A formal Model - The statistical Learning Framework

Domain Set : An arbitrary set, X. This is the set of objects that we may wish to label. These data points are represented by a vector of features. This set is from some probability distribution D.

Label Set : An arbitrary set of all possible labels, Y.
As an example, in a two-element set we have $Y = \{0, 1\}$

Training data :

$S = ((x_1, y_1) \cdots (x_m, y_m))$, Sample set $X \times Y$ used to train the learner.

Learner's Output : The learner will output a ~~learner~~ hypothesis (h) which is a function $h: X \to Y$, used to predict the label of new data points.

Target Function : $(f : X \to Y)$

Target function maps the X to Y. Target function is unknown. Hypothesis, h actually approximates the target function.

Probability Distribution : Suppose $A \subset X$, the probability distribution D will assign a number $D(A)$, which determines how likely it is to observe a point $x \in A$.

Measure of Success :

The error of h is the probability that $h(x) \neq f(x)$ for any given x taken from the underlying distribution.

Error is given by

$$L_{D,f}(h) \equiv \underset{x \sim D}{P}[h(x) \neq f(x)] = D(\{x : h(x) \neq f(x)\})$$

$L_{D,f}(h)$ is also called Generalization error, the risk or the true error of h.

Empirical Risk Minimization

The goal of the learning algorithm is to find h_S that minimizes the error w.r.t to the unknown D & f.

Since D & f are unknown, the true error is not available to the learner. A useful notion of error that can be calculated by the learner is the training error — the error the classifier incurs over the training sample

$$L_S(h) = \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m}$$

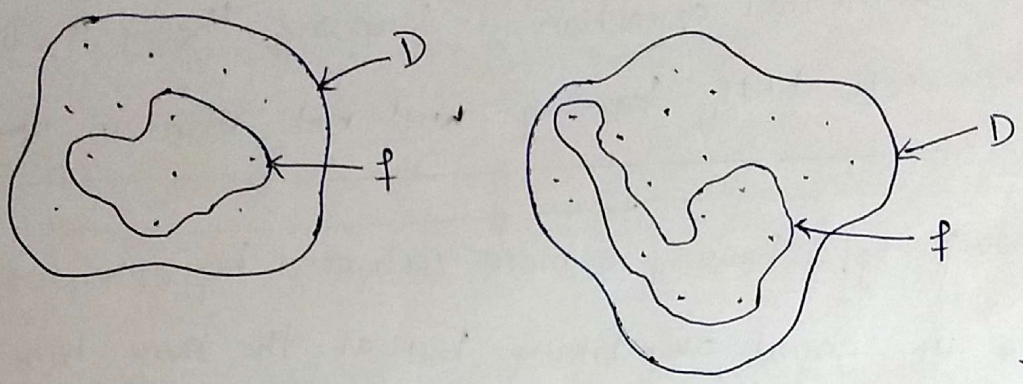
where $[m] = \{1, 2, \cdots, m\}$

$L_S(h)$ is termed as the Empirical error | Empirical risk

The learning paradigm - coming up with a predictor h that minimizes $L_S(h)$ - is called Empirical Risk Minimization | ERM

Note: Since the training sample is the snapshot of the world that is available to the learner, it makes sense to search for a solution that works well on that data. This can result in over-fitting

Side Note:



Overfitting model

$$h_S(x) = \begin{cases} y_i & \text{if } \exists i \in [m] \text{ s.t. } x_i = x \\ 0 & \end{cases}$$

Empirical Risk Minimization with Inductive Bias

A common solution to overfitting is to apply the ERM learning rule over a restricted search space. Formally, the learner should choose in advance a set of predictors. This set is called hypothesis set H. Each $h \in H$ is a function mapping from $X \to Y$. For a given class H, and a training sample, S, the ERM_H learner uses the ERM rule to choose a predictor $h \in H$, with the lowest possible error over S. Formally

$$ERM_H(S) \in \underset{h \in H}{\operatorname{argmin}} L_S(h)$$

① Learning has lot to do with prior knowledge
Ex: Bait Shyness

② - Prior knowledge
Less sample data points
- No PK - more

where argmin stands for the set of hypotheses in H that achieve the minimum value of $L_S(h)$ over H. By restricting the learner to choosing a predictor from H, we bias it toward a particular set of predictors. Such restrictions are often called an inductive bias.

A fundamental question in learning theory is, over which hypothesis set ERM_H learning will not result in over fitting.

Intuitively, choosing a more restricted hypothesis class better protects us against over fitting but at the same time might cause us a stronger inductive bias.

