

PAC Learning and VC Classes

Stephan Winkler

April 11, 2005

Introduction to Learning Theory

- Feature space $F \subseteq \mathbb{R}^d$
- Concept class \mathcal{C} of subsets of F , e.g. half-spaces
- An example $(x, y) \in F \times \{\pm 1\}$ is called consistent with $C \in \mathcal{C}$ if: $x \in C$ iff $y = +1$
- Training examples $\{(x_i, y_i) : i = 1 : n\}$ consistent with some unknown concept $C \in \mathcal{C}$
- Problem: Find a concept $C' \in \mathcal{C}$, aka hypothesis, that is consistent with the training examples

PAC Learning

- Many concepts may be consistent with training data, so we can only hope to learn the concept approximately.
- Probably Approximately Correct (PAC) learning: Assume X_1, \dots, X_n iid $\sim P$ for some unknown distribution P , each labeled consistently with some unknown concept $C \in \mathcal{C}$, learn concept $C' = C'(\mathbf{X}) \in \mathcal{C}$ such that

$$\mathbb{P}\{P(C \Delta C') \leq \epsilon\} \geq \delta \quad (1)$$

- Question: For particular concept class \mathcal{C} , how many examples do we need so that every consistent C' satisfies equation 1?

Vapnik-Chervonenkis

- Empirical distribution P_n is "random" probability measure on F that puts mass $1/n$ at each X_i :

$$P_n(C) = P_{n,\omega}(C) = \frac{1}{n} \sum_{i \leq n} \{X_i(\omega) \in C\}$$

Theorem. (*Vapnik-Chervonenkis*) Let P be any probability distribution on the features space $F = \mathcal{R}^d$, let \mathcal{C} be the concept class of closed half-spaces, and let $\epsilon, \delta > 0$ arbitrary. Then for a suitably large sample X_1, \dots, X_n iid $\sim P$

$$\mathbb{P}\{\sup_{C \in \mathcal{C}} |P(C) - P_n(C)| \geq \epsilon\} \leq \delta.$$

Uniform Strong Law of Large Numbers

- Note $nP_n(C) \sim \text{Bin}(n, P(C))$, so that by the usual SLLN, $|P_n(C) - P(C)| \rightarrow 0$ a.s. for each fixed C .
- We will prove the much stronger result that for sufficiently large n , $\sup_{C \in \mathcal{C}} |P_n(C) - P(C)| \rightarrow 0$ a.s., which implies VC.
- Strategy: Establish exponential decay for the tail probability $\mathbb{P}\{\sup_{C \in \mathcal{C}} |P_n(C) - P(C)| \geq \epsilon\} \rightarrow 0$ and apply Borel-Cantelli.
- Idea: Symmetrize expression by replacing P by an independent copy P'_n of P_n , based on a second, independent sample $X'_1, \dots, X'_n \sim P$.

Symmetrization Proof (Vapnik-Chervonenkis):

- Observe that $P(C) = \mathbb{P}(P'_n(C)) = \mathbb{P}_{\mathbf{X}}(P'_n(C))$. Then

$$\begin{aligned} |P_n(C) - P(C)| &= |P_n(C) - \mathbb{P}_{\mathbf{X}}(P'_n(C))| \\ &\leq \mathbb{P}_{\mathbf{X}}|P_n(C) - P'_n(C)| \end{aligned}$$

- Take $\sup_{C \in \mathcal{C}}$ inside the expectation on the RHS:

$$|P_n(C) - P(C)| \leq \mathbb{P}_{\mathbf{X}} \sup_{C \in \mathcal{C}} |P_n(C) - P'_n(C)|$$

- Finally take $\mathbb{P} \sup_{C \in \mathcal{C}}$ of both sides

$$\begin{aligned} \mathbb{P} \sup_{C \in \mathcal{C}} |P_n(C) - P(C)| &\leq \mathbb{P}(\mathbb{P}_{\mathbf{X}} \sup_{C \in \mathcal{C}} |P_n(C) - P'_n(C)|) \\ &= \mathbb{P} \sup_{C \in \mathcal{C}} |P_n(C) - P'_n(C)| \end{aligned}$$

Reduction to a simple stochastic process (1/2)

