Homework 2 for Machine learning: Theory and Algorithms

Gabriele Genovese

11 October 2024

1 Exercise 1

Consider the following definition of learnability: A hypothesis class H is learnable if there exist a learning algorithm A with the following property: For every distribution D over X , and for every labeling function $f : X \to \{0, 1\}$, if the realizable assumption holds with respect to H, D, f, then when running the learning algorithm on m i.i.d. examples generated by D and labeled by f, it holds $\lim_{m\to\infty} E_{S\sim D^m}[L_D(A(S))] = 0$.

To prove that the definitions are equivalent, we need to show that PAC-learnability is equivalent to the above definition and vice versa.

Let's first prove that PAC-learnability implies the above definition.

1. Let's first prove that PAC-learnability implies the above definition. Then $\forall \varepsilon, \delta > 0$ there exists a sample S of size m such that $L_D(A(S)) < \varepsilon$ with probability $\geq 1 - \delta$. We can also express this as $\Pr_{S\sim D^m}[L_D(A(S)) \leq \varepsilon] \geq 1 - \delta$. As the semple size m goes to infinity, ε and δ will reach 0 and by using the Markov's inequality we can conclude that $\lim_{m\to\infty} E_{S\sim D^m}[L_D(A(S))] = 0$, which matches the above definition.

Now let's prove that the above definition implies PAC-learnability.

2. Assume there exists a learning algorithm A with the property stated above. So, it holds that $\lim_{m\to\infty} E_{S\sim D^m}[L_D(A(S))] = 0$. Using Markov's inequality, for any probability $\epsilon > 0$, we have $\Pr(L_D(A(S)) \geq \epsilon) \leq \frac{E_{S\sim D^m}[L_D(A(S))]}{\epsilon}$. Since the expectation tends to 0 as the sample size m goes to infinity, we define $\delta > 0$ such that $\Pr(L_D(A(S)) \geq \epsilon) \leq \delta$, which is the definition of PAC-learnability.

2 Exercise 2

Assume, by contradiction, that H is PAC-learnable. By the definition of PAC-learnability, there exists an algorithm A that can produce a learner for the hypothesis class H with an accuracy $L_{D(A(S))} \leq \varepsilon$ with probability greater than $1 - \delta$. Let's choose some $\varepsilon < \frac{1}{8}$ and some $\delta < \frac{1}{7}$. We can now compute $m = m(\varepsilon, \delta)$. Since $\text{VCdim}(H) = +\infty$, for any training set size m, there exists a shattered set of size $2m$. Applying the NFLT, since $|X| > 2m$, for any learning algorithm (in particular for A), there exists a distribution D and a predictor $h \in H$ such that $L_D(h) = 0$, but with a probability of at least $\frac{1}{7}$ over the choice of $S \sim D^m$, we have that $L_D(A(S)) \geq \frac{1}{8}$, which contradicts the initial assumption.