \tag{2}$$

which is equivalent to

$$P(S; R(\alpha) = err_P(f_\alpha) > \epsilon) < 1-\delta \tag{3}$$

We called a hypothesis *consistent* if this hypothesis was correct with respect to the training set S. This means that the empirical risk, R_{emp} must be 0 and hence the bound on $R(\alpha)$ will be

$$R(a) \leq \sqrt{\frac{h(\ln\frac{2m}{h}+1)-\ln(\frac{\epsilon}{4})}{m}} \tag{4}$$

where

- h denotes the VC dimension of the hypothesis space H from which f_α comes,
- m is the size of the training set S.

NOTE *For consistent hypotheses the bound on $R(\alpha)$ depends only on the ratio $\frac{h}{m}$.*

Let now f_α be a hypothesis consistent to a training example $(x_0, y_0) \in S \times \{-1, +1\}$ such that

$$P((x_0,y_0);\ \ R(\alpha) > \epsilon) < 1-\epsilon \tag{5}$$

Therefore, by the iid assumption, for all $(x,y) \in S \times \{-1,+1\}$

$$P((x,y);R(\alpha)>\epsilon)=\Pi_{(x,y)}P((x,y);R(\alpha)>\epsilon)=(P((x,y);R(\alpha)>\epsilon))^m<(1-\epsilon)^m \tag{6}$$

and since $(1-\epsilon)^m < e^{-\epsilon m}$ (easy to check) it follows that

$$P((x,y)\in S\times\{-1,+1\};R(\alpha)>\epsilon)<e^{-\epsilon m} \tag{7}$$

Denote by $|H|$ the cardinality of the hypothesis space, that is the number of functions f_α (we assume that H is finite).

Then the probability that a hypothesis f_α which satisfies (7) exists is:

$$P(\text{that there exists a hypothesis } h\in H \text{ consistent to } S;\ R(\alpha)>\epsilon)=$$

$$=P\left(\bigcup_{f\in H}\{(x,y)\in S;\ h \textit{ consistent to } S;\ R(\alpha)>\epsilon\}\right)=$$

$$=\sum_{f\in H}P\left(\{(x,y)\in S;\ h \textit{ consistent to } S;\ R(\alpha)>\epsilon\}\right)\leq |H|\cdot e^{-\epsilon m}$$

In order for equation (2) (or (3)) to hold we require then (take ln on both sides) we need

$$|H|e^{-\epsilon m}=\delta \Leftrightarrow ln(|H|e^{-\epsilon m})=ln\delta$$

$$\Leftrightarrow$$

$$ln(|H|)+ln(e^{-\epsilon m})=ln\delta \Leftrightarrow ln(|H|)-ln(\epsilon m)=ln\delta$$

$$\Leftrightarrow$$

$$ln(|H|)-ln\delta=\epsilon\cdot m \Leftrightarrow ln\frac{|H|}{\delta}=\epsilon\cdot m$$

$$\Leftrightarrow$$

$$\epsilon=\frac{1}{m}ln\frac{|H|}{\delta} \tag{8}$$

or,

$$m=\frac{1}{\epsilon}ln\frac{|H|}{\delta} \tag{9}$$

That is, the sample size is proportional to the size of the hypothesis space. Thus,

$$P\left(S;\ f_\alpha \text{ consistent to } S;\ R(\alpha) > \frac{1}{m} ln \frac{|H|}{\delta}\right) < \delta \tag{10}$$

What does this mean from the point of view of PAC learning?

Assume that $\delta = 0.05$ and $\epsilon = 0.01$. Then, according to the equation (9), we need

$$m = \frac{1}{0.01} \ln \frac{|H|}{0.05} = 100 \ln(500|H|).$$

(Again, that is, the sample size is proportional to the size of the hypothesis space.)

For example,

- If $|H| = 3$ then we obtain $m \simeq 100 \times 7.31 = 731$;
- If $|H| = 10$ we obtain $m \simeq 100 \times 8.51 = 851$, etc.

Note however, that in general, $|H|$ will be much larger.

