Foundations for Machine Learning

L. Y. Stefanus

TU Dresden, June-July 2018

Slides 03

Reference

- Shai Shalev-Shwartz and Shai Ben-David. UNDERSTANDING MACHINE LEARNING: From Theory to Algorithms. Cambridge University Press, 2014.

Finite Hypothesis Classes

- The simplest type of restriction on a class is imposing an upper bound on its size (that is, the number of predictors h in H).
- We will show that if H is a finite class then ERM_H will not overfit, provided it is based on a sufficiently large training sample (this size requirement will depend on the size of H).

Finite Hypothesis Classes

- Limiting the learner to prediction rules within some finite hypothesis class may be considered as a reasonably mild restriction.
- For example, H can be the set of all predictors that can be implemented by a Python program written in at most 10^9 bits of code.

Finite Hypothesis Classes

- Another example of H is the class of axis aligned rectangles for the papaya learning problem, with discretized representation.

Performance Analysis of ERM_H

- H is a finite class.
- For a training sample, S, labeled according to some $f : X \rightarrow Y$, let h_S denote a result of applying ERM_H to S, namely,

$$h_S \in \underset{h \in H}{\operatorname{argmin}}\, L_S(h)$$

Performance Analysis of ERM_H

The Realizability Assumption:

There exists $h^* \in H$ such that $L_{(D,f)}(h^*) = 0$.

Note that this assumption implies that with probability 1 over random samples, S, where the instances of S are sampled according to D and are labeled by f,

we have $L_S(h^*) = 0$.

Performance Analysis of ERM_H

- Any guarantee on the error with respect to the underlying distribution D, for an algorithm that has access only to a sample S, should depend on the relationship between D and S.
- The common assumption in statistical machine learning is that the training sample S is generated by sampling points from the distribution D independently of each other.
- Expressed formally:

the i.i.d assumption

The examples in the training set are independently and identically distributed (i.i.d.) according to the distribution D. That is, every x_i in S is freshly sampled according to D and then labeled according to the labeling function, f. We denote this assumption by $S \sim D^m$ where m is the size of S, and D^m denotes the probability over m-tuples induced by applying D to pick each element of the tuple independently of the other members of the tuple.

- Intuitively, the training set S is a window through which the learner gets partial information about the distribution D over the world and the labeling function, f. The larger the sample gets, the more likely it is to reflect more accurately the distribution and labeling used to generate it.

Confidence Parameter $(1-\delta)$

- Since the training set S is picked by a random process, it is not realistic to expect that with full certainty S will suffice to direct the learner toward a good predictor (from the point of view of D), as there is always some probability that S happens to be very nonrepresentative of D.
- In the papaya tasting example, there is always some chance that all the papayas we have happened to taste were not tasty, in spite of the fact that, say, 75% of the papayas in our island are tasty. In such a case, $ERM_H(S)$ may be the constant function that labels every papaya as not tasty (and has 75% error on the true distribution of papayas in the island). Therefore …

Confidence Parameter (1-δ)

- Therefore, we will address the probability to sample a training set for which $L_{(D,f)}(h_S)$ is not too large. Usually, we denote the probability of getting a non-representative sample by δ, and call $(1 - \delta)$ the confidence parameter of our prediction.

Accuracy Parameter $\in$

- Furthermore, since we cannot guarantee perfect label prediction, we need another parameter for the quality of prediction, the accuracy parameter, commonly denoted by $\in$.
- We interpret the event $L_{(D,f)}(h_S) > \in$ as a failure of the learner, while if $L_{(D,f)}(h_S) \leq \in$ we view the output of the algorithm as an approximately correct predictor.
- Therefore …

Accuracy Parameter $\in$

- Therefore, we are interested in upper bounding the probability to sample m-tuple of instances that will lead to failure of the learner. The labeling function $f : X \to Y$ is fixed.
- Let $S|_x = (x_1, \ldots, x_m)$ be the instances of the training set. We would like to upper bound
$$D^m(\{S|_x : L_{(D,f)}(h_S) > \epsilon\})$$
- Let H_B be the set of bad hypotheses:
$$H_B = \{h \in H : L_{(D,f)}(h) > \epsilon\}$$
and M be the set of misleading samples:
$$M = \{S|_x : \exists h \in H_B, L_S(h) = 0\}$$
- M is misleading, because $\forall S|_x \in M$, there is a bad hypothesis that looks like a "good" hypothesis on $S|_x$.

