Statistical learning theory

Setup:

- Sample space X. (also called instance space).
  Its points are called samples. (or instances)
  Usually we have $X \subseteq \mathbb{R}^n$, where each component represents a feature.
  i.e. each sample is represented by a vector of features.

- Label set y.
  For multiclassification y is finite.
  For regression $y = \mathbb{R}$.

- Training data: ~~finite $S \subseteq$~~
  With $Z = X \times y$. the training data will be represented by ~~some finite $S \subseteq Z^m$~~
  $S \in Z^m$.

Thus the training data is a pt. $S = ((x_1, y_1), \ldots, (x_m, y_m))$ consisting of labeled samples.

Prediction set.

- A learner is requested to output a prediction function $h: X \to Y$, predicting the relationship between samples and labels - for example. if there is a function relation $f: X \to Y$, then the learner is asked to approximate this function.

Assumptions on data

- The training data is not completely random. But rather. it is sampled from an underlying distribution

That is: we have a Random variable

$$(X, Y): \Omega \to X \times Y \subset \mathbb{R}^n \times Y.$$

which induces a ~~Random on X×Y~~ push-forward measure on $X \times Y$.

So, we have a joint distribution on samples and labels. called D

Intuitively this joint distribution decomposes into

- marginal distribution
- conditional distribution.

~~Intuit~~

Assuming ~~densities~~ we have a joint density $f_{X,Y}$ we obtain the marginal density

$$f_X(x) = \int f_{X,Y}(x,y)\, dy$$

and define $f_{Y|X}(y) = \dfrac{f_{X,Y}(x,y)}{f_X(x)}$

Note: In general we do not have access to this distribution.

Interpretation/ Example:

- The marginal distribution describes the distribution of the samples as they are not explicitly modeled
- The conditional distribution describes the probability given a sample x, which ~~subset~~ labels it might have.

Say we want to classify papayas into

- 1 = "tasty"
- 0 = "non-tasty"

We represent each papaya by the features

- Color
- softness

Thus $X \subseteq \mathbb{R}^2$ is the space of all Papayas represented by color and softness.

Clearly: X does not contain arbitrary points as there are limits to color and softness.

The marginal distribution describes how likely it is to have papayas with given color and softness.

The conditional distribution describes given a particular Papaya, how likely is it that it's tasty.

So the joint distribution describes the probability that a tasty papaya w/ given color and softness occurs.

In practice:

We expect outliers to occur less often!

- ~~Error functions~~ / loss functions.

~~loss functions~~:

Hypothesis class:

~~This~~ A set $\mathcal{H}$ of prediction functions.
This could be, all functions
· linear functions.
⋮

Error functions

Given $\mathcal{H}$ and $\mathcal{Z}$, a loss function
a loss function is a function

$$\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$$

s.t. $\ell(h, -) : \mathcal{Z} \to \mathbb{R}_{\geq 0}$ is measurable

In our supervised learning setting we have $Z = X \times Y$. but this formulation includes unsupervised learning.

The risk function / true error is the expected loss of a prediction function h over Z.

i.e. $L_D(h) : H \to \mathbb{R}^+$

$$h \mapsto \mathbb{E}_{z \sim D}[\ell(h, z)]$$

The empirical risk / empirical error is the expected loss over a given sample $S \in Z^m$

$$L_S : H \to \mathbb{R}^+$$

$$h \mapsto \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i)$$

Example: . 0-1 loss for classification:

$$\ell_{0-1}(h, (x,y)) = \begin{cases} 0 & h(x) = y \\ 1 & h(x) \neq y \end{cases} \Rightarrow L_D(h) = \mathbb{P}_{(x,y) \sim D}[h(x) \neq y] = D(h(x) \neq y)$$

. Square loss $\ell_q(h, (x,y)) = (h(x) - y)^2$

$$L_D(h) = \int (h(x) - y)^2 \, dx \, dy$$

i.i.d.:

Remember that our training data is given by a pt $S \in Z^m = (X \times Y)^m$

We assume that the samples are all independent.

Notation $S \sim D^m$.

Empirical risk minimization

Remember that

Given a training set $S \in Z^m$ sampled from an unknown D we want to output a predictor $h_S : X \to Y$ which depends on S.

Since D is unknown we only have access to the empirical risk.

$$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i)$$

Note: That is simply the training error!

Clearly, minimizing wrto the empirical risk when $\mathcal{H}$ = all will lead to overfitting.

We want to search for conditions, which guarantee that ERM does not overfit but is also likely to perform well on unseen data.

The obvious problem is that $\mathcal{H}$ consists of many functions at this point.