Let us now look at the following theorems:

Theorem 1 *Let H be a hypothesis space with finite VC dimension h. Then* **for any distribution** P *on $X \times \{-1, +1\}$:*

$$P\left(S; \exists f \in H;\ R_{emp} = 0;\ R(\alpha) < \frac{2}{m}\left(h\frac{\ln m + 1}{h} + \ln \frac{2}{\epsilon}\right)\right) = 1 - \delta$$

provided that $h \leq m,\ m > \frac{2}{\epsilon}$.

Theorem 2 *Let H be fixed with finite VC dimension h. Then* **for every** $f \in H$ **there exists P** *such that*

$$P\left[S; R_{emp} = 0;\ R(\alpha) \geq max\left\{\frac{h-1}{32m},\ \frac{1}{m} \ln \frac{1}{\delta}\right\}\right] \geq \delta \tag{11}$$

Looking at equation (11) on the complementary we have

$$P\left[S; R_{emp} = 0;\ R(\alpha) < max\left\{\frac{h-1}{32m},\ \frac{1}{m} ln \frac{1}{\delta}\right\}\right] < 1 - \delta \tag{12}$$

In general, the quantity $\frac{1}{m} ln \frac{1}{\delta}$ will be very big (since we want δ very small).

However, if h is big enough, more precisely, if $\frac{h-1}{32m} > \frac{1}{m} ln \frac{1}{\delta}$ (which is equivalent to $h > 1 + 32 ln \frac{1}{\delta}$ which can be very large!), then the maximum of the two quantities in the equations above is actually $\frac{h-1}{32m}$ and so by setting this quantity equal to ϵ we obtain $m = \frac{h-1}{32\epsilon}$ which means that for large VC dimension, to obtain small error, M must be very large.

Example 1 *For example,* $\epsilon = 0.01$, $\delta = 0.02$ *and assume* $h > 1 + 32 ln \frac{1}{\delta} \simeq 126$, *say* $h = 127$. *We need* $m = \frac{126}{32 \cdot 0.01} = \frac{99}{0.32} = 393$ *(or,* $m \simeq 3h$*)*

Theorems 1 and 2 should be "read" together: In both the hypothesis space, that is its VC dimension h is fixed and it is finite.

Theorem 1 says that **for every P there exists f** such that with large probability its risk is small;

Theorem 2 says that **for every f in H there exists a probability distribution P** such that probability to have a small risk is less than 1.

So, this means that once we fix H we should look for an approach which can find a "good" distribution (this distribution would have to be in a feature space not on the initial example space where the distribution, unkown as it maybe, is fixed). Such an approach is taken in **Support Vector Machines**(SVM).

The idea behind the SVM approach is to fix the class of hypotheses, $H = \{wx + b; (w, b)\}$ (eventually map the original feature space into another feature space – of much higher dimension) and find that "good" distribution (still unknown) for which Theorem 1 holds.

Then learning with SVM means finding that feature space where such a distribution exists.

Let us now look at the case in which $R_{emp} \neq 0$.

$$R_{emp} = \frac{1}{2m} \sum_{i=1}^{m} |f_\alpha(x_i) - y_i| = \frac{k}{m}$$

where k is the cardinality of the subset of the example set for which $f_\alpha(x_i) \neq y_i$, i.e. k is the number of training errors.

Then we have

$$P\left(R(\alpha) \leq \frac{k}{m} + \sqrt{\frac{h(\ln \frac{2m}{h} + 1) - \ln \frac{\epsilon}{4}}{m}} \right) = 1 - \epsilon$$

Note that the quantity

$$\sqrt{\frac{h(\ln(\frac{2m}{h}) + 1) - \ln(\frac{\epsilon}{4})}{m}}$$

is fixed once H and ϵ are fixed, and so the only way to keep a small bound on $R(\alpha)$ is to minimize k, the number of training errors.

This is nice, isn't it, since it kind of confirms our intuition when we try to develop any kind of learning that we should minimize the training errors.

