Statistical Learning Framework:

Domain set : $\mathcal{X} \to$ set of objects that we wish to label. The points in $\mathcal{X}$ will often be described by feature vectors.

Label set: $\mathcal{Y}$ want to label elements of $\mathcal{X}$ with labels from $\mathcal{Y}$.

Learning alg:

Input $\to$ training "set" a sequence from $(\mathcal{X} \times \mathcal{Y})^*$.

Output $\to$ $h: \mathcal{X} \to \mathcal{Y}$ a prediction rule also called a classifier.

The training data is generated by sampling from some <u>unknown</u> distribution $\mathcal{D}$ over $\mathcal{X}$. We assume there is a "correct" labeling f" $f: \mathcal{X} \to \mathcal{Y}$ which is unknown to the learner but is used to label the training data.

The <u>error</u> of $h: \mathcal{X} \to \mathcal{Y}$ is

$$L_{\mathcal{D},f}(h) \stackrel{\text{def}}{=} \mathbb{P}_{x \sim \mathcal{D}} [h(x) \neq f(x)] = \mathcal{D}(\{x : h(x) \neq f(x)\}).$$

The error depends on $\mathcal{D}$ and f. The learner cannot compute this, it can compute

$$L_S(h) \stackrel{\text{def}}{=} \frac{|\{i \in [m] \mid h(x_i) \neq y_i\}|}{m}$$

This is called <u>empirical</u> <u>error</u> or <u>empirical</u> <u>risk</u>.

Basic paradigm: empirical risk minimization.

Problem overfitting: Suppose $\mathcal{X}$ is a square of area 2 in the plane and there is a smaller square inside it of area 1.

Assume the labelling f f maps points in the inner sq to 1 and other points to 0. Suppose $\mathcal{D}$ is uniform over the larger square. We define h from a sample S by

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

Now this will always give $L_S = 0$! But $L_D = \frac{1}{2}$ so the performance is terrible.

Inductive bias: force the learner to choose from a predefined set of possible h & then impose ERM. What classes of possible h will prevent overfitting? Let the set of possible rules (hypotheses) be $\mathcal{H}$.

Simple restriction $|\mathcal{H}| \leq n$.

So ERM gives $h_S = \underset{h \in \mathcal{H}}{\operatorname{argmin}} L_S(h)$.

We make a simplifying assumption $\exists h^* \in \mathcal{H}$ s.t. $L_{\mathcal{D},f}(h^*) = 0$. This means that with prob 1 $L_S(h^*) = 0$.

Now with this assumption we get that with prob 1 ERM produces h_S with $L_S(h_S) = 0$. What is $L_{\mathcal{D},f}(h_S)$?

We need to understand sampling. Basic assumption the samples are i.i.d according to $\mathcal{D}$: $S \sim \mathcal{D}^m$.

If we are unlucky and get an unrepresentative sample we get large true error. We write δ for the prob. of getting a bad sample. Even if we get a good sample we won't get an exact rule: we introduce an accuracy parameter ϵ. We view $L_{D,f}(h_S) > \epsilon$ as a failure of the algorithm. We want bounds on the prob. of failure.

$\mathcal{D}^m$ is the distribution on i.i.d samples of size m.
Want to bound $\mathcal{D}^m(\{S \mid L_{\mathcal{D},f}(h_s) > \varepsilon\})$.
We have a set of "bad" hypotheses

$$\mathcal{H}_B \stackrel{\text{def}}{=} \{h \in \mathcal{H} \mid L_{(\mathcal{D},f)}(h) > \epsilon\}.$$

Now a sample is "misleading" if it makes one of the bad hypotheses look good:

$$M = \{S \mid \exists h \in \mathcal{H}_B \;\; L_S(h) = 0\}.$$

The simplifying assumption implies $L_S(h) = 0$ so if $L_{(\mathcal{D},f)}(h_s) > \epsilon$ it must be because $S \in M$.

Thus $\{S \mid L_{(\mathcal{D},f)}(h_s) > \epsilon\} \subseteq M$.