The simplest restriction is that $\mathcal{H}$ is finite.

We show that for $\mathcal{H}$ finite we can guarantee not to overfit while retaining good generalization performance.

To this end we will make the following:

Definition (Realizability assumption)

There exists $h^* \in \mathcal{H}$ s.t. $L_D(h) = 0$

Note that this implies w/ probability 1 that w.r. a random sample S we have

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

We moreover assume that $L_D(h_S)$ is a random variable in S.

Now:

Analysis of finite $\mathcal{H}$ | Given $h_S \in \operatorname{argmin}_{h \in \mathcal{H}} L_S(h)$, give a bound on $L_D(h_S)$

- It is not realistic to expect w/ full certainty that S will suffice to direct the learner to a good predictor.

It could always be that the sample is non-representative of $\mathcal{D}$.

Thus we can only speak of the probability that $L_{\mathcal{D}}(h_S)$ is not too large.

Def: The probability of getting a non-rep. sample by δ.

Call $(1-\delta)$ the confidence parameter.

However, we cannot guarantee perfect label prediction.

Def: So we have an accuracy parameter $\varepsilon > 0$.

Def: For $\varepsilon > 0$ we interpret $L_{\mathcal{D}}(h_S) > \varepsilon$ as failure to learn.

$L_{\mathcal{D}}(h_S) < \varepsilon$ is an approximately correct predictor.

We want to estimate the size of the event $L_D(h_S) > \varepsilon$ in S.

More precisely:

~~We want $S \in Z^m$~~

We want to estimate

$$D^m(\{S \in Z^m \mid L_D(h_S) > \varepsilon\}) \quad \text{where } L_D(h_-) \text{ is random variable}$$

Let $H_B \subset H$ be the set of "bad" hypotheses

i.e. $H_B = \{h \in H \mid L_D(h) > \varepsilon\}$

Set $M = \{S \in Z^m \mid \exists h \in H_B, L_S(h) = 0\}$

the set of "bad" samples.

i.e. for $S \in M$, the bad hypothesis looks good.

The realizability assumption implies

$L_S(h_S) = 0$.

$\Rightarrow$ $L_D(h_S) > \varepsilon$ can only happen, if

for some $h \in \mathcal{H}_B$ we have $L_S(h) = 0$.

($\Leftrightarrow$ $S \subseteq M$.

We have

$$\{ S \in Z^m \mid L_D(h_S) > \varepsilon \} \subseteq M.$$

Now write

$$M = \bigcup_{h \in \mathcal{H}_B} \{ S \in Z^m \mid L_S(h) = 0 \}.$$

Hence:

$$D^m(\{ L_D(h_S) > \varepsilon \}) \leq D^m(M) = D^m\left(\bigcup_{h \in \mathcal{H}_B} \{ S \in Z^m \mid L_S(h) = 0 \} \right)$$

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

Now we bound each summand:

Fix $h \in \mathcal{H}_B$. Then $L_S(h) = 0$ is equiv to $h(x_i) = y_i \quad \forall i$.

Hence

$$D^m(\{S \in \mathcal{X}^m \mid L_S(h) = 0\})$$

$$= D^m(\{S \in \mathcal{X}^m \mid h(x_i) = y \ \forall i\})$$

$$= \prod_{i=1}^{m} D(\{(x_i, y_i) \mid h(x_i) = y_i\})$$

$L_D(h) = D(\{x_i y_i\} \mid h(x_i) \neq y_i)$
b/c. classification.

we have

$$D(\{(x_i, y_i) \mid h(x_i) = y_i\}) = 1 - L_D(h)$$

$$\leq 1 - \varepsilon$$

since $h \in \mathcal{H}_B$ and $L_D(h) \geq \varepsilon$

Now use $1 - \varepsilon \leq e^{-\varepsilon}$ we have,

$$D^m(\{S \in \mathcal{X}^m \mid L_S(h) = 0\}) \leq (1-\varepsilon)^m \leq e^{-\varepsilon m}$$

and

we conclude that

$$D^m(|S \in \mathcal{Z}^m : L_D(h_S) > \varepsilon|) \leq |\mathcal{H}_B| e^{-\varepsilon m} \leq |\mathcal{H}| e^{-\varepsilon m}$$

We have proven:

Prop: Let $\mathcal{H}$ be finite. $\delta \in (0,1)$ and $\varepsilon > 0$

Let $m \in \mathbb{N}$ s.t. $m \geq \frac{\log(|\mathcal{H}|/\delta)}{\varepsilon}$

Then for any D for which realizability holds

with probability $1-\delta$ over the choice of

sample S of size m we have

$$L_D(h_S) \leq \varepsilon$$