Finite Hypothesis Set (then ERM_H will not over fit)

The hypothesis set contains infinite hypothesis. The simplest type of restriction is imposing an upper bound on its size (the number of predictors h in H)

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

Realizability assumption says there exists $h^* \in H$ s.t. $L_{D,f}(h^*) = 0$, this implies $L_S(h^*) = 0$ for any sample drawn from D. & labelled by f. This means $h^* = f$

The i.i.d assumption: (Independently & Identically Distributed).

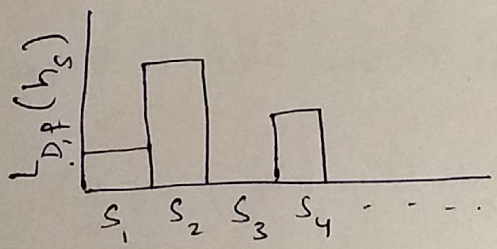
Every x_i in S is freshly sampled according to the distribution D & then labelled according to the labelling function, f.

denote this assumption by $S \sim D^m$, where m is the size of S & D^m, the probability distribution over m tuples.

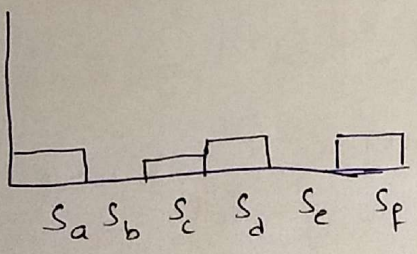
Intuitively, the training set S is a window through which the learner gets partial information about the distribution D & the labelling function, f. So, the larger the sample, the better it is.

Since $L_{D,f}(h_S)$ depends on the training set, S, and that training set is picked by a random process, there is randomness in the choice of the predictor h_S & consequently, in the risk $L_{D,f}(h_S)$. Formally, we say $L_{D,f}(h_S)$ is a random variable.

Random Variable : is a variable whose possible values are outcomes of a random phenomenon.



Here $S_1, S_2, \cdots$ all have the same sample size.



Here $S_a, S_b, \cdots$ all have same sample size but bigger than $S_1, S_2 \cdots$

Let us denote probability of getting a non representative sample by δ, and call $(1-\delta)$ the confidence parameter of our prediction.

On top of that, since we cannot guarantee perfect label prediction, we introduce another parameter for the quality of prediction, the accuracy parameter, commonly denoted by ϵ.

We interpret the event $L_{D,f}(h_S) > \epsilon$ as failure of the learner, while if $L_{D,f}(h_S) \leq \epsilon$ we view the output of the algorithm as an approximately correct predictor.

Therefore we are interested in upper bounding the probability to sample of instances that will lead to failure of the learner.

Formally, let $S|_x = (x_1, x_2, \cdots, x_m)$ be the instances of the training set. We would like to upper bound

$$D^m(\{S|_x : L_{D,f}(h_S) > \epsilon\}) \quad \{\text{find out the optimal } m\}$$

Let H_B be the set of "bad" hypotheses, that is,

$$H_B = \{h \in H : L_{D,f}(h) > \epsilon\}$$

In addition, let

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

be the set of misleading samples: Namely, for every

$S|_x \in M$, there is a "bad" hypothesis, $h \in H_B$, that looks like a "good" hypothesis on $S|_x$. Now, recall that we would like to bound the probability of the event $L_{D,f}(h_S) > \epsilon$. But since the realizability assumption implies that $L_S(h_S) = 0$, it follows that the event $L_{D,f}(h_S) > \epsilon$ 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. Formally, we have shown that

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

Note that we can rewrite M as

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

Hence,

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

Next, we upper bound the right hand side of the preceding equation using the union bound - a basic

property of probabilities.

Union Bound For any two sets A, B and a distribution D we have $D(A \cup B) \leq D(A) + D(B)$

Applying the union bound to the right hand side yields

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

$$D^m(\{S|_x : L_S(h) = 0\}) = D^m(\{S|_x : \forall i, h(x_i) = f(x_i)\})$$

$$= \prod_{i=1}^{m} D(\{x_i : h(x_i) = f(x_i)\})$$

For each individual sampling of an element of the training set we have

$$D(\{x_i : h(x_i) = y_i\}) = 1 - L_{D,f}(h) \leq 1 - \epsilon,$$

Combining the previous equation & using the inequality $1 - \epsilon \leq e^{-\epsilon}$ we obtain that for every $h \in H_B$,

$$D^m(\{S|_x : L_S(h) = 0\}) \leq (1-\epsilon)^m \leq e^{-\epsilon m}$$

Finally $D^m(\{S|_x : L_{D,f}(h_S) > \epsilon\}) \leq |H_B| e^{-\epsilon m} \leq |H| e^{-\epsilon m}$

Important

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

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

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

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

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

Finally, the Machine Learning Diagram