Now $M = \bigcup_{h \in \mathcal{H}_B} \{S \mid L_S(h) = 0\}$ so

$$\mathcal{D}^m(\{S \mid L_{(\mathcal{D},f)}(h_s) > \epsilon\}) \leq \mathcal{D}^m(M) = \mathcal{D}^m\Big(\bigcup_{h \in \mathcal{H}_B} \{S \mid L_S(h) = 0\}\Big)$$

$$\leq \sum_{h \in \mathcal{H}_B} \mathcal{D}^m(\{S \mid L_S(h) = 0\}).$$

Let us bound each term. Now $L_S(h) = 0$ means $\forall i \;\; h(x_i) = f(x_i)$ & since the sampling is i.i.d.

$$\mathcal{D}^m(\{S \mid \forall i \; h(x_i) = f(x_i)\})$$
$$= \prod_{i=1}^m \mathcal{D}(\{x_i \mid h(x_i) = f(x_i)\})$$
$$= \prod_{i=1}^m \mathcal{D}(\{x_i \mid h(x_i) = y_i\}) = \prod_{i=1}^m (1 - L_{(\mathcal{D},f)}(h)) \leq (1-\epsilon)^m$$

since h is a bad hypothesis.

so $\mathcal{D}^m(\{S \mid \cdots\}) \leq (1-\epsilon)^m \leq e^{-m\epsilon}$

Thus our bound is

$$\mathcal{D}^m(\{S \mid L_{(\mathcal{D},f)}(h_s) > \epsilon\}) \leq |\mathcal{H}_B| \cdot e^{-m\epsilon} \leq |\mathcal{H}| \cdot e^{-m\epsilon}.$$

<u>Cor</u> If $\delta \in (0,1)$ & $m \geq \dfrac{\log(|\mathcal{H}|/\delta)}{\epsilon}$ then with prob $(1-\delta)$ we have $L_{(\mathcal{D},f)}(h_s) \leq \epsilon$.

Prop Note in general that $\mathbb{E}[L_S(h)] = L_{(\mathcal{D},f)}(h)$.

Proof
$$L_S(h) = \frac{1}{m}\sum_{i=1}^{m} 1_{[h(x_i)\neq f(x_i)]}$$

so
$$\mathbb{E}[L_S(h)] = \frac{1}{m}\sum_{i=1}^{m} \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[1_{\{h(x_i)\neq f(x_i)\}}]$$

$$= \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[1_{\{h(x)\neq c(x)\}}] = L_{(\mathcal{D},f)}(h).$$

Def of PAC Learning ~~A set of~~ A labelling f" $f: \mathcal{X}\to\mathcal{Y}$ is called a concept. A concept class C is said to be PAC-learnable if $\exists$ an alg. $\mathcal{A}$ and a polynomial $p(\cdot,\cdot,\cdot,\cdot)$ s.t. $\forall\, \epsilon>0, \delta>0$, $\forall\, \mathcal{D}$ on $\mathcal{X}$ and $\forall f\in C$ $\forall$ samples of size $m \geq p(1/\epsilon, 1/\delta, n, cost(f))$

$$\Pr_{S\sim\mathcal{D}^m}\Big[\underbrace{L_{(\mathcal{D},f)}(h_s)\leq\epsilon}_{\text{APPROXIMATELY}}\Big] \geq \underbrace{1-\delta}_{\text{PROBABLY}}$$

n is cost of representing an element of $\mathcal{X}$ & $cost(f)$ is the cost of representing f computationally.

If, in addition, $\mathcal{A}$ runs in time polynomial in $1/\epsilon$ & $1/\delta$ we say C is efficiently PAC learnable.

δ : confidence in the result

ϵ : accuracy of the result.

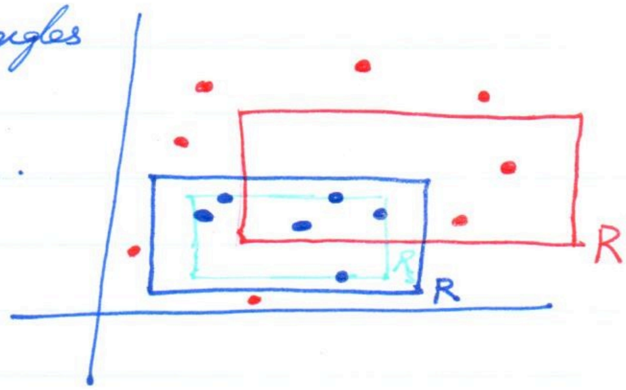
Note we quantify over all distributions. The training sample (which produces h_S) & the test sample (used to estimate $L_{(\mathcal{D},f)}$) are both drawn according to the distribution $\mathcal{D}$. The concept class C is known to $\mathcal{A}$ but of course not f.

Learning axis aligned rectangles

$X = R^2$

C: axis aligned rectangles.

From a labelled sample of points determine a rectangle.



$R \cap (R')^c$: false negatives

$R' \cap R^c$: false positives

Alg A: given a sample S return the tighest rectangle containing points labelled by 1. R_S

R_S will never give false positives. So the errors generated by R_S are all inside R.

We have some distribution of points in R^2.

Let us assume $P_r[R] > \epsilon$ otherwise the algorithm is trivially going to work regardless of S.

Define 4 rectangular strips along the edges of R s.t each of r_i has prob $\geq \epsilon/4$.

