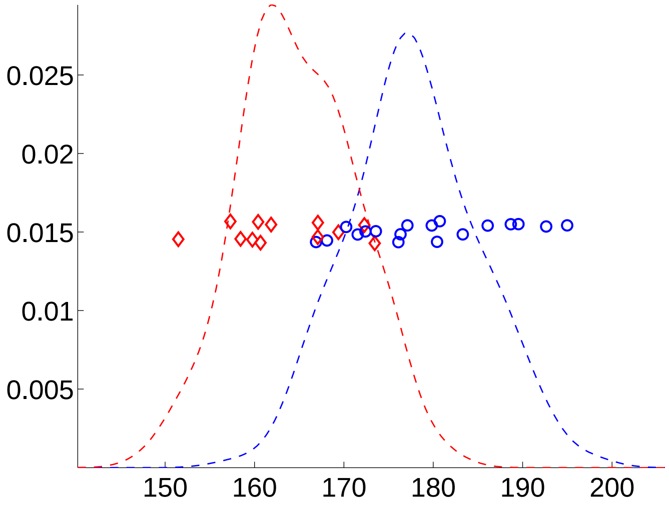
Foundation of Intelligent Systems, Part I

Statistical Learning Theory

mcuturi@i.kyoto-u.ac.jp

Previous Lecture : Hoeffding's Bound



- Hoeffding's Inequality: $P\left(|P_nf - Pf| > \varepsilon\right) \leq 2e^{-\frac{2n\varepsilon^2}{(b-a)^2}}$.
- With probability at least $1-\delta$,

$$|P_nf - Pf| \leq (b-a)\sqrt{\frac{\log\frac{2}{\delta}}{2n}}$$

Today: VC-dimension, SVM's

- Continue where we left:
  - Hoeffding's bound for finite families
  - Hoeffding's bound for countable families
  - Hoeffding's bound for arbitrary families of functions
    - Growth function
    - VC dimension
- VC-dimension for linear classifiers
- SVM

Obtaining Uniform Bounds

- Simple example with two functions f_1 and f_2.
- Define the two sets of n-uples,

$$C_{\mathbf{1}} = \{\{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\} \mid Pf_{\mathbf{1}} - P_n f_{\mathbf{1}} > \varepsilon\}$$

and

$$C_{\mathbf{2}} = \{\{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\} \mid Pf_{\mathbf{2}} - P_n f_{\mathbf{2}} > \varepsilon\}$$

- These sets are the "bad" sets for which empirical risk is much lower than the real risk.

Obtaining Uniform Bounds

- For each, we have the Hoeffing's inequalities **(no absolute value)**, that

$$P(C_1) \leq \delta, P(C_2) \leq \delta \text{ where } \delta = e^{-2n\varepsilon^2}.$$

- Note that whenever a n-uple is in $C_1 \cup C_2$, then either

$$Pf_{\mathbf{1}} - P_n f_{\mathbf{1}} > \varepsilon \text{ or } Pf_{\mathbf{1}} - P_n f_{\mathbf{1}} > \varepsilon.$$

- Of course, $P(C_1 \cup C_2) \leq P(C_1) + P(C_2) \leq 2\delta$.
- Thus, with probability smaller than 2δ at least one of f_1 or f_2 will be such that $Pf_{\mathbf{1}} - P_n f_{\mathbf{1}} > \varepsilon$.

Generalizing to N functions

- Consider $f_1, \cdots, f_N$ functions.
- Define the corresponding sets of n-uples, $C_1, \cdots, C_N$ with ε fixed.
- Of course,

$$P(C_1 \cup C_2 \cup \cdots \cup C_N) \leq \sum_{i=1}^{N} P(C_i)$$

- Use now Hoeffding's inequality

$$P(\exists f \in \{f_1, \cdots, f_N\} \,|\, Pf - P_n f > \varepsilon) = P\left(\bigcup_{i=1}^{N} C_i\right)$$

$$\leq \sum_{i=1}^{N} P(C_i) \leq N\delta = Ne^{-2n\varepsilon^2}$$

Error bound for finite families of functions

- We thus have that for **any** family of N functions,

$$P(\sup_{f\in\mathcal{F}} Pf - P_n f \geq \varepsilon) \leq N e^{-2n\varepsilon^2},$$

- or equivalently, that if $\mathcal{G} = \{g_1, \cdots, g_N\}$, with probability at least $1-\delta$,

$$\forall g \in \mathcal{G}, \quad R(g) \leq R_n(g) + \sqrt{\frac{\log N + \log\frac{1}{\delta}}{2n}}$$

Estimation bound for finite families of functions

- Recall that $g^\star$ is a function in $\mathcal{G}$ such that $R(g^\star) = \min_{g\in\mathcal{G}} R(g)$.
- The inequality

$$R(g^\star) \leq R_n^{\text{emp}}(g^\star) + \sup_{g\in\mathcal{G}} \left(R(g) - R_n^{\text{emp}}(g)\right),$$

- combined with $R_n^{\text{emp}}(g^\star) - R_n^{\text{emp}}(g_n) \geq 0$ by definition of g_n, we get

$$\begin{aligned} R(g_n) = R(g_n) - R(g^\star) + R(g^\star) &\leq \underbrace{R_n^{\text{emp}}(g^\star) - R_n^{\text{emp}}(g_n)}_{\geq 0} + R(g_n) - R(g^\star) + R(g^\star) \\ &\leq 2\sup_{g\in\mathcal{G}} |R(g) - R_n^{\text{emp}}(g)| + R(g^\star) \end{aligned}$$

