L4: Hardness of Estimation: Learning Theory

Data X, y of size n so each $(x_i, y_i) \in \mathbf{R}^d \times \{-1, +1\}$

$X \sim P$ for probability distribution on $\mathbb{R}^d$ and $y \sim \sigma$.
$X, y \sim P, \sigma$ a *joint* distribution, y_i depends on x_i (may be functional relation).

Goal is to learn $f : \mathbb{R}^d \to \mathbb{R}$ so if $f(x) > 0$, predict $+1$, and otherwise predict -1.
Want $y \approx \mathsf{sign}(f(X))$

Function class F (e.g., family of all halfspaces, or neural net with fixed architecture).
$h_{u,b} \in H$ (halfspaces), we can define $h_{u,b}(x) =< x, u > -b$

Goal 1: Find $f \in F$ on (X, y) so that

$$err(f, X, y) = \frac{1}{n} \sum_i (\mathsf{sign}(f(x_i)) \neq y_i)$$

is as small as possible.

But this only works with existing data (X, y).
The real goal is to understand P, σ, and potential new data drawn again from there.

$$err(f, P, \sigma) = E_{(x,y)\sim(P,\sigma)}(\mathsf{sign}(f(x_i)) \neq y_i)$$

Sample Complexity for Learning Bounds

Separable Data:
Assume first there exists some $h \in H$ so that $err(h, P, \sigma) = 0$.
Let $h \in H$ satisfies $err(h, X, y) = 0$.

Let $n = \Omega((\nu/\varepsilon)\log(\nu/\varepsilon\delta))$ for $\varepsilon, \delta \in (0, 1)$; we will explain ν later.
Then with probability at least $1 - \delta$, $err(h, P, \sigma) \leq \varepsilon$.

Non-Separable Data:
Let $h \in H$ satisfies $err(h, X, y) = \gamma$.

Let $n = \Omega((1/\varepsilon^2)(\nu + \log(1/\delta))$ for $\varepsilon, \delta \in (0, 1)$
Then with probability at least $1 - \delta$, $err(h, P, \sigma) \leq \gamma + \varepsilon$.

VC (Vapnik-Chervonenkis) Dimension

Let (X, F) be a *range space*, where (in this class) $X \subset \mathbb{R}^d$, and F provides a family of subsets of X (e.g., H, those defined by inclusion in a halfspace).

We say a range space (Y, F) for $Y \subset X$, can be *shattered* if all subsets of Y exist.
That is, for each $Z \subset Y$, there exists some "shape" $f \in F$ so the $f \cap Y = Z$.

Any subset of size 3 points in the $\mathbb{R}^2$ can be shattered by halfspaces (unless they are co-linear). But no set of 4 points can be shattered by halfspaces.

The **VC-dimension** of a range space (X, F) is the size of the largest subset $Y \subset X$ which can be shattered.
For *halfspaces* in $\mathbb{R}^d$, the VC-dimension is $d + 1$.

More generally, let (X, F) for $X \subset \mathbb{R}^d$ and F be a family of functions (e.g., a neural net) which can be evaluated with t *simple operations* with the follow structure:

- +, -, x, /
- jumps using <, <=, >, >=, =, != on real numbers
- return 0, 1

Then the VC-dimension of (X, F) is at most $4d(t+2)$.

If you also allow $q > 1$ exponential $\exp(\cdot)$ operations in functions in F
then the VC-dimension of (X, F) is $O(d(q^2 + q(t + \log(dq)))$

Take-away: the number of samples needed to generalize grows linearly (if not quadratically) with dimension d.

Which Function Class?

So are simpler (lower VC-dim) function classes better?

If we only use halfspaces, on the first 3 coordinates, then we get better generalization with same samples, right?
Then only need $n = O((1/\varepsilon^2)(4 + \log(1/\delta))$.
–> But $\gamma = err(h, X, y)$ is larger!

Let the model error

$$\gamma_F = \min_{f \in F} err(f, P, \sigma)$$

be the minimal amount of error from a function class F. Simple classifiers tend to have larger γ_F.
Complicated (high-dimensional) classifiers have smaller γ_F.

- If $d = n$, then for halfspaces $\gamma_H = 0$. Since we can shatter X.
- In general for (X, F) with VC-dimension ν if $n = \nu$, we might be able to shatter X, in which case $\gamma_F = 0$.
- For H_p described as polynomials of sufficiently degree p, it can approximate any function f. But VC dimension $O(d^p)$.
- Even 2-layer neural networks with sufficiently wide second layer, can also approximate any function.

But then if $n \approx \nu$, it does not satisfy $n = \Omega(\nu/\varepsilon^2)$, so do not get sample complexity bound, and $|err(f, X, y) - err(f, P, \sigma)|$ can be large – it is not controlled.