COMP 585 **Spring 2024**

Probabilistic Toolkit for Learning and Computing *Instructor: Maryam Aliakbarpour*

March 26, 2024 *Scriber: Jefferson Hernandez*

Lecture 23

1 Learning Boolean Conjunctions

In this lecture, we revisit the problem of learning conjunctions, when there was no noise in the data. The algorithm started with a hypothesis consisting of a conjunction of all $2n$ literals; thus it begins with a hypothesis that always predicts 0. Then, for every positive example $(a, 1)$, it deletes the literals present in its hypothesis, which causes this example to be classified as negative. As long as sufficiently many examples are used, this algorithm is guaranteed to produce a hypothesis with an error at most ε.

This algorithm does not work when there is noise in the data. For instance, if a negative example was observed with label 1 (due to noise), the algorithm may drop several literals from the hypothesis that are required. The decisions made by the algorithm are not robust as they are based on a single example. We will design a more robust algorithm for learning conjunctions. To begin with, let us continue to assume that the data we receive is noise-free; later, we'll discuss how this more robust algorithm can also be used when the data is noisy.

Let $y^{(i)} = h^\star(x^{(i)})$ be the target conjunction and let h be a literal that appears in y. We will use the notation $h(x^{(i)}) = 1$ to indicate that the literal h evaluates to 1 (true) on the instance $x^{(i)} \in X$. For any literal h that is present in the target conjunction, it holds that $\Pr_{x\sim D}[x_{1,z} = 0 \wedge h^\star(x) = 1] = 0$. We would like to identify all such literals and put them in the output hypothesis. Of course, it is only important to do this for literals that have a significant probability mass of being false under the distribution. Let us make this idea more concrete.

- A literal z is said to be significant if $\Pr_{x\sim D}[x_{1,z} = 0] = \Pr_0(z) \geq \frac{\varepsilon}{8n}$
- A literal z is harmful if $\Pr_{x\sim D}[x_{1,z} = 0 \wedge h^\star(x) = 1] = \Pr_{01}(z) \geq \frac{\varepsilon}{8n}$

We want to prove the following lemma:

Lemma 1.1. *If $\mathcal{H}$ contains all significant but not harmful z's $\implies$ $err(h) \leq \varepsilon$*

Proof. First, notice that all harmful literals are also significant. Let h be a hypothesis that is a conjunction of all literals that are significant, but not harmful. Let us analyze the error of h.

$$
\begin{aligned}
\mathrm{err}(h) &= \Pr_{x\sim D}[h(x) \neq h^\star(x)] \\
&= \Pr_{x\sim D}[h(x) = 0 \wedge h^\star(x) = 1] + \Pr_{x\sim D}[h(x) = 1 \wedge h^\star(x) = 0] \\
&\leq \frac{\varepsilon}{2} = \frac{\varepsilon}{4} + \frac{\varepsilon}{4}
\end{aligned}
$$

where we have set (1) $\Pr_{x\sim D}[h(x) = 0 \wedge h^\star(x) = 1] \leq \frac{\varepsilon}{4}$, and (2) $\Pr_{x\sim D}[h(x) = 1 \wedge h^\star(x) = 0] \leq \frac{\varepsilon}{4}$

To prove (1) notice that

$$\text{Given } h(x) = 0, \exists z \in h \quad \text{s.t.} \quad x_{1,z} = 0 \wedge z \notin h^\star, h^\star(x) = 1$$

$$
\begin{aligned}
\Pr_{x\sim D}[h(x) = 0 \wedge h^\star(x) = 1] &\leq \Pr_{x\sim D}[\exists z \in\notin h^\star x_{1,z} = 0 \wedge h^\star(x) = 1] \\
&\leq \sum_z \Pr_{x\sim D}[x_{1,z} = 0 \wedge h^\star(x) = 1] \\
&\leq \sum_{z\in h\notin h^\star} \Pr_{01}(z) \leq \sum_{z\in h} \Pr_{01}(z) \\
&\leq \frac{\varepsilon}{8n}(2n) \leq \frac{\varepsilon}{4}
\end{aligned}
$$

To prove (2) notice that

$$\exists z \in h^\star \quad \text{s.t.} \quad x_{1,z} = 0 \wedge z \in h^\star \wedge z \notin h$$

This implies that z can not be harmful and since it is removed from h, it implies that z is insignificant

$$
\begin{aligned}
\Pr_{x\sim D}[h(x) = 1 \wedge h^\star(x) = 0] &\leq \Pr_{x\sim D}[\exists z \text{ is on insignificant }, z \in h^\star \text{s.t} x_{1,z} = 0 \wedge h(x) = 1] \\
&\leq \sum_z \Pr[x_{1,z} = 0] \\
&\leq \frac{\varepsilon}{8n}(2n) \leq \frac{\varepsilon}{4}
\end{aligned}
$$

□

2 Statistical Query Model

For a concept class $\mathcal{C}$, we say $\mathcal{C}$ is realizable (and efficiently) learnable in the statistical query model if and only if there exists an algorithm A that for all ε, D, receives ε and makes $\chi_1, \cdots, \chi_t$ queries to the oracle and receives $\hat{P}_{\chi,\tau}$, then algorithm A outputs $\hat{c} \in \mathcal{C}$ such that $\mathrm{err}(\hat{c}) \leq \varepsilon$. A statistical query is a tuple, (χ, τ), where $\chi : X \times 0, 1 \to 0, 1$ is a boolean function that takes as input an instance $x \in X$ and a target $y \in 0, 1$ (one of the two possible labels of the instance), and τ is the tolerance parameter.

For the next lecture, we will prove that learning an algorithm in the statistical query model implies the learnability in the PAC setting with noise.