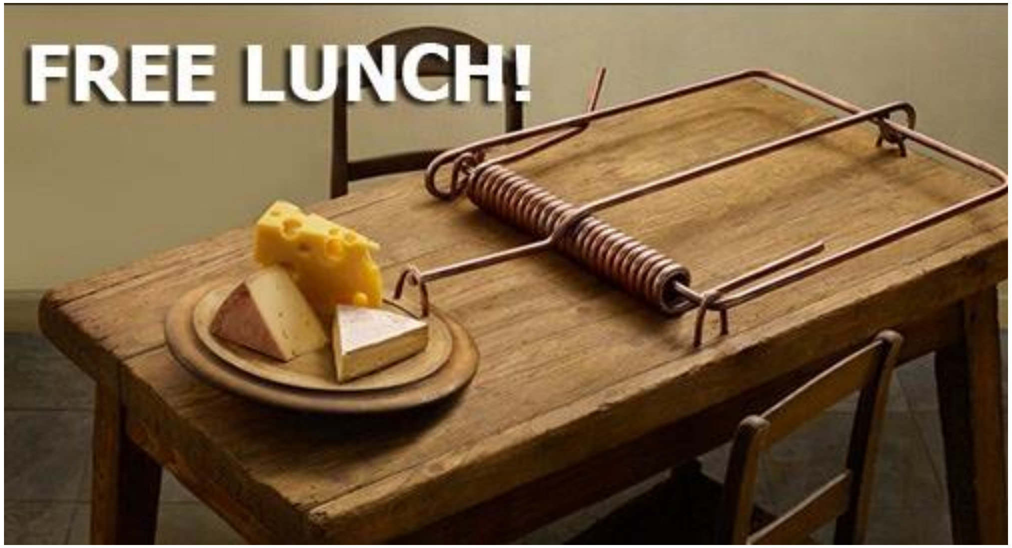
Chapter 5
No Free Lunch



- There are many No-Free-Lunch theorems.

- The one we prove in this chapter only says that there is no universal learner.

- If the hypothesis class is not restricted then there is ALWAYS a distribution that causes the algorithm to overfit (not only ERM!)

- (**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 X. Let m be any number smaller than **|X|/2**, representing a training set size. Then, there exists a distribution D over X ×{0,1} such that:

- There exists a function f:X →{0,1} with $L_D(f) = 0$. (i.e. task can be learned)
- With probability of at least $\frac{1}{7}$ over the choice of $S \sim D^m$ we have that $L_D(A(S)) \geq \frac{1}{8}$. (i.e at least 1/7 chance to have true error > 1/8)

- Lemma:

  Z r.v. in [0,1] with E[Z]=m. Then $\forall a \in (0,1)$

$$\boldsymbol{P}[Z > 1 - a] \geq \frac{m - (1 - a)}{a}$$

- Proof:

  Y=1-Z. Applying Markov we have

$$\boldsymbol{P}[Z > 1 - a] = 1 - \boldsymbol{P}[Y \geq a] \geq 1 - \frac{1 - m}{a}$$

- The above shows that

$$E_{S\sim D^m}[L_D\big(A(S)\big)] \geq \frac{1}{4} \rightarrow \boldsymbol{P}[L_D(A(S)) \geq \frac{1}{8}] \geq \frac{1}{7}$$

- It suffices to prove the below (by Markov)
- $\forall\, A\ \exists D\ such\ that\ E_{S\sim D^m}[L_D(A(S))] \geq \frac{1}{4}$

- In other words every algorithm has a distribution on which it fails ¼ of the time in expectation.
- Intuition:

- Equivalently we want to show

$$\max_{i\in[T]} E_{S\sim D_i{}^m}[L_{D_i}\big(A(S)\big)] \geq \frac{1}{4}$$

- Denote S_j^i the training sequence of size m labeled by the function f_i corresponding to distribution D_i. There are m^{2m} possible training sets that can be sampled at equal probability.
- Therefore expected loss for a fixed i is equal to

$$\frac{1}{k}\sum_{j=1}^{k} L_{D_i}\left(A\left(S_j^i\right)\right)$$

where k = m^{2m}

- $\max_{i\in[T]} \frac{1}{k}\sum_{j=1}^{k} L_{D_i}(A(S_j^i)) \geq \frac{1}{T}\sum_{i=1}^{T} \frac{1}{k}\sum_{j=1}^{k} L_{D_i}\left(A(S_j^i)\right) = \frac{1}{k}\sum_{j=1}^{k} \frac{1}{T}\sum_{i=1}^{T} L_{D_i}\left(A(S_j^i)\right) \geq \min_{j\in[k]} \frac{1}{T}\sum_{i=1}^{T} L_{D_i}\left(A(S_j^i)\right)$ **(I)**