Upper Bounding the Probability of Learner's Failure

- We want to bound the probability of the event $L_{(D,f)}(h_S) > \in$.
- Since the realizability assumption implies that $L_S(h_S) = 0$, it follows that the event $L_{(D,f)}(h_S) > \in$ can only happen if for some $h \in H_B$ we have $L_S(h) = 0$. In other words, this event will only happen if our sample is in the set of misleading samples, *M*. So, formally, we have shown that

$$\{S|_x : L_{(D,f)}(h_S) > \in\} \subseteq M.$$

- Rewriting *M* as

$$M = \bigcup_{h \in H_B} \{S|_x : L_S(h) = 0\}$$

we have …

Upper Bounding the Probability of Learner's Failure

we have

$$D^m(\{S|_x: L_{(D,f)}(h_S) > \epsilon\}) \le D^m(M) = D^m(\cup_{h\in H_B}\{S|_x: L_S(h) = 0\})$$

- Applying the union bound property from the Probability Theory to the right-hand side of the preceding equation yields

$$D^m(\{S|_x: L_{(D,f)}(h_S) > \epsilon\}) \le \sum_{h\in H_B} D^m(\{S|_x: L_S(h) = 0\}) \qquad (*)$$

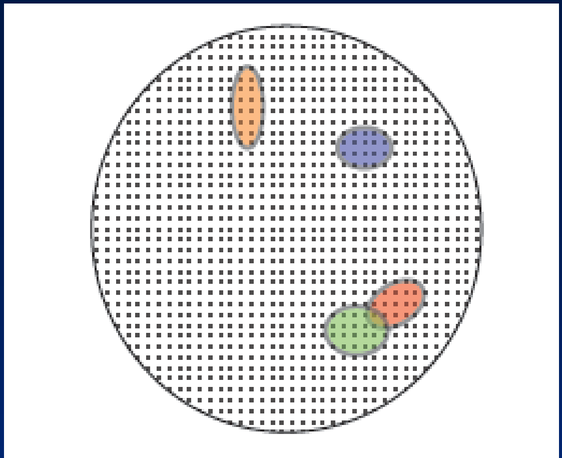
- Next, let us bound each summand of the right-hand side of the preceding inequality. Fix some "bad" hypothesis $h \in H_B$. The event $L_S(h) = 0$ is equivalent to the event $\forall i, h(x_i) = f(x_i)$. Since the examples in the training set are sampled i.i.d. we get

$$\begin{aligned}\mathcal{D}^m(\{S|_x : L_S(h) = 0\}) &= \mathcal{D}^m(\{S|_x : \forall i, h(x_i) = f(x_i)\}) \\ &= \prod_{i=1}^{m} \mathcal{D}(\{x_i : h(x_i) = f(x_i)\}).\end{aligned} \qquad (**)$$

Upper Bounding the Probability of Learner's Failure

- For each individual sampling of an element of the training set we have
$$D(\{x_i : h(x_i) = f(x_i)\}) = 1 - L_{(D,f)}(h) \leq 1 - \epsilon,$$
where the last inequality follows from the fact that $h \in H_B$ such that $L_{(D,f)}(h) > \epsilon$. Combining the previous equation with Equation (**) and using the inequality $1 - \epsilon \leq e^{-\epsilon}$ we obtain that for every $h \in H_B$,
$$D^m(\{S|_x : L_S(h) = 0\}) \leq (1 - \epsilon)^m \leq e^{-\epsilon m}.$$
Combining this inequality with Inequality (*) we conclude that $D^m\left(\left\{S|_x : L_{(D,f)}(h_S) > \epsilon\right\}\right) \leq |H_B| e^{-\epsilon m} \leq |H| e^{-\epsilon m}$.

A graphical illustration of the union bound result



Each point in the large circle represents a possible m-tuple of instances. Each colored oval represents the set of misleading m-tuple of instances for some bad predictor $h \in H_B$. The ERM can potentially overfit whenever it gets a misleading training set S. That is, for some $h \in H_B$ we have $L_S(h) = 0$. The result of the union bound guarantees that for each individual bad hypothesis, at most $(1-\epsilon)^m$-fraction of the training sets would be misleading. In particular, the larger m is, the smaller each of these colored ovals becomes. The union bound formalizes the fact that the area representing the training sets in M is at most the sum of the areas of the colored ovals. Therefore, it is bounded by $|H_B| \bullet$ (the maximum size of a colored oval). Any sample S outside the colored ovals cannot cause the ERM to overfit.

So we have derived the following theorem about learnability.

Theorem 1

Let *H* be a finite hypothesis class. Let $\delta \in (0,1)$ and $\epsilon > 0$ and let *m* be an integer that satisfies

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

Then, for any labeling function, *f*, and for any distribution, *D*, for which the realizability assumption holds (that is, for some $h \in H, L_{(D,f)}(h) = 0$), with probability of at least $1 - \delta$ over the choice of an i.i.d. sample *S* of size *m*, we have that for every ERM hypothesis, h_S, it holds that

$$L_{(D,f)}(h_S) \leq \epsilon.$$

- Theorem 1 tells us that for a sufficiently large sample m, the ERM_H rule with a finite hypothesis class H will be probably (with confidence $1 - \delta$) approximately (up to an error of ϵ) correct.