3.1. PAC Learning

DEFINITION 3.1 (PAC learnability) A hypothesis class $\mathcal{H}$ is PAC learnable if there exist a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm that: For every $\epsilon, \delta \in (0,1)$, for every distribution D over $\mathcal{X}$ and every labeling function $f : \mathcal{X} \to (0,1)$, and the realizability assumption holds, then the algorithm using $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ returns a hypothesis h such that, with probability $1 - \delta$: $L_{(D,f)}(h) \leq \epsilon$

Accuracy parameter ϵ means how far the output classifier is from the optimal one. Accuracy is about forgiving the classifier's errors, even when the realizability's assumption is met. Confidence paramerer δ means how likely the classifier is to meet the accuracy requirement. It is about getting a representative sample such that the accuracy parameter can be met. Function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ maps accuracy and confidence to complexity, represents how many examples are needed to achieve a probably (δ) aproximately correct (ϵ) solution.

Since there are many functions that satisfy the PAC learnability condition, the interesting one is the minimal function. Now we can impose the upper bound on the previously (chapter 2) derived complexity function

COROLARY 3.2. Every finite hypothesis class is PAC learnable with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\log(|\mathcal{H}|/\delta}{\epsilon} \right\rceil$$

3.2. A More General Learning Model

Relaxing two conditions in PAC: 1. The realizability assumption is strong (it is hard to guarantee that a distribution representative sample is possible to find). 2. Going beyond binary classification

3.2.1 Releasing the Realizability Assumption – Agnostic PAC Learning

In practice, it is not realistic to assume that labels can be mapped from input features 100% of the time, therefore a relaxation is introduced.

D is now a joint probability distribution $\mathcal{X} \times \mathcal{Y}$ over domain points and labels. An alternative perspective is to see this joint distribution as a composition of a *marginal* distribution over domain points D_x and a *conditional* distribution of labels on points $D((x,y)|x)$. This relaxation no longer requires the same instance features to map to the same label.

Now, we redefine true error as (the difference is that we now sample the pair (x, y) instead of x):

$$L_D(h) \doteq \mathbb{P}_{(x,y)\sim D}[h(x) \neq y] \doteq D((x, y) : h(x) \neq y)$$

The empirical risk on sample S stays the same:

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

The goal is to find a hypothesis h that minimizes true risk L_D. The optimal predictor is aware of the data generating function D a will assign labels according to the true distribution:

$$f_D(x) = 1 \rightarrow \mathbb{P}[y = 1|x] \geq 1/2$$
$$f_D(x) = 0 \rightarrow \mathbb{P}[y = 1|x] < 1/2$$

DEFINITION 3.3 (Agnostic PAC Learnability)

A hypothesis class $\mathcal{H}$ is agnostic PAC learnable if for a complexity function $m_{\mathcal{H}} : (0, 1)^2 \rightarrow \mathbb{N}$, accuracy ϵ and realizability $1 - \delta$, exists an algorithm that running on $m \geq m_{\mathcal{H}}$ examples from D returns a hypothesis h such that:

$$L_D(h) \leq min_{h' \in \mathcal{H}} L_D(h') + \epsilon$$

In other words: if there exists a hypothesis with a loss that's optimal and some chosen accuracy ϵ, the hypothesis class of that found hypothesis is considered agnostic PAC learnable.

- TODO why is it not $(1 - \epsilon)$ on the right hand side
- TODO why is the loss using h less or equal, why not equal? (there can't be anything less than optimal h loss and ϵ is arbitrary, my guess is that, this is a math thing)

With respect to non-agnostic PAC learning, we no longer guarantee a perfect prediction, but compare ourselves to some ϵ, making agnostic PAC learning a generalization of PAC learning ($min(L) = 0$)

3.2.2. Scope of Learning Problems Modeled

Multiclass classification and regression can also be modeled as learning tasks. Regression aims to find a functional relationship between $\mathcal{X}$ and $\mathcal{Y}$. Loss can be measured differently for regression problems. One way is the *expected square difference* between the true labels and their predicted values:

$$L_D(h) \doteq \mathbb{E}_{(x,y)\sim D}(h(x) - y)^2$$

Loss functions are functions that map from a hypothesis and domain to a real nonnegative value: $l : \mathcal{H} \times Z \to \mathbb{R}_+$ Risk function is an expected loss of a classifier $h \in \mathcal{H}$. Empirical risk is expected loss over a sample S. Two basic types of loss are 0-1 loss and square loss.

DEFINITION 3.4. (Agnostic PAC Learnability for General Loss Functions)

Same as definition 3.3, but $L_D(h) = \mathbb{E}_{z\sim D}[l(h, z)]$. $l(h, .) : Z- > \mathbb{R}$ is a random variable and $L_D(h)$ is the expected value of that random variable.

- TODO why is loss upper/lower case. What exactly is loss upper case?
- Why is upper case loss an expectation of l, shouldn't it range (0, 1)

Remark 3.2. Representation independent learning. We might be operating on a hypothesis space $\mathcal{H}'$ ($\mathcal{H}$ is a superset of it). If we're searching through $\mathcal{H}$ that is called representation independent learning.