2 Two examples

Example 2 *Consider the k*th *nearest classifier. The basic idea in this classifier is as follows:*

- *For each test point x we look at its k nearest neighbors which are in the training set and assign the test point the highest frquency label which occurs in this group.*
- *When $k = 1$ the algorithm assigns to each point the label of its nearest training example. Of course this requires us to exclude the case when the nearest neighbors are points of different classes.*

Then **any number of points will be learned** *by this algorithm, and hence the the corresponding function set has $VC - dimension = \infty$ and $R_{emp} = 0$. This means that the bound on R_α provides no information. Yet this type of classifier can still perform well.*

Example 3 *(*The notebook classifier*) This is a classifier for which the bounds should hold but which violates the bound).*

To achieve thes two conditions, when we would want R_α to be as large as possible and the bound on it, $R_{emp} + VC - confidence$ to be as small as possible.

That is, we want a family of classifiers which gives the worst possible actual risk, and this is 0.5*, the best empirical risk, $R_{emp=0}$, up to some number m of training examples, and such that its VC dimension, $h \leq l$ (where l is the total number of training examples) is easy to compute.*

The notebook *classifier is as follows: assume that the notebook contains enough room to write down the classes for m training examples.*

For all others examples, the classifier will assign the same class. It is assumed that there as many positive and negative examples (equal probability to draw one of tham at random). It is clear that this clasifier will have :

1. $R_{emp} = 0$ *for up to m examples;*
2. $R_{emp} = 0.5$ *for the remaining examples;*
3. $R_\alpha = 0.5$*;*
   and the VC dimension,
4. $h = m$.

From item (1) and (2) above it follows that $R_{emp} = \frac{l-m}{2 \cdot m}$
Therefore, it follows that

$$\frac{m}{4 \cdot m} \leq \ln \frac{2 \cdot l}{m} + 1 - \frac{1}{m} \ln \epsilon 4 \tag{13}$$

Now if we take :

$$f(t) = C \exp^{\frac{t}{4}-1} \leq 1, \;\; 0 \leq t \leq 1$$

since $f(t)$ is monotonic increasing and $f(1) = 0.236$ it follows that (13) holds for all ϵ.

3 Structural Risk Minimization (SRM)

We see that the terms that appear in the bound on the actual risk R_α depend on the chosen class of functions (the VC confidence) and on a **particular** function selected by the training step (the empirical risk).

The idea behind SRM is to find a subset of the set of functions which minimizes the empirical risk. Since the VC dimension h is an integer, the VC confidence is not a smooth function of h.

So, the hypothesis space is divided into nested subsets of functions ordered by the VC dimension: $H_1 \subset, ..., \subset H_n$ with corresponding VC dimensions $h_1 \leq, \ldots, \leq h_n$ and such that for each subset either the VC dimension or a bound on it can be calculated. For each subset, the goal of training is to minimize the empirical risk. Let $R_{emp}(H_i)$ be the minimum empirical risk for the ith subset. Then the final selection of the function f (the learning machine) is i_0 such that
$R_{emp}(H_{i_0}) + VC(h_0) = min_i\{R_{emp}(H_i) + VC(h_0)\}$.

Example 4 *(Conjunction of Booleans are PAC learnable) C: the class of target concepts – conjunction of boolean literal – (e.g. $a \wedge b \wedge \not c$).*

Claim 1 *C is PAC-learnable*

Proof:

Any consistent learner requires only a polynomial number of training examples: L consistent learner, H hypothesis space is identical to C.

Then from equation (9) we can compute m. All is needed is size of H, $|H|$ which is 3^n since each variable can (i) be included in a hypothesis, (ii) its negation can be included, or (iii) ignored.

Then from (9) with $H = 3^n$ we obtain

$$m \geq \frac{1}{\epsilon}\left(n \ln 3 + \ln\left(\frac{1}{\delta}\right)\right)$$

For example, $n = 10$, $\delta = 0.05$, $\epsilon = 0.1$), that is, with probability 0.95 the hypothesis will learn the concept with error up to 0.1, then

$$m = \frac{1}{0.1}\left(10 \ln 3 + \ln\left(\frac{1}{0.05}\right)\right) = 140$$

m grows linearly with n, $\frac{1}{\epsilon}$ and logarithmically in $\frac{1}{\delta}$.