1 Approximate Correctness

- Two languages are ϵ-close under a distribution P iff $P(L_1 \triangle L_2) \leq \epsilon$
  - $w \in L_1 \triangle L_2$ iff $w \in L_1 - L_2$ or $w \in L_2 - L_1$
  - when H is the hypothesis, and T the target of learning, we call $P(H \triangle T)$ is also called the *error* of the hypothesis, and write $\mathsf{error}(H)$ (leaving T and P implicit)
- This means that the chances of randomly encountering a sentence that distinguishes the languages ($w \in L_1 \triangle L_2$) is no greater than ϵ
- If your hypothesis is ϵ-close to the true language, you are *approximately* correct, in the sense that your chances of making an error are no greater than ϵ

2 Probably Correct

- It is always possible to collect an unrepresentative sample
  - Even though getting all heads when flipping a fair coin 10 times is unlikely, it is still possible
    * we expect it to happen $.5^{10} \times 100 \approx 0.1$ percent of the time; i.e. 1 time out of a thousand
- if the sample *is actually* unrepresentative, we will surely make a mistake
- There is no way of knowing whether we got an unrepresentative sample
- The larger the sample, the less likely it is to be unrepresentative
- This means that as your sample grows in size, you grow more confident that it is representative

3 Probably, Approximately Correct

- Assume a space of observables Ω, and a distribution P over these

- Given a *concept* $c \subseteq \Omega$, we write χ_c for the characteristic function of c

$$\chi_c(\omega) := \begin{cases} 1 & \text{if } x \in c \\ 0 & \text{if } x \notin c \end{cases}$$

- We write $\mathsf{EX}(c, P)$ for a random source of elements of Ω, drawn according to P, and labeled according to χ_c
  - so if sampling from Ω we draw $\omega_1, \omega_2, \omega_3$, EX returns $\langle \omega_1, \chi_c(\omega_1) \rangle, \langle \omega_2, \chi_c(\omega_2) \rangle, \langle \omega_3, \chi_c(\omega_3) \rangle$
- $C \subseteq \wp(\Omega)$ is **PAC learnable** iff
  - $\forall c \in C$
  - $\forall P : \Omega \to [0, 1]$
  - $\forall 0 < \epsilon < \frac{1}{2}$ (*margin of error* - **approximately**)
  - $\forall 0 < \delta < \frac{1}{2}$ (*confidence* - **probably**)

  there is some learner ϕ such that $\phi(\mathsf{EX}(c, P), \epsilon, \delta) = h$, where the probability that h is ϵ-close to c under P is at least $1-\delta$ (i.e. P(error(h) $\leq \epsilon$) $\geq$ 1 - δ)
- If $C = \bigcup_{n \in \mathbb{N}} C_n$, we say that C is learnable w.r.t. *size* or *dimension* n iff for all n, C_n is PAC learnable
- If there is a PAC learner ϕ that achieves the PAC success criterion $P(\mathsf{error}(h) \leq \epsilon) \geq 1 - \delta$ for each C_n in time polynomial in $(\frac{1}{\epsilon}, \frac{1}{\delta}, n)$, then we say that C is **efficiently PAC learnable**.

3.1 Example: Monomials

- Consider concepts which can be described using some number of binary features
  - A given concept might be underspecified for a particular feature
- A concept can then be given as a list of *literals*, where a literal for a feature f is either `f` if f should have a value of 1, or $\overline{\texttt{f}}$, if f should have a value of 0
  - if the concept is underspecified for f, neither `f` nor $\overline{\texttt{f}}$ appears
- An example: $\texttt{syl}, \overline{\texttt{cons}}, \texttt{son}$ is the concept of vowels

- There is a natural measure of *size*, namely, how many possible features there are.
- Our instance space is the set of sequences of 1s and 0s of length n
- Our learner begins with the (inconsistent) hypothesis: $x_1 \wedge \overline{x_1} \wedge \ldots \wedge x_n \wedge \overline{x_n}$, where x_1 through x_n are the possible features
- on a positive input $i = b_1 \cdots b_n$, remove literals that contradict the data
  - if $b_i = 1$, then remove $\overline{x_i}$ from the hypothesis
  - if $b_i = 0$, then remove x_i from the hypothesis
- Note that after one positive example the learners current hypothesis is consistent
- Let us now analyze the learner's error.
  - the learner's hypothesis always denotes a subset of the actual concept
  - so the learner will only disagree with the concept on positive examples
  - this manifests itself as the learner having a literal in its hypothesis that shouldn't be there
  - so the learner requires that feature f have a certain value, but actually it doesn't have to
  - if the learner has such an erroneous literal f, it causes h to be wrong only on postive examples in which $f = 0$.[1]
  - We do not care how many of these there are, only how likely we are to see them. Thus we define

$$\texttt{bad}(f) := P(\{a : \chi_c(a) = 1 \wedge f \text{ is } 0 \text{ in } a\})$$

  - As every error of h is due to at least one literal, we have that the total error of h is no greater than the sum of the badness of each of its literals:

$$\mathsf{error}(h) \leq \Sigma_{f \in h} \mathsf{bad}(f)$$

[1] If f is $\texttt{f}$, then $f = 0$ means that there is a 0 in the example. If f is $\overline{\texttt{f}}$, then $f = 0$ means that there is a 1 in the example.

- We want the error of h to be no larger than ϵ, which we can upper bound if $\Sigma_{f \in h}\mathtt{bad}(f) \leq \epsilon$. But there are only ever at most $2n$ literals in a hypothesis.
- so as long as no $\mathtt{bad}(f)$ is greater than $\epsilon/2n$, we have that $\Sigma_{f \in h}\mathtt{bad}(f) \leq 2n(\epsilon/2n) = \epsilon$
- A literal whose badness exceeds $\epsilon/2n$ is **truly bad**. We want to be confident that we do not have any of these.
- Consider a particular truly bad literal *f*.
- Because it is truly bad, the probability that it is removed after seeing a single example is $\mathtt{bad}(f) \geq \epsilon/2n$.
- Thus the probability of **not** removing it after m examples is at most $(1 - \epsilon/2n)^m$
- And so the probability of there being **some** bad literal which is **not** removed after m examples is at most $2n(1 - \epsilon/2n)^m$.
- Now we want to see how large m should be to make $2n(1 - \epsilon/2n)^m \leq \delta$.
  - as $1 - x \leq e^{-x}$, we can choose m so that $2ne^{-m\epsilon/2n} \leq \delta$
  - which yields $m \geq (2n/epsilon)(ln(2n) + ln(1/\delta))$

- Letting $n = 4$, if we want to be 99% sure that our hypothesis is 99% right, we need to draw 17 examples
- Letting $n = 8$, if we want to be 99% sure that our hypothesis is 99% right, we need to draw 45 examples

4 VC Dimension

Given $S \subseteq \Omega$, if

$$\{L \cap S \mid c \in C\} = \wp(S)$$

then S is **shattered** by C.

- The VC dimension of C is the size of the largest set shattered by C

$$\mathsf{VC}(C) = \mathtt{max}\{|S| \mid S \text{ is shattered by } C\}$$

- A class is PAC learnable iff it has finite VC dimension