- Observe that

$$P_n(C) - P'_n(C) = \frac{1}{n} \sum_{i \leq n} (\{X_i \in C\} - \{X'_i \in C\})$$

has the same distribution as

$$\frac{1}{n} \sum_{i \leq n} \sigma_i(\{X_i \in C\} - \{X'_i \in C\}).$$

Hence

$$\begin{aligned} & \mathbb{P} \sup_{C \in \mathcal{C}} |P_n(C) - P(C)| \\ & \leq \mathbb{P} \sup_{C \in \mathcal{C}} \left| \frac{1}{n} \sum_{i \leq n} \sigma_i(\{X_i \in C\} - \{X'_i \in C\}) \right| \\ & \leq 2\mathbb{P}(\mathbb{P}_{\mathbf{X}} \sup_{C \in \mathcal{C}} \left| \frac{1}{n} \sum_{i \leq n} \sigma_i \{X_i \in C\} \right|) \end{aligned}$$

Reduction to a simple stochastic process (1/2)

- Conditional on the set of points \mathbf{X} , we may identify $C \in \mathcal{C}$ with vertices $\mathbf{h} \in \mathbb{H}$ of the discrete hypercube $\{0,1\}^n$ via $h_i = \{X_i \in C\}$.
- Key idea: Conditional on \mathbf{X} , we may focus on the study of the process $Z_{\mathbf{h}} = \sum_{i \leq n} \sigma_i h_i$, indexed by $\mathbf{h} \in \mathbb{H}$.
- We are reduced to bounding

$$\mathbb{P}_{\mathbf{X}} \left\{ \sup_{\mathbf{h} \in \mathbb{H}} |Z_{\mathbf{h}}| \geq n\epsilon \right\}$$

Hoeffding tail bound for subgaussian random variables

- The process $\{Z_{\mathbf{h}} : \mathbf{h} \in \mathbb{H}\}$ is subgaussian with scale factors $\sigma_{\mathbf{h}}^2 = \sum_{i \leq n} h_i^2$.

$$\mathbb{P}_{\mathbf{X}} e^{tZ_{\mathbf{h}}} = \prod_{i \leq n} \left(\frac{1}{2} e^{th_i} - \frac{1}{2} e^{-th_i} \right) \leq \prod_{i \leq n} e^{\frac{1}{2} t^2 h_i^2} = e^{\frac{1}{2} t^2 \sigma_{\mathbf{h}}^2}$$

- We thus obtain the tail bound

$$\mathbb{P}_{\mathbf{X}} \{Z_{\mathbf{h}} \geq n\epsilon\} \leq e^{-tn\epsilon} \mathbb{P} e^{tZ_{\mathbf{h}}} \leq e^{-tn\epsilon} e^{\frac{1}{2} t^2 \sigma_{\mathbf{h}}^2}$$

- Choosing $t = n\epsilon/\sigma_{\mathbf{h}}^2$, and noting that $\sigma_{\mathbf{h}}^2 \leq n$, we get

$$\mathbb{P}_{\mathbf{X}} \{Z_{\mathbf{h}} \geq n\epsilon\} \leq e^{-\frac{1}{2} n\epsilon^2}.$$

Putting things together

- Let $N = \#\mathbb{H}$ and bound the supremum over $\mathbf{h} \in \mathbb{H}$ by the sum over all such terms, we obtain

$$\begin{aligned} \mathbb{P}\{\sup_{C \in \mathcal{C}} |P_n(C) - P(C)| \geq \epsilon\} &\leq 2\mathbb{P}\left(\sum_{\mathbf{h} \in \mathbb{H}} \mathbb{P}_{\mathbf{X}}\{|Z_{\mathbf{h}}| \geq n\epsilon\}\right) \\ &\leq 2Ne^{-\frac{1}{2}n\epsilon^2}. \end{aligned}$$

- Observe the trivial bound $N \leq 2^n$. We would be in bad shape if N could be exponentially large. Fortunately, we can show for many concept classes \mathcal{C} that N grows only polynomially in n . Such classes are called VC classes.

