Now we want to bound each term in $(\star\star)$

Recall: $L_{\mathcal{D}}(h) = \mathbb{E}_{z\sim\mathcal{D}}[\ell(h,z)]$

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

Important: each z_i is sampled i.i.d. from $\mathcal{D}$

$$\mathbb{E}[\ell(h,z_i)] = \mathbb{E}_{z\sim\mathcal{D}}[\ell(h,z)] = L_{\mathcal{D}}(h)$$

Therefore: $\mathbb{E}[L_S(h)] = \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} \ell(h,z_i)\right]$

by def. of $L_S(h)$

by linearity of expectation

$$= \frac{1}{m}\sum_{i=1}^{m} \underbrace{\mathbb{E}[\ell(h,z_i)]}_{L_{\mathcal{D}}(h)}$$

$$= \frac{1}{m}\cdot m \cdot L_{\mathcal{D}}(h) = L_{\mathcal{D}}(h)$$

Let θ_i be the r.v. given by $\ell(h, z_i)$ i-th element of S

Since h is fixed and z_i is sampled i.i.d. from $\mathcal{D}$

$\Rightarrow \theta_1, \theta_2, \dots, \theta_m$ are i.i.d. r.v.

Note that: $L_S(h) = \frac{1}{m}\sum_{i=1}^{m} \theta_i$. Let's define $\mu = L_{\mathcal{D}}(h)$

Given assumption that $\ell : \mathcal{H} \times Z \to [0,1]$

we have $\theta_i \in [0,1]$, $\forall i = 1, \dots, m$.

We can apply Hoeffding's inequality with $a_i = 0$, $b_i = 1$ $\forall i=1..m$

$$\mathcal{D}(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}) = \Pr\left[\left|\left(\frac{1}{m}\sum_{i=1}^{m}\theta_i\right) - \mu\right| > \varepsilon\right]$$

by Hoeffding's ineq. $\to \leq 2 \cdot e^{-2m \cdot \varepsilon^2}$

Combining the inequality above with (★★)

$$\mathcal{D}(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon) \leq \sum_{h \in \mathcal{H}} 2e^{-2m\varepsilon^2}$$

$$= 2|\mathcal{H}| e^{-2m\varepsilon^2}$$

By choosing $m \geq \lg\left(\frac{2|\mathcal{H}|}{\delta}\right) \cdot \frac{1}{2\varepsilon^2}$ then

$$\mathcal{D}\left(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}\right) \leq$$

$$\leq 2|\mathcal{H}| e^{-\cancel{2\varepsilon^2} \lg\left(\frac{2|\mathcal{H}|}{\delta}\right) \cdot \frac{1}{\cancel{2\varepsilon^2}}}$$

$$= 2|\mathcal{H}| e^{-\lg\left(\frac{2|\mathcal{H}|}{\delta}\right)}$$

$$= 2|\mathcal{H}| \cdot \frac{\delta}{2|\mathcal{H}|} = \delta$$

■

for example: $m = \left\lceil \lg\left(\frac{2|\mathcal{H}|}{\delta}\right) \frac{1}{2\varepsilon^2} \right\rceil$

Machine Learning

Bias-Complexity Trade-off

Fabio Vandin November 10th, 2023

Our Goal in Learning

Given:

- training set: $S = ((x_1, y_1), \ldots, (x_m, y_m))$
- loss function: $\ell(h, (x, y))$

Want: a function $\hat{h}$ such that $L_{\mathcal{D}}(\hat{h})$ *is small*

We can pick: the learning algorithm A, that given S will produce $\hat{h} = A(S)$

Note: A comprises:

- the hypothesis set $\mathcal{H}$
- the procedure to pick $\hat{h} = A(S)$ from $\mathcal{H}$

Question: is there a *universal learner*, i.e., an (implementable) algorithm A that predicts the best $\hat{h}$ for any distribution $\mathcal{D}$?

The No Free Lunch Theorem

The following answers the previous question for some specific settings.

Theorem (No-Free Lunch)

Let A be any learning algorithm for the task of binary classification with respect to the 0-1 loss over a domain $\mathcal{X}$. Let m be any number smaller than $|\mathcal{X}|/2$, representing a training set size. Then, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that:

- *there exists a function $f\colon \mathcal{X} \to \{0,1\}$ with $L_{\mathcal{D}}(f) = 0$*
- *with probability of at least $1/7$ over the choice of $S \sim \mathcal{D}^m$ we have that $L_{\mathcal{D}}(A(S)) \geq 1/8$.*

Note: there are similar results for other learning tasks.

No Free Lunch and Prior Knowledge

Corollary

Let $\mathcal{X}$ be an infinite domain set and let $\mathcal{H}$ be the set of all functions from $\mathcal{X}$ to $\{0, 1\}$. Then, $\mathcal{H}$ is not PAC learnable.

What's the implication?

We need to use our prior knowledge about $\mathcal{D}$ to pick a *good* hypothesis set.

How do we choose $\mathcal{H}$?

- we would like $\mathcal{H}$ to be *large*, so that it may contain a function h with small $L_{\mathcal{D}}(h)$
- no free lunch $\Rightarrow$ $\mathcal{H}$ cannot be too large!

Error Decomposition

Let h_S be an $\mathrm{ERM}_{\mathcal{H}}$ hypothesis.

Then

$$L_{\mathcal{D}}(h_S) = \epsilon_{\mathrm{app}} + \epsilon_{\mathrm{est}}$$

where

- $\epsilon_{\mathrm{app}} = \min\limits_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$ (approximation error)
- $\epsilon_{\mathrm{est}} = L_{\mathcal{D}}(h_S) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$ (estimation error)

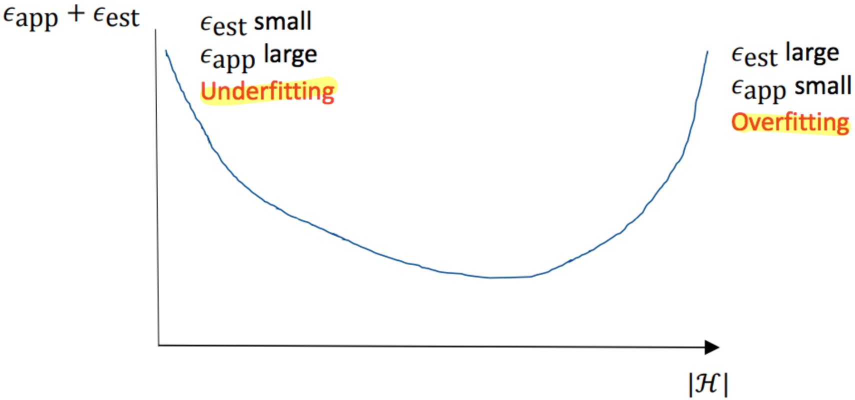
Approximation error: $\epsilon_{\text{app}} = \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$

- derives from our choice of $\mathcal{H}$
- once we have chosen $\mathcal{H} \Rightarrow \epsilon_{\text{app}}$ is unavoidable!
- to decrease it, chose a "larger" $\mathcal{H}$

Estimation error: $\epsilon_{\text{est}} = L_{\mathcal{D}}(h_S) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$

- derives from our inability to choose (with ERM) the best hypothesis in $\mathcal{H}$
- could be avoided if had chosen the best hypothesis!
- to decrease, we need a low number of hypotheses in $\mathcal{H}$ so that training error is good estimate of generalization error for all of them $\Rightarrow$ need a "small" $\mathcal{H}$

Complexity of $\mathcal{H}$ and Error Decomposition



Low complexity
Large inductive bias

High complexity
Small inductive bias

Estimating $L_{\mathcal{D}}(h_S)$

How can we estimate the generalization error $L_{\mathcal{D}}(h)$ for a function h, for example $h_S \in \mathrm{ERM}_{\mathcal{H}}$?

We can use a **test set**: new set of samples not used for picking h_S (=the training set).

Notes:

- the test must not be looked at until we have picked our **final** hypothesis!
- in practice: we have 1 set of samples and we split it in *training set* and *test set*.

Machine Learning

Model Selection and Validation

Fabio Vandin

November 10th, 2023

Model Selection

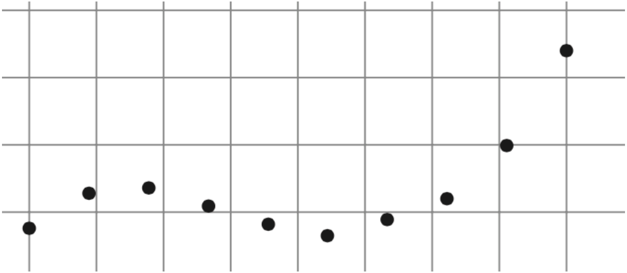
When we have to solve a machine learning task:

- there are different algorithms/classes
- algorithms have parameters

Question: how do we choose a algorithm or value of the parameters?

Example

Regression task, $\mathcal{X} = \mathbb{R}$, $Y = \mathbb{R}$



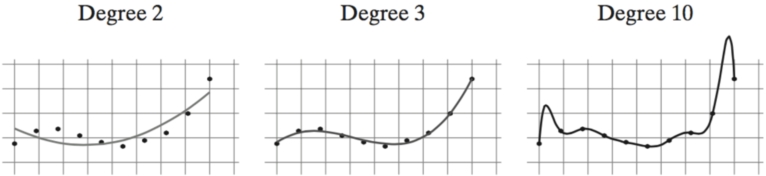
Decision: $\mathcal{H} =$ polynomials.

Note: can be done using the linear regression machinery we have seen!

How do we pick the degree d of the polynomial?

What about considering the empirical risk of best hypothesis of various degrees (e.g., d=2, 3, 10)?

Best hypotheses for degree $d \in \{2, 3, 10\}$



Degree 2

Degree 3

Degree 10

Empirical risk is not enough!

Approach we will consider: validation!