- Hence, with probability at least $1-\delta$,

$$R(g_n) \leq R(g^\star) + 2\sqrt{\frac{\log N + \log\frac{2}{\delta}}{2n}}$$

Hoeffding's bound for countable families of functions

- Suppose now that we have a countable family $\mathcal{F}$
- Suppose that we assign a number $\delta(f) > 0$ to each $f \in \mathcal{F}$, which we use to set

$$P\left(|Pf - P_n f| > \sqrt{\frac{\log \frac{2}{\delta(f)}}{2n}} \right) \leq \delta(f),$$

- Using the union bound on a **countable set** (basic probability axiom),

$$P\left(\exists f \in \mathcal{F} : |P_n f - Pf| > \sqrt{\frac{\log \frac{2}{\delta(f)}}{2n}} \right) \leq \sum_{f \in \mathcal{F}} \delta(f).$$

- Let us set $\delta(f) = \rho p(f)$ with $\rho > 0$ and $\sum_{f \in \mathcal{F}} p(f) = 1$.
- Then with probability $1 - \rho$,

$$\forall f \in \mathcal{F}, Pf \leq P_n f + \sqrt{\frac{\log \frac{1}{p(f)} + \log \frac{1}{\rho}}{2n}}.$$

Hoeffding's bound for general families of functions

- Two problems:
  - Most interesting families of functions are not countable.
  - Defining the weights $p(f)$ is not so obvious.
- However, what really matters for a sample $\mathbf{z}_1, \cdots, \mathbf{z}_n$ is

$$\mathcal{F}_{\mathbf{z}_1,\cdots,\mathbf{z}_n} = \{(f(\mathbf{z}_1), f(\mathbf{z}_2), \cdots, f(\mathbf{z}_n)),\ f \in \mathcal{F}\}$$

- $\mathcal{F}_{\mathbf{z}_1,\cdots,\mathbf{z}_n}$ is a large set of binary vectors $\subset \{0,1\}^N$
- The more complex $\mathcal{F}$, the larger $\mathcal{F}_{\mathbf{z}_1,\cdots,\mathbf{z}_n}$ with maximum 2^n possible elements.

Definition 1 (Growth Function). *The growth function of $\mathcal{F}$ is equal to*

$$S_{\mathcal{F}}(n) = \sup_{(\mathbf{z}_1,\cdots,\mathbf{z}_n)} |\mathcal{F}_{\mathbf{z}_1,\cdots,\mathbf{z}_N}|$$

Vapnik-Chervonenkis

Theorem 1 (Vapnik-Chervonenkis). *For any $\delta > 0$, with probability at least $1-\delta$,*

$$\forall g \in \mathcal{G}, R(g) \leq R_n(g) + 2\sqrt{2\frac{\log S_{\mathcal{G}}(2n) + \log\frac{2}{\delta}}{n}}$$

- To prove it, we will need two lemmas,

Lemma 1 (Symmetrization). *For any $t > 0$ such that $nt^2 \geq 2$, and any n' more independent samples of P,*

$$P(\sup_{f\in\mathcal{F}} Pf - P_nf \geq t) \leq 2P(\sup_{f\in\mathcal{F}} P'_nf - P_nf \geq t/2)$$

Lemma 2 (Chebyshev's Inequality). *For any $t > 0$,*

$$P(|X - \mathbb{E}[X]| \geq t| \leq \frac{\textbf{var}\, X}{t^2}$$

Vapnik-Chervonenkis Entropy

- The VC bound holds for any probability distribution.
- As a result, it might be too loose. A density dependent result is given, using

  Definition 2. *The VC entropy is defined as*

  $$H_{\mathcal{F}}(n) = \log \mathbb{E}[|\mathcal{F}_{\mathbf{z}_1,\cdots,\mathbf{z}_N}|]$$

- The bound is then

  Theorem 2. *For any* $\delta > 0$*, with probability at least* $1-\delta$*,*

  $$\forall g \in \mathcal{G}, R(g) \leq R_n(g) + 2\sqrt{2\frac{\boldsymbol{H_{\mathcal{G}}(2n)} + \log\frac{2}{\delta}}{n}}$$

Vapnik-Chervonenkis Dimension

Definition 3 (VC Dimension)**.** *The VC dimension of a class $\mathcal{G}$ is the largest n such that*

$$S_{\mathcal{G}}(n) = 2^n.$$

- Since n points can have 2^n configurations, the VC dimension is the largest number of points which can be *shattered* ($i.e.$split arbitrarily) by the function class.
- The VC dimension of linear classifiers in $\mathbb{R}^d$ is $d + 1$.

Vapnik-Chervonenkis

- Given the VC dimension h of a family $\mathcal{G}$, we can prove

$$\forall g \in \mathcal{G}, R(g) \leq R_n(g) + 2\sqrt{2\frac{h \log \frac{2en}{h} + \log \frac{2}{\delta}}{n}}$$

Lemma 3 (Vapnik and Chervonenkis, Sauer, Shelah). *Let $\mathcal{G}$ be a class of functions with finite VC-dimension h. Then,*

$$\forall n \in \mathbb{N}, S_{\mathcal{G}}(n) \leq \sum_{i=0}^{h} \binom{n}{i},$$

$$\forall n \geq h, S_{\mathcal{G}}(n) \leq \left(\frac{en}{h}\right)^h.$$

- Combining with VC theorem, we obtain the result given above.
- Important thing: difference between true and empirical risks is at most of the order of

$$\sqrt{\frac{h \log n}{n}}$$