- Let $u_1, u_2, \ldots, u_n$ be the examples not in the training set (so p≥m). The true loss is at least half as much as the loss on the unknown examples.

- $L_{D_i}(h) = \frac{1}{2m} \sum_{x \in C} \mathbf{1}_{[h(u_r) \neq f_i(u_r)]} \geq \frac{1}{2m} \sum_{r=1}^{p} \mathbf{1}_{[h(u_r) \neq f_i(u_r)]} \geq \frac{1}{2p} \sum_{r=1}^{p} \mathbf{1}_{[h(u_r) \neq f_i(u_r)]}$

- Bounding true loss from below and changing summation order

- $\frac{1}{T}\sum_{i=1}^{T} L_{D_i}(A(S_j^i)) \geq$
  $\frac{1}{T}\sum_{i=1}^{T} \frac{1}{2p}\sum_{r=1}^{p} \mathbf{1}_{\left[A\left(S_j^i\right)(u_r) \neq f_i(u_r)\right]} =$
  $\frac{1}{2p}\sum_{r=1}^{p} \frac{1}{T}\sum_{i=1}^{T} \mathbf{1}_{\left[A\left(S_j^i\right)(u_r) \neq f_i(u_r)\right]} \geq$
  $\frac{1}{2} * \min_{r\in[p]} \frac{1}{T}\sum_{i=1}^{T} \mathbf{1}_{\left[A\left(S_j^i\right)(u_r) \neq f_i(u_r)\right]}$ **(II)**

- For a fixed r, all functions $f_i, f_{i'}$ can be paired according to their classification of u_r
- $\mathbf{1}_{[A(S_j^i)(u_r) \neq f_i(u_r)]} + \mathbf{1}_{[A(S_j^{i'})(u_r) \neq f_{i'}(u_r)]} = 1$

$\rightarrow \frac{1}{T}\sum_{i=1}^{T} \mathbf{1}_{[A(S_j^i)(u_r) \neq f_i(u_r)]} = \frac{1}{2}$ **(III)**

(I),(II),(III) $\rightarrow \max_{i \in [T]} E_{S \sim D_i{}^m}[L_{D_i}\left(A(S_j^i)\right)] \geq \frac{1}{4}$

- Corollary:

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

- Proof:

Assume that H is learnable, choosing ε<1/8 and δ<1/7. By PAC definition ∃ A,m=m(ε,δ) such that A given training size ≥m, with P>1-δ, L_{D_i}(A(S))≤ε.

However, by No-Free-Lunch theorem since X>2m (i.e learner knows at most half of universe) ∃ D such that $\boldsymbol{P}[L_D(A(S)) \geq \frac{1}{8}] \geq \frac{1}{7}$. Contradiction.

- To prevent this we must avoid distributions that can deceive us. (i.e increase our bias about the underlying model).
- On the other hand, we need to keep our hypothesis class rich enough to contain the zero error f (or smallest in APAC setting).

- Bias vs Variance.

- In particular, we can decompose the error of an ERM hypothesis $L_D(h_s) = \varepsilon_{app} + \varepsilon_{est}$
- ε_{app}=$\min_{h \in H} L_D(h)$

(bias, price of restricting our class, sample size cant reduce this)

- ε_{est}=$L_D(h_s) - \min_{h \in H} L_D(h)$

(variance, needing more data to train our model)

- The less we restrict the class the more data we need (no restriction=all the data)

- Therefore we need to restrict our class somewhat with educated guesses. Less sacrifices => need more data to counteract estimation error.

- Example:
- Basic Euclidean classification can generalize better (81%) than more sophisticated Naïve Bayes (75%)