Lecture 2

—

Empirical Risk Minimization and Complexity

Agenda

- Binary classification - Goal and Probabilistic setup
- The Principle of Empirical Risk Minimization (ERM)
- A first go - the finite case
- Concentration Bounds - McDarmid's Inequality
- The Vapnik-Chervonenkis inequality
- Complexity (Combinatorial) - VC Dimension

Probabilistic setup for binary classification

- Random pair $= (X, Y) \sim P$ unknown
- $X =$ observation vector in $\mathcal{X}$ (ex: $\mathbb{R}^d$ with $d >> 1$)
- $Y =$ binary label in $\mathcal{Y} = \{-1, +1\}$
- **Our goal:** guess the *output* Y from the *input* observation X
- Examples:
  - Diagnosis medical support: predicting the occurence of a disease
    X measurements (blood pressure, medical history, age, *etc.*)
    $Y = 1$ "heart attack" vs $Y = 0$ "no heart attack"
  - Credit risk screening: predicting bankruptcy
    X socio-economic profile (turnover, headcount, *etc.*)
    $Y = 1$ "default" vs $Y = 0$ "debt repayment"
  - Spam filtering: X features of the e-mail (*e.g.* tf-idf vector)
    $Y = 1$ "spam" vs $Y = 0$ "not spam"

Probabilistic setup for binary classification

- **Classifier**: $C : x \in \mathcal{X} \mapsto C(x) \in \{-1, 1\}$ in a **class** $\mathcal{G}$
- Risk functional (unknown!) $=$ **Expected prediction error**

$$L(C) = \mathbb{E}[\mathbb{I}\{Y \neq C(X)\}]$$

  to minimize over $C \in \mathcal{G}$
- $\mathcal{G}$ is in 1-to-1 correspondence with the class of sets $\{\{x \in \mathcal{X} : \ C(x) = +1\} : \ C \in \mathcal{G}\}$

Theoretical Risk Minimization

- Let $\eta(x) = \mathbb{P}(Y = +1|X = x)$ **regression function**
- Let $p = \mathbb{P}(Y = +1)$
- Compute $C^* = \arg\min_{C \in \mathcal{G}} L(C)$
- Calculations yields the **Naive Bayes Classifier**

$$C^*(x) = 2 \cdot \mathbb{I}\{\eta(x) > 1/2\} - 1, \quad x \in \mathcal{X}$$

$\Rightarrow$ affects the likeliest label given the observation $X = x$

- Minimum theoretical risk: $L^* = L(C^*) = 1/2 - \mathbb{E}[|\eta(X) - 1/2|]$
- How close $\eta(X)$ is to $1/2$ governs the difficulty of the problem!

Theoretical Risk Minimization

- Theoretical **excess of risk**:

$$L(C) - L^* = \mathbb{E}[|\eta(X) - 1/2|\mathbb{I}\{X \in G^* \Delta G_C\}]$$

where G^*, G_C denote the subsets of the input space $\mathcal{X}$

$$\begin{aligned} G^* &= \{\eta(X) > 1/2\} \\ G_C &= \{C(X) = +1\} \end{aligned}$$

and $A \Delta B = (A \cap \bar{B}) \cup (\bar{A} \cap B)$ the *symmetric difference*.

- Insights: when a little of X's mass is concentrated around the **margin** $\{\eta(x) = 1/2\}$, the problem gets simpler.

Empirical Risk Minimization (ERM)

- Data $= D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$
- Classifier candidate: $C : \mathcal{X} \to \{-1, 1\}$ in a class $\mathcal{G}$
- Empirical risk functional $=$ Training (misclassification) error

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

  to minimize over $C \in \mathcal{G}$.
- Solution **"empirical risk minimizer"**: $\hat{C}_n = \arg\min_{C \in \mathcal{G}} L_n(C)$
- OK for the training data, now for **future data** (X, Y)?