The corners of R are $(\cancel{b}\ l, b)$, (l, t), (r, b), (r, t).

$R = \cancel{(l,r)}\ [l, r] \times [b, t]$.

r_4 is $[l, s_4] \times [b, t]$ where $s_4 = \inf\{s \mid P_r\, [l,s]\times[b,t] \geq \frac{\epsilon}{4}\}$

Define $\bar{r}_4 = [l, s_4) \times [b, t]$, $P_r\, \bar{r}_1 \leq \epsilon/4$.

If R_S meets all 4 regions then it has a side in each region so its error region is included in the union of the strips and thus $\leq 4 \cdot \epsilon/4 \leq \epsilon$. If R_S has error $> \epsilon$ then R_S must miss one of the r_i completely

$$\Pr_{S\sim\mathcal{D}^m}[L(R_S)>\epsilon] \le \Pr_{S\sim\mathcal{D}^m}\left[\bigvee_{i=1}^{4} (R_S \cap r_i)=\phi\right]$$

$$\le \sum_{i=1}^{4} \Pr_{S\sim\mathcal{D}^m}[R_S \cap r_i = \phi]$$

$$\le 4\cdot(1-\tfrac{\epsilon}{4})^m \quad [\Pr[r_i] \ge \epsilon/4]$$

$$\le 4\exp(-m\epsilon/4). \quad [1-x \le e^{-x}]$$

So for any $\delta>0$ to ensure $\Pr_{S\sim\mathcal{D}^m}[L_S(R_S)>\epsilon]\le\delta$

we set $4e^{-m\epsilon/4} \le \delta$

or $m \ge \frac{4}{\epsilon}\log\frac{4}{\delta}$.

So if the sample size is bigger than this we get our PAC bounds. We say the sample complexity is $O(\frac{4}{\epsilon}\log\frac{4}{\delta})$.

We can equivalently say: if the sample size is m then error ϵ is bounded by $\frac{4}{m}\log\frac{4}{\delta}$.

Sample bounds for finite hypothesis sets - consistent case

Thm Let H be a finite set of concepts $f: \mathcal{X}\to\mathcal{Y}$. Let A be an alg that returns a consistent hypothesis h_S for any target concept f and sample S. Then $\forall\,\epsilon,\delta>0$ the inequality $\Pr_{\mathcal{D}}[L(h_S)\le\epsilon] > 1-\delta$ holds if $m \ge \frac{1}{\epsilon}(\log|H| + \log\frac{1}{\delta})$.

Proof Fix $\epsilon>0$: $\Pr[\exists h\in H \mid L_S(h)=0 \,\&\, L(h)>\epsilon]$

$$\le \sum_{h\in H}\Pr[L_S(h)=0 \,\&\, L(h)>\epsilon] \quad \text{(union bd)}$$

$$\le \sum_{h\in H}\Pr[L_S(h)=0 \mid L(h)>\epsilon] \quad \text{(cond. prob)}$$

$$\le \sum_{h\in H}(1-\epsilon)^m \le \sum_{h\in H} e^{-m\epsilon} = |H|\cdot e^{-m\epsilon} \le \delta$$

so $m \ge \frac{1}{\epsilon}[\log|H| + \log\frac{1}{\delta}]$

Conjunction of boolean literals:

C_n conjunctions of at most n boolean literals

e.g. $x_1 \wedge x_2 \wedge \bar{x}_3 \wedge x_4$.

$1 \quad 1 \quad 1 \quad 1 \to 0$

$1 \quad 1 \quad 0 \quad 1 \to 1$ } the only pos. example

$0 \quad 0 \quad 0 \quad 1 \to 0$

Positive examples tell you much more.

Alg: If $(b_1, \ldots b_n)$ is a positive example then for all i s.t. $b_i = 1$ $\bar{x}_i$ is ruled out & for all i s.t. $b_i = 0$ x_i is ruled out.

Return the conjunction of all literals not ruled out.

$|H| = 3^n$. Using the bound

$m \geq \frac{1}{\epsilon}((\log 3)\, n + \log \frac{1}{\delta})$.

So this class is PAC learnable.

Boolean vectors: $\mathcal{X} = \{0,1\}^n$. $U_n = 2^{\mathcal{X}}$. Is this class PAC learnable? $|H| = 2^{2^n}$.

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

We need samples exponential in n.

(k,n) term DNF disjunctions of at most k terms each term is a conj of at most n boolean literals.

$|H| = (3^n)^k$ so $m \geq \frac{1}{\epsilon}((\log 3)\, nk + \log \frac{1}{\delta})$

so PAC learnable. But we can show that for $k=3$ learning is in RP so ~~unlikely to be~~ we have no clue! [Mohri's remarks seem wrong here.]