36-705 (L3)

Supplementary review (lecture Notes 3)

(*) To invoke and make argument that $P(|\hat{F}_n(t) - F(t)| > \epsilon)$; i.e. state absolute approximation error, Hoeffding's inequality is used.

- Recall Hoeffding requires IID observations $Y_1, \ldots, Y_n$ with $\mathbb{E}[Y_i] = \mu$ and that each r.v. is bounded $a \leq Y_i \leq b$.

- For each t and for any $\epsilon > 0$:-

$$P(|F_n(t) - F(t)| > \epsilon) \leq 2e^{-2n\epsilon^2/(b-a)^2}$$

- LW: Argue that $F_n(t)$ is equivalent to the sample average $\bar{Y}_n$ IID. of Bernoulli r.vs. ; with $0 \leq Y_i \leq 1$ and $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$

- Hence $(b-a)^2 = 1$

$$\Rightarrow P(|F_n(t) - F(t)| > \epsilon) \leq 2e^{-2n\epsilon^2}$$ (empirical CDF)

(*) Can you see intuitively why $F_n(t) = \bar{Y}_n$ for Bernoulli r.vs.

(*) use indicator notation:-
- can see how it works, but it still does not quite 'fit together' for me

- (!)(1) → overspill → come back to this → WIKI on empirical CDF or review Wasserman

(*) Going from pointwise → uniform convergence

(*) Have printed out some WIKI articles, as alternative source of exp.

(!) I am more comfortable with pointwise/uniform convergence distinction in context of Example 3 (class.); Example 2 (prob); than in context of Example 1 (empirical CDF)

- (!)(2) - Example 1 in context of understanding pointwise; uniform convergence distinction

- Lecture presents uniform convergence through emphasising distinction between approximation error and maximum approx. error (for emp CDF.)

(*) yields Vapnik-Chervonenkis theorem

- states a uniform convergence probability bound

(*) next part motivating VC dimension is thinking how the rate of increase/decrease in shattering coefficient $s_n(A)$ and exponential term $e^{-n\epsilon^2/32}$ offsets each other in n.

(Q) intuitive interpretation of n (?)

(*) VC dimension: $d = d(A) := $ largest n : $s_n(A) = 2^n$

- d is size of largest set that can be shattered i.e. $\overset{n:}{S(A,F)} = 2^n$

(*) Sauer's theorem is a comment on how the shattering coefficient $s_n(A)$ increases before and after the critical point given by the VC-dimension.
(complexity of class of sets A)

- $s_n(A)$ increases exponentially for $n<d$
  polynomially for $n>d$

(*) list of key points in lecture

36-705 10/05/2020

Lecture Notes 3 - Uniform Bounds (supplement)

Q: Why are events $B_1, B_2 \ldots$ NOT disjoint? A1 → stackexchange

- Note $B_j = \{|P_n(A_j) - P(A_j)| > \epsilon\}$

W: I am comfortable with this reasoning that

i) As B_j not disjoint, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) \neq \sum_{i=1}^{\infty} P(A_i)$$

(*) Union bound (covered in class notes)

- For any finite/countable set of events $B = \{B_1, \ldots, B_k\}$; the probability that at least one of the events happens is no greater than the sum of probabilities of individual events

- i.e. for $B = \{B_1, \ldots, B_k\}$

$$P\left(\bigcup_i B_i\right) \leq \sum_i P(B_i)$$

(*) comfortable with how this is invoked to prove uniform convergence. -

$$P\left(\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| > \epsilon\right) \leq \sum_{j=1}^{N} P(|P_n(A_j) - P(A_j)| > \epsilon) \underset{(i)}{=} 2Ne^{-2n\epsilon^2}$$

$$\Rightarrow P\left(\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| > \epsilon\right) \leq 2Ne^{-2n\epsilon^2}$$

(i) Each $P(|P_n(A_j) - P(A_j)| > \epsilon) \leq 2e^{-2n\epsilon^2}$ via Hoeffding

---

(*) Extension to classes of sets of infinite cardinality

- Invoke additional tools
- For the infinite class of sets in question $\mathcal{A} = \{A_1, A_2, \ldots\}$
  - Define an arbitrary finite set $F = \{x_1, x_2, \ldots\}$
  - Define a subset of this finite set $G \subset F$

· Then say the class $\mathcal{A}$ picks out G if:-

$A \cap F = G$ for some $A \in \mathcal{A}$

(*) There will an exhaustive list of all possible subsets of my finite set F.

(*) I can set my A anyway I like, within the constraints defined by the class $\mathcal{A}$ of which A is a member.

(*) I then take the intersection of A with F to try and get as many possible subsets of F as possible.

(*) each of the subsets G are the subsets which I can get by intersecting A with F ; $A \cap F = G$

- and we say $\mathcal{A}$ picks out G if $A \cap F = G$ for some $A \in \mathcal{A}$

(*) The maximal number of subsets of F is 2^n (powersets)

(*) often, the maximal number of subsets of F will not be the same as the number of subsets G that $\mathcal{A}$ picks out

(*) we define the number of subsets picked out by collection/class $\mathcal{A}$ as $s(\mathcal{A}, F)$

(*) $s(\mathcal{A}, F) \leq 2^n$ for a finite set F of size n $|F| = n$

(*) The shattering coefficient :- (measure of complexity of class of (infinite) sets $\mathcal{A}$)

$$s_n(\mathcal{A}) = \sup_{F \in \mathcal{F}_n} s(\mathcal{A}, F)$$

(*) need to repeat above for all finite sets of size n (what is the algorithmic way to do this?)

(*) And the finite set F is shattered if $s(\mathcal{A}, F) = 2^n$ where n is no. of points in F.

(*) in context of earlier; this occurs when for a particular finite set of size n; we can choose our A such that our class $\mathcal{A}$ picks out every possible subset of F.