VC-Dimension and Sampling

1 Sampling

VC-dimension and sampling are combinatorial properties of set systems that can let us do better than a union bound. Recall that an event $X \subseteq \Omega$ is a subset of the sample space Ω. If an event X is sufficiently likely and we draw enough samples, we are likely to hit the event.

Proposition 1. *(Simple sampling). Let $X \subseteq \Omega$ with $\Pr[X] \geq \varepsilon$, and let S be obtained by drawing a set of $\varepsilon^{-1} \ln \delta^{-1}$ samples from Ω independently. Then*

$$\Pr[S \cap X] \geq 1 - \delta.$$

In other words, repeating the experiment enough times boosts the probability of it succeeding at least once.

Proof. The probability of never hitting X is at most $(1-\varepsilon)^{\varepsilon^{-1} \ln \delta^{-1}} \leq e^{-\ln \delta} = \delta$. □

Is it true that all high probability events are likely to happen? Draw a set S of some number of points (small-ish, like $\varepsilon^{-1} \ln \delta^{-1}$). Is it true that with some good probability (*e.g.,* $1 - \delta$), for every set X with $\Pr[X] \geq \varepsilon$, we have $S \cap X \neq \emptyset$?

This is actually false! Given S, define $X = \bar{S}$. Then $X \cap S = \emptyset$, but $\Pr[X] \geq \varepsilon$ unless we drew more than a $1 - \varepsilon$ fraction of the sample space.

Typically, we go from sampling to the universal statement we wanted via a union bound. If we care about too many events, we only get trivial bounds; if we only care about a small set R of events, then an additive term of $\sim \varepsilon^{-1} \ln R$ in the number of samples we draw is enough to enable a union bound.

Our goal is now to side-step the crude union bound by exploiting some structure in R. Throughout, $R \subseteq \mathcal{P}(\Omega)$ is a collection of events $A \subseteq \Omega$ (*i.e.,* a set system).

2 Shattering and VC-Dimension

Definition 2. A set $X \subseteq \Omega$ is shattered by R iff for every set $Y \subseteq X$, there exists an $A \in R$ with $Y = X \cap A$. That is, every subset of X can be obtained by using a set in R to "clip off" a part of X.

Example 3. Let Ω be the unit square $[0,1]^2$ and R the set of all closed halfspaces. (i) The set of three points is shattered because every subset of points can be separated from its complement by a line. (ii) The set of four points in a square is not shattered because there is no line that can separate the diagonals. (iii) The set of four points in a star is not shattered because there is no line that can separate the middle point.

Definition 4. The VC (Vapnik-Chervonenkis) dimension of R is the maximum size of any set X that is shattered by R. This is a natural measure of the complexity of R). The VC-dimension could be unbounded if the underlying set Ω is infinite. To show VC-dimension equals k, we have to show at least one shatterable subset of size k, then show that no subset of size $k+1$ is shattereable.

Example 5. Let Ω be the unit square $[0,1]^d$ and R the set of all closed halfspaces. (i) $d = 1$: we can shatter 2 points but not 3, so the VC-dimension is 2. (ii) $d = 2$: we showed we can shatter 3 points but not 4, so the VC-dimension is 3. (iii) any d: we can easily construct hyperplanes to shatter a set of $d + 1$ points in general position. To show we cannot shatter $d + 2$ points, we need the following:

Theorem 6. *[Radon]. Let $S = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{d+2}\}$ in $\mathbb{R}^d$. Then S can be partitioned as $S = X \cup Y$ with $X \cap Y = \emptyset$ such that $conv(X) \cap conv(Y) \neq \emptyset$.*

From this theorem we see that no hyperplane can separate X from Y, because with X it would contain $conv(X)$, but the intersection is nonempty so it must also contain some point in Y. So halfspaces in $\mathbb{R}^d$ have a VC-dimension of $d + 1$.

Definition 7. The restriction of R to $\Omega' \subseteq \Omega$ is

$$R\,|_{\Omega'} = \{A \cap \Omega' : A \in R\}.$$

It is obvious that the VC-dimension of $(\Omega', R\,|_{\Omega'})$ is at most (Ω, R).

3 ε-nets and ε-samples

Definition 8. For a set system (Ω, R), a set $N \subseteq \Omega$ is called an ε-net iff $N \cap A \neq \emptyset$ for all $A \in R$ with $\Pr[A] \geq \varepsilon$. In other words, an ε-net hits all sufficiently large (*w.r.t.* Pr) sets. Note: there are other definitions of ε-nets, *e.g.* containing a point close to every input point.

So, an ε-net would be a good set of samples in the sense that it hits all high-probability events A that we care about.

Theorem 9. *[ε-net theorem]. Let R be a set system over Ω with VC-dimension d. Let S be obtained by drawing*

$$\Theta\big(\frac{d}{\varepsilon}\log\frac{d}{\varepsilon} + \frac{1}{\varepsilon}\log\frac{1}{\delta}\big)$$

samples from Ω i.i.d. according to Pr. *Then S is an ε-net for (Ω, R) with probability at least $1 - \delta$. Notice that this bound is very good when we restrict both d and ε.*

Example 10. Consider n points in the plane. We are interested in hitting all halfspaces that contain at least εn points. That is, we want to color a small number of points red such that, with high probability, every halfspace containing enough points also contains at least one red point. How many red points are needed? Because $d = 3$, if we think of ε, δ as constants, $\mathcal{O}(1)$ red points is enough, regardless of n. By drawing $\mathcal{O}(1)$ points, with probability $\geq 90\%$, we hit every set separable by a line that contains at least 20% of the points – pretty remarkable!

Example 11. Consider $R = \mathcal{P}(\Omega)$ with $|\Omega| = n$. How small is the smallest ε-net for (Ω, R)? We need to intersect all sets of size at least εn. Well, any ε-net must have $|N| > 1 - \varepsilon n$, otherwise $\bar{N}$ would be a set of size greater than εn and would not intersect n. We get nothing useful because the VC-dimension of R is n.

The other key notion besides an ε-net is an ε-sample.

Definition 12. A set $S \subseteq \Omega$ is an ε-sample for (Ω, R) iff

$$|\Pr[A] - \frac{\Pr[S \cap A]}{\Pr[S]}| < \varepsilon$$

for all $A \in R$. Intuitively, an ε-sample contains, for each $A \in R$, up to additive error ε, the right fraction of samples compared to the actual fraction of probability mass in A. In other words, the sample is representative.

Proposition 13. *If S is an ε-sample for (Ω, R) then S is an ε-net for (Ω, R).*

Proof. Let $A \in R$ be "large", with $\Pr[A] \geq \varepsilon$. We need to show $A \cap S \neq \emptyset$. If $S \cap A = \emptyset$, then $\Pr[S \cap A] = 0$, so

$$|\Pr[A] - \frac{\Pr[S \cap A]}{\Pr[S]}| = \Pr[A] \geq \varepsilon$$

so S would not be an ε-sample. □

Theorem 14. *[ε-sample theorem]. Let (Ω, R) be a set system of VC dimension d. Let S be a set of*

$$\Omega(\frac{d^2}{\varepsilon^2} \log \frac{d}{\varepsilon} + \frac{1}{\varepsilon^2} \log \frac{1}{\delta})$$

points drawn i.i.d. from Ω according to $\Pr$. Then S is an ε-sample with probability at least $1 - \delta$.