Empirical Risk Minimization (ERM) - Heuristics

- $L_n(C)$ must be close to $L(C)$, uniformly over $C \in \mathcal{G}$
- For any fixed C, this is true (SLLN, CLT, Berry-Esseen, *etc.*)
- This should remain true, provided that $\mathcal{G}$ is **not too complex**, whatever (X, Y)'s distribution

Investigating the properties of the ER Minimizer

- Don't forget that $\hat{C}_n$ is **random** (depending on the data D_n)
- Let $(X, Y) \sim P$ be a **new random pair**, independent from D_n
  Will $\hat{C}_n$ performs well as a classifier for this novel pair?

$$\Rightarrow \text{ compute } L(\hat{C}_n) = \mathbb{P}(Y \neq \hat{C}_n(X) \mid D_n)$$

- $L(\hat{C}_n)$ is a **random variable**! It depends on the data D_n.
- **Deviation** between the r.v. $L(\hat{C}_n)$ and the min. error L^* (cst)

$$\Rightarrow \text{ Study the excess of risk } 0 \leq \mathcal{E}(C) = L(\hat{C}_n) - L^*$$

- Learning Theory: compute explicit **confidence bounds**, $\forall \epsilon > 0$

$$\mathbb{P}_{D_n}(L(\hat{C}_n) - L^* \geq \epsilon) \leq ?$$

Learning Bounds

- Consider $C_0 = \arg\min_{C \in \mathcal{G}} L(C)$ (theoret. minimizer over $\mathcal{G}$)
- Check the **"bias-variance"** decomposition

$$L(\hat{C}_n) - L^* \quad \leq \quad 2 \sup_{C \in \mathcal{G}} |L(C) - \hat{L}_n(C)| + L(C_0) - L^*$$

as $\mathcal{G}$ "increases" $\nearrow$ $\searrow$

- The second term depends on the model $\mathcal{G}$ solely (bias)
- The 1st term (estimation) involves **concentration** of

$$Z = \{L(C) - \hat{L}_n(C)\}_{C \in \mathcal{G}}$$

$\Rightarrow$ theory of **empirical processes**

Empirical processes - Basics

- Let $X_1, \ldots, X_n$ be i.i.d. r.v.'s drawn as P
- Let $P_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$ the empirical df
- Let $\mathcal{F}$ be a class of functions $f : \mathbb{R} \to \mathbb{R}$
- **Empirical process** $\{P_n f\}_{f\in\mathcal{F}}$: $P_n f = n^{-1}\sum_{i=1}^n f(X_i)$, $f \in \mathcal{F}$
- Investigate which conditions on $\mathcal{F}$ allow to **control**

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

- Ex.: recall **Donsker's theorem**, $\mathcal{F} = \{\mathbb{I}\{. \leq x\},\ x \in \mathbb{R}\}$

$$\sqrt{n}\sup_{x\in\mathbb{R}} |n^{-1}\sum_{i\leq n} \mathbb{I}\{X_i \leq x\} - P(]-\infty, x])| \Rightarrow \sup_{t\in[0,1]} |B(t)|$$

Basics inequalities

- Finite class: $Card(\mathcal{F}) = N$.

  "Union's bound" combined with **Chernoff's method**

  $$\mathbb{P}(\sup_{f\in\mathcal{F}} |P_n f - Pf| \geq \epsilon) \leq 2N \cdot e^{-2n\epsilon^2}$$

  if $\forall f \in \mathcal{F}:\ 0 \leq f \leq 1$

- Cumulative distribution functions: **Dvoretsky-Kiefer-Wolfowitz**

  $$\mathbb{P}(\sqrt{n}\sup_{x\in\mathbb{R}} |\frac{1}{n}\sum_{i\leq n}\mathbb{I}\{X_i \leq x\} - P(]-\infty, x])| \geq \epsilon) \leq 2e^{-2\epsilon^2}$$

- McDarmid (1989)

The finite situation

- **Hoeffding inequality :** $X_1, \ldots, X_n$ independent r.v.'s such that $-\infty < a_i \le X_i \le b_i < +\infty$ almost-surely. Let $S_n = \sum_{i \le n} X_i$. Then, for any $\epsilon > 0$,

$$\mathbb{P}\{S_n - \mathbb{E}[S_n]\} \le \exp\left\{-\frac{2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right\}$$

- **Application to the ERM paradigm:** if $\#\mathcal{G} = N$, then: $\forall \epsilon > 0$,

$$\mathbb{P}\{\sup_{C \in \mathcal{G}} |\hat{L}_n(C) - L(C)| \ge \epsilon\} \le 2Ne^{-2n\epsilon^2}$$

- