Theory of classification:

Basic Model :

- $x \in \mathcal{X}$, $\mathcal{X}$: a measurable space equipped with a σ-algebra.
- $y \in \{-1, 1\}$ a class.
- x is an observation.
- $g : \mathcal{X} \to \{-1, 1\}$: a classifier.
- (X, Y) : a random pair.
- $\eta(x) = P\{Y = 1 \mid X = x\}$: a posteriori probability.
- measure perf of classifier g by its probability of error:

$$L(g) = P\{g(X) \neq Y\}.$$

- Given η, classifier with minimal prob. of error is:

$$g^*(x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2 \\ -1 & \text{otherwise} \end{cases}.$$

then $L(g^*) \leq L(g)$ $\forall g$.

- The minimal risk $L^* \triangleq L(g^*)$ is called Bayes risk.
- We have: $L(g) - L^* = E\left[1_{\{g(X) \neq g^*(X)\}} \left| 2\eta(X) - 1 \right| \right] \geq 0$.
- g^* can be called the Bayes classifier.
- data is denoted $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ i.i.d.
- A classifier constructed using D_n is g_n.
- $g_n(X) = g_n(X; X_1, Y_1, \ldots, X_n, Y_n)$.
- Performance of g_n : $L(g_n) = P[g_n(X) \neq Y \mid D_n]$.

3. Empirical risk minimization and Rademacher averages.

1. Consider a class C of classifiers $g : \mathcal{X} \to \{-1, 1\}$ and use data-based estimates of the probabilities of error $L(g)$ to select a classifier from the class.

2. The error count is a natural estimate of $L(g)$.

$$L_n(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{g(X_i) \neq Y_i\}$$

$L_n(g)$ is called the empirical error of the classifier g.

3. Denote by g_n^* the classifier that minimizes the estimated probability of error over the class:

$$L_n(g_n^*) \leq L_n(g) \quad \forall g \in C$$

4. We have:

- $$L(g_n^*) - \inf_{g \in C} L(g) \leq 2 \sup_{g \in C} | L_n(g) - L(g) |$$
- $$L(g_n^*) \leq L_n(g_n^*) + \sup_{g \in C} | L_n(g) - L(g) |$$

5. $nL_n(g)$ is a r.v. binomially distributed with parameters n & $L(g)$.

6. Thus, to obtain bounds for the success of ~~e~~ empirical error minimization we need to study uniform deviations of binomial r.v. from their means.

7. Let: $X_1, \ldots, X_n$ be i.i.d r.v. $X_i \in \mathcal{X}$.
   - $\mathcal{F}$ be a class of bounded fnc $\mathcal{X} \to [-1, 1]$.
   - $Pf = E[f(X_1)]$ denotes expectation.
   - $P_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i)$ denotes empirical average.

we are interested in upper bound for the maximal deviation:

$$\sup_{f \in \mathcal{F}} (Pf - P_n f)$$

- Theorem 3.1 (bounded differences inequality). Let $g: \mathcal{X}^n \to R$ be a fun. of n vars | $\exists\ c_i$

$$\sup_{\substack{x_1,\dots,x_n\\ x_i' \in \mathcal{X}}} \left| g(x_1,\dots,x_n) - g(x_1,\dots,x_{i-1},x_i',x_{i+1},\dots,x_n) \right| \le c_i, \quad 1 \le i \le n.$$

Let: $X_1,\dots,X_n$ be n ~~iid~~ independent r.v.

- r.v. $Z = g(X_1,\dots,X_n)$ satisfies

$$P\left[\, |Z - E[Z]| > t \,\right] \le 2 e^{\frac{-2t^2}{C}}$$

where $C = \sum_{i=1}^{n} c_i^2$

- An example of a fun. that satisfies the bounded differences assumption is:

$$Z = \sup_{f \in \mathcal{F}} |Pf - P_n f|$$

- Z satisfies the bounded diff ass. with $c_i = \frac{2}{n}$.
- With prob at least $1-\delta$:

$$\sup_{f \in \mathcal{F}} |Pf - P_n f| \le E\left[\sup_{f\in\mathcal{F}} |Pf - P_n f|\right] + \sqrt{\frac{2\log\frac{1}{\delta}}{n}}.$$

- Introduce a "ghost sample" $X_1',\dots,X_n'$, independent of the X_i and distributed identically. If $P_n' f = \frac{1}{n}\sum_{i=1}^{n} f(X_i')$, then by Jensen's ~~ineq~~ inequality.

$$E\left[\sup_{f\in\mathcal{F}} |Pf - P_n f|\right] = E\left[\sup_{f\in\mathcal{F}} \left(E\left[\,|P_n' f - P_n f| \,\middle|\, X_1,\dots,X_n\right]\right)\right]$$

$$\le E\left[\sup_{f\in\mathcal{F}} |P_n' f - P_n f|\right].$$

- It can be shown that:

$$E\left[\sup_f |P_n' f - P_n f|\right] \le 2E\left[\sup_f \frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i f(X_i)\right|\right].$$

- Let: $A \in R^n$ be a bounded set of vectors $a = (a_1, \dots, a_n)$ ~~and~~
  - $R_n(A) = E\left[\sup_{a\in A} \frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i a_i\right|\right]$

- $R_n(A)$ is called the Rademacher average associated with A.
- for a given seq. $x_1, \dots, x_n \in \mathcal{X}$, we write $\mathcal{F}(x_1^n)$ for the class of n-vectors $(f(x_1), \dots, f(x_n))$ with $f \in \mathcal{F}$.

  so: $\mathcal{F}(x_1^n) = \{ (f(x_1), \dots, f(x_n)) : f \in \mathcal{F} \}$.
- With prob. at least $1-\delta$:

$$\sup_f |Pf - P_n f| \leq 2E\left[R_n(\mathcal{F}(X_1^n)) \right] + \sqrt{\frac{2\log\frac{1}{\delta}}{n}}.$$

We also have: $\sup_f |Pf - P_n f| \leq 2 R_n(\mathcal{F}(X_1^n)) + \sqrt{\frac{2\log\frac{2}{\delta}}{n}}$.

this is a data-dependent bound.

- <u>Properties of Rademacher averages</u>. let A, B be bounded subsets of $\mathbb{R}^n$ & let $c \in \mathbb{R}$ be a constant. Then:
  - $R_n(A \cup B) \leq R_n(A) + R_n(B)$.
  - $R_n(c \cdot A) = |c| R_n(A)$. $\quad c \cdot A = \{ca : a \in A\}$
  - $R_n(A \oplus B) \leq R_n(A) + R_n(B)$. $\quad A \oplus B = \{a + b : a \in A, b \in B\}$
- Moreover, if $A = \{a^{(1)}, \dots, a^{(N)}\} \subset \mathbb{R}^n$ is a finite set, then

$$R_n(A) \leq \max_{j=1,\dots,N} \|a^{(j)}\| \frac{\sqrt{2\log N}}{n}.$$

$\|\cdot\|$ is Euclidean norm.

- If $\text{absconv}(A) = \left\{ \sum_{j=1}^{N} c_j a^{(j)} : N \in \mathbb{N}, \sum_{j=1}^{N} |c_j| \leq 1, a^{(j)} \in A \right\}$

  is the absolute convex hull of A, then:

$$R_n(A) = R_n(\text{absconv}(A)).$$

- The contraction principle states that if $\phi: \mathbb{R} \to \mathbb{R}$ is a fn. with $\phi(0) = 0$ and Lipschitz constant L_ϕ and $\phi \circ A$ is the set of vectors of form $(\phi(a_1), \dots, \phi(a_n)) \in \mathbb{R}^n$ with $a \in A$.

$$R_n(\phi \circ A) \leq L_\phi R_n(A)$$

- Consider the case when $\mathcal{F}$ is a class of indicator functions.

For any collections of points $x_1^n = (x_1, \dots, x_n)$, $\mathcal{F}(x_1^n)$ is a finite subset of $\mathbb{R}^n$ whose cardinality is denoted by $S_{\mathcal{F}}(x_1^n)$ and is called the VC shatter coefficient.

- Obviously, $S_{\mathcal{F}}(x_1^n) \leq 2^n$.
- We have, $\forall x_1^n$, $R_n(\mathcal{F}(x_1^n)) \leq \sqrt{\dfrac{2\log S_{\mathcal{F}}(x_1^n)}{n}}$ (we used the fact that $\sum_i f(X_i)^2 \leq n$)
- In particular,
$$E\left[\sup_f |Pf - P_n f|\right] \leq 2E\left[\sqrt{\frac{2\log S_{\mathcal{F}}(X_1^n)}{n}}\right]$$
- The log. of the VC shatter coefficient may be upper bounded in terms of a combinatorial quantity, called the VC dimension.
- If $A \subset \{-1,1\}^n$, then the VC dimension of A is the size V of the largest set of indices $\{i_1, \dots, i_V\} \subset \{1, \dots, n\}$ | for each binary V-vector $b = (b_1, \dots, b_V) \in \{-1,1\}^V$, $\exists$ an $a = (a_1, \dots, a_n) \in A$ | $(a_{i_1}, \dots, a_{i_V}) = b$.
- A key inequality establishing a rel. b/t shatter coefficient & VC dimension is known as Sauer's lemma which states that the cardinality of any set $A \subset \{-1,1\}^n$ may be upper bounded as
$$|A| \leq \sum_{i=0}^{V} \binom{n}{i} \leq (n+1)^V$$
where V is the VC dimension of A.
- In particular, $\log S_{\mathcal{F}}(x_1^n) \leq V(x_1^n)\log(n+1)$.
where we denote by $V(x_1^n)$ the VC dimension of $\mathcal{F}(x_1^n)$.
- Thus: $$E\left[\sup_f \left[Pf - P_n f\right]\right] \leq 2E\left[\sqrt{2V(X_1^n)\frac{\log(n+1)}{n}}\right]$$

- To obtain distribution-free upper bounds, introduce the VC dim. of a class of binary funs. $\mathcal{F}$, defined by:

$$V = \sup_{n, x_1^n} V(x_1^n).$$

- VC inequality. For all distributions, one has

$$E\left[\sup_f (Pf - P_n f)\right] \leq 2\sqrt{\frac{2V\log(n+1)}{n}}$$

also,

$$E\left[\sup_f (Pf - P_n f)\right] \leq C\sqrt{\frac{V}{n}}$$

for a universal constant C.

- One useful property: let G be an m-dimensional vector space of real-valued functions defined on $\mathcal{X}$. The class of indicator functions:

$$\mathcal{F} = \{f(x) = 1_{g(x) \geq 0} : g \in G\}$$

has VC dimension $V \leq m$.