Theoretical Statistics, STOR 655

Subadditivity

Andrew Nobel

February 2022

Fekete's Lemma

Definition: A sequence $a_1, a_2, \ldots \in \mathbb{R}$ is *subadditive* if $a_{m+n} \leq a_m + a_n$ for every $m, n \geq 1$

Fact: If the sequence $a_1, a_2, \ldots$ is subadditive then

$$\lim_{n\to\infty} \frac{a_n}{n} \text{ exists and is equal to } \inf_{n\geq 1} \frac{a_n}{n}$$

Note: The limit may be $-\infty$, is non-negative if $a_1, a_2, \ldots$ are non-negative

Stationary Sequences

Recall: A random sequence $X_1, X_2, \ldots \in \mathcal{X}$ is *stationary* if for all $k, j \geq 1$,

$$(X_1, \ldots, X_k) \stackrel{\mathsf{d}}{=} (X_{j+1}, \ldots, X_{j+k})$$

In other words, the distribution of k-blocks is invariant under shifts

Application: Averages of Stationary Sequences

Example: Let $X_1, X_2, \ldots \in \mathcal{X}$ be stationary and let $\mathcal{F}$ be a family of functions $f : \mathcal{X} \to [-1, 1]$. Then

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f(X) \right| \right]$$

converges to a non-negative constant. If the constant is zero, the family $\mathcal{F}$ satisfies a uniform law of large numbers with respect to $\{X_i\}$

Example: If $X_1, X_2, \ldots \in \mathbb{R}$ is stationary with $\mathbb{E}|X_i|$ finite, then

$$\mathbb{E}\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mathbb{E} X \right|$$

converges to a non-negative constant. The constant is zero if the usual LLN holds in expectation

Application: Bin Packing

For $n \geq 1$ define $f_n : [0,1]^n \to \mathbb{N}$ by $f_n(x_1^n) =$ minimum number of size 1 bins needed to hold n objects of size $x_1, \ldots, x_n$. For example

- $f_3(3/4, 3/4, 2/3) = 3$
- $f_3(1/2, 1/3, 1/4) = 2$

Note that $0 \leq f_n \leq n$ and that

$$f_{m+n}(x_1^{m+n}) \leq f_m(x_1^m) + f_n(x_{m+1}^{m+n})$$

Fact: If $X_1, X_2, \ldots \in [0,1]$ is stationary, then the expected number of bins per item converges, that is

$$\frac{\mathbb{E} f_n(X_1^n)}{n} \to \inf_k \frac{\mathbb{E} f_k(X_1^k)}{k}$$

Induced Subsets

Setting: Set $\mathcal{X}$ and a family $\mathcal{A} \subseteq 2^{\mathcal{X}}$ of subsets of $\mathcal{X}$

Let points $x_1, \ldots, x_n \in \mathcal{X}$ be given. Every set $A \in \mathcal{A}$ induces a subset

$$A \cap \{x_1, \ldots, x_n\}$$

of the points, namely, the set of x_i that belong to A

Question: How many subsets of $x_1, \ldots, x_n$ are induced by sets $A \in \mathcal{A}$?

Shatter Coefficients

Definition: The *shatter coefficient* of the family $\mathcal{A}$ on points $x_1, \ldots, x_n$ is the number of *distinct* subsets of $x_1, \ldots, x_n$ induced by sets in A,

$$S_{\mathcal{A}}(x_1^n) = |\{A \cap \{x_1, \ldots, x_n\} : A \in \mathcal{A}\}|$$

Note: Always have $1 \leq S_{\mathcal{A}}(x_1^n) \leq 2^n$, even if family $\mathcal{A}$ is infinite

Definition: The family $\mathcal{A}$ *shatters* $x_1, \ldots, x_n$ if $S_{\mathcal{A}}(x_1^n) = 2^n$, or equivalently, the elements of $\mathcal{A}$ induce every subset of $x_1, \ldots, x_n$

Sub-Multiplicative Property of Shatter Coefficients

Fact: For every $m, n \geq 1$ and $x_1, \ldots, x_{n+m} \in \mathcal{X}$

$$S_{\mathcal{A}}(x_1^{m+n}) \leq S_{\mathcal{A}}(x_1^m)\, S_{\mathcal{A}}(x_{m+1}^{m+n})$$

Fact: If the sequence $X_1, X_2, \ldots \in \mathcal{X}$ is stationary, then

$$\frac{\mathbb{E} \log S_{\mathcal{A}}(X_1^n)}{n} \to \inf_k \frac{\mathbb{E} \log S_{\mathcal{A}}(X_1^k)}{k}$$

In other words, the expected exponential growth rate of the shatter coefficients converges to a constant

Digression: The Vapnik-Chervonenkis (VC) Dimension

Recall: Family $\mathcal{A} \subseteq 2^{\mathcal{X}}$ shatters $x_1, \ldots, x_n \in \mathcal{X}$ if $S(x_1^n : \mathcal{A}) = 2^n$

Definition: The VC-dimension $\dim(\mathcal{A})$ is the largest n such that $\mathcal{A}$ shatters some n-point set. If $\mathcal{A}$ shatters arbitrarily large finite sets then $\dim(\mathcal{A}) = \infty$

Examples

- $\mathcal{X} = \mathbb{R}$, family $\mathcal{A} = \{(-\infty, t] : t \in \mathbb{R}\}$ has $\dim(\mathcal{A}) = 1$
- $\mathcal{X} = \mathbb{R}^2$, family $\mathcal{A} =$ all open disks has $\dim(\mathcal{A}) = 3$
- $\mathcal{X} = \mathbb{R}^d$, family $\mathcal{A} =$ all half-spaces has $\dim(\mathcal{A}) = d + 1$
- $\mathcal{X} = \mathbb{R}^d$, family $\mathcal{A} =$ all rectangles has $\dim(\mathcal{A}) = 2d$
- $\mathcal{X} = \mathbb{R}^2$, family $\mathcal{A} =$ all convex sets has $\dim(\mathcal{A}) = \infty$

The Vapnik-Chervonenkis (VC) Dimension

Sauer's Lemma: If $\mathcal{A}$ has VC-dimension d then for all $n \geq 1$ and all $x_1, \ldots, x_n \in \mathcal{X}$

$$S_{\mathcal{A}}(x_1^n) \;\leq\; \sum_{k=0}^{d} \binom{n}{k} \;\leq\; (n+1)^d$$

which is a polynomial in n with degree d