Dosnker and VC Classes

BIO 251,
Lab 9

April 14, 2014

Symmetrization

Lemma (Symmetrization).

For every convex, nondecreasing $\Phi : \mathbb{R} \mapsto \mathbb{R}$ and a class of measurable functions $\mathcal{F}$,

$$\mathrm{E}^* \Phi(||\mathbb{P}_n - P||_{\mathcal{F}}) \leq \mathrm{E}^* \Phi(2||\mathbb{P}_n^0||_{\mathcal{F}})$$

Measurable Class

Definition (Measurable class).

A class $\mathcal{F}$ of measurable functions $f : \mathcal{F} \mapsto \mathbb{R}$ on a probability space $(\mathcal{X}, \mathcal{A}, P)$ is called a P-measurable class if the map:

$$(X_1, X_2, \ldots, X_n) \mapsto \left\| \sum_{i=1}^{n} e_i f(X_i) \right\|_{\mathcal{F}}$$

is measurable for all $\{e_1, \ldots, e_n\} \in \mathbb{R}^n$ on the completion of the space $(\mathcal{X}^n, \mathcal{A}^n, P^n)$.

Example

The use of the above definition is that we can claim that $\mathrm{E}_X \mathrm{E}_\varepsilon = \mathrm{E}^*$

GC Class Sufficient Condition 2

Theorem

Let $\mathcal{F}$ be a P-measurable class of measurable functions with envelope F such that $P^*F < \infty$. Let $\mathcal{F}_M$ be the class of functions $f\mathbb{1}_{F \leq M}$ when f ranges over $\mathcal{F}$. If $\log N(\varepsilon, \mathcal{F}_M, L_1(\mathbb{P}_n)) = o_P^*(n)$ for every ε and $M > 0$, then $||\mathbb{P}_n - P||_{\mathcal{F}}^* \to 0$ both almost surely and in mean. In particular $\mathcal{F}$ is a GC class.

A Donsker Theorem

Theorem

Theorem. Let $\mathcal{F}$ be a class of measurable functions, with envelope F, that satisfies the following uniform entropy bound:

$$\int_0^\infty \sup_Q \sqrt{\log N(\varepsilon ||F||_{Q,2}, \mathcal{F}, L_2(Q))} d\varepsilon < \infty$$

The supremum is taken over all finitely discrete probability measures Q on $(\mathcal{X}, \mathcal{A})$, such that $||F||^2_{Q,2} = \int F^2 dQ > 0$. Let the classes $\mathcal{F}_\delta = \{f - g : f, g \in \mathcal{F}, ||f - g||_{P,2} < \delta\}$ and $\mathcal{F}^2_\infty = \{f^2 : f \in \mathcal{F}_\infty\}$ be P-measurable for any $\delta > 0$. If $P^* F^2 < \infty$, then F is P-Donsker.

Example

Example

The set $\mathcal{F}$ of all indicator functions $\mathbb{1}_{(-\infty,t]}$ of cells in $\mathbb{R}$.

Some Definitions

- Let $\mathcal{C}$ be a collection of subsets of a set $\mathcal{X}$.
- Say that $\mathcal{C}$, *picks out* a certain subset from $\{x_1, \ldots, x_n\}$ if this subset takes the form $C \cap \{x_1, \ldots, x_n\}$ for some $C \in \mathcal{C}$.
- $\mathcal{C}$ is said to *shatter* $\{x_1, \ldots, x_n\}$ if all possible 2^n subsets can be picked out by $\mathcal{C}$.

VC-index

The *VC-index* $V(\mathcal{C})$ of the collection $\mathcal{C}$ is the smallest n for which there is no set of size n, which is shattered by $\mathcal{C}$.

VC-Index

$$\Delta_n(\mathcal{C}, x_1, \ldots, x_n) = \#\{C \cap \{x_1, \ldots, x_n\} : C \in \mathcal{C}\}$$

$$V(\mathcal{C}) = \inf\{n : \max_{x_1, \ldots, x_n} \Delta_n(\mathcal{C}, x_1, \ldots, x_n) < 2^n\}$$

Example

Example

- The collections of all cells of the form $(-\infty, c]$
- The collection of sets $(a, b]$ for $a, b \in \mathbb{R}$

A Combinatorial Lemma

Lemma (Sauer-Shelah).

Let $\{x_1, \ldots, x_n\}$ be arbitrary points. Then the total number of subsets $\Delta_n(\mathcal{C}, x_1, \ldots, x_n)$ picked out by $\mathcal{C}$ is bounded above by the number of subsets of $\{x_1, \ldots, x_n\}$ shattered by $\mathcal{C}$.

Corollary.

For a VC-class of sets of index $V(\mathcal{C})$, one has:

$$\max_{x_1,\ldots,x_n} \Delta_n(\mathcal{C}, x_1, \ldots, x_n) \leq \sum_{j=0}^{V(\mathcal{C})-1} \binom{n}{j}$$