A simple CV class: closed half-spaces

Theorem. *Let $\mathcal{C} = \mathcal{H}_d$ be the collection of closed half spaces in $F = \mathbb{R}^d$, and let P a set of n points in \mathbb{R}^d . Then the number N of subsets of the form $P \cap H$, with $H \in \mathcal{H}_d$, is $N \leq \sum_{i=0}^d \binom{n}{i} \leq (d+1)n^d$.*

- Consider any $H \in \mathcal{H}_d$. Let i be the affine dimension of $P \cap H$. Then we may perturb H so that i affinely independent points of P lie on the boundary of H without changing $P \cap H$.
- To enumerate all subsets of P of the form $P \cap H$, pick each subset R of P of cardinality at most i , $i = 1, \dots, d$, and then if the points in R are affinely independent, take any hyperplane through them.

Towards shatter dimension

- Develop general framework for counting arguments that lead to bounds on the number of subsets $P \cap C$ picked out by concepts $C \in \mathcal{C}$ from a finite set $P = \{x_1, \dots, x_n\} \subset F$.
- Given $\mathbf{x} = (x_1, \dots, x_n)$, a pattern is specified by a subset \mathbb{J} of $1 : n$ and a concept $C_{\mathbb{J}}$ from \mathcal{C} for which $x_j \in C_{\mathbb{J}}$ iff $j \in \mathbb{J}$. Of course $C_{\mathbb{J}}$ need not be unique.
- Identify a concept class \mathcal{C} that picks out N distinct patterns from \mathbf{x} with an $N \times n$ binary matrix $V = V_{\mathbf{x}, \mathcal{C}}$.

Shatter dimension and VC classes of sets

- A nonempty subset \mathbb{J} of $1 : n$ of cardinality $k = \#\mathbb{J}$ is shattered if all 2^k possible k -tuples of 0's and 1's appear at least once as a row of $V[, \mathbb{J}]$.
- The shatter dimension, $\text{sd}(V)$, of V is the largest k for which for which there is a shattered \mathbb{J} with $\#\mathbb{J} = k$. If V equals $V_{\mathbf{x}, \mathcal{C}}$, the matrix indicating which patterns a concept class \mathcal{C} picks out from \mathbf{x} , say \mathcal{C} shatters $(x_j : j \in \mathbb{J})$ if V shatters \mathbb{J} , and write $\text{sd}(\mathbf{x}, \mathcal{C}) = \text{sd}(V)$.
- Define the VC dimension, $\text{vcd}(\mathcal{C})$, of \mathcal{C} as the supremum of $\text{sd}(\mathbf{x}, \mathcal{C})$ over all $\mathbf{x} \in F$. Call \mathcal{C} a VC class of sets if $\text{vcd}(\mathcal{C}) < \infty$.

The VC lemma for binary matrices (1/2)

Theorem. *Let V be an $N \times n$ matrix. If $\text{sd}(V) \leq l$, then*

$$N \leq \sum_{i=0}^l \binom{n}{i}.$$

Proof:

- I will establish the contrapositive, showing that if $N > \sum_{i=0}^l \binom{n}{i}$, then $\text{sd}(V) > l$.
- Define downshift for j -th column: for $i = 1 : N$, if $V[i, j] = 1$, change it to a 0 unless the resulting matrix V' would no longer have distinct rows.

The VC lemma for binary matrices (2/2)

- Observe that no new shattered sets of columns can be created by a downshift. (Assume V' shatters \mathbb{J} , and show that V also shatters \mathbb{J} .)
- When no more downshifting is possible, the final matrix has more rows than the $\sum_{i=0}^d \binom{n}{i}$ that could be created by allowing d or fewer 1's per row.
- So there must be a row with 1's in columns \mathbb{J} with $\#\mathbb{J} > d$. Since each possible downshift is blocked, the final matrix must shatter \mathbb{J} .

Summary

- The proof of the VC Theorem stated at the very beginning does not use any explicit characteristics of the feature space F or concept class \mathcal{C} .
- What is needed is a polynomial bound on the number of subsets N of a given set of n points in feature space that can be picked out by concepts in \mathcal{C} .
- If the concept class \mathcal{C} has finite VC dimension l , by the VC Lemma, this number of subsets N is bounded by $N \leq \sum_{i=0}^l \binom{n}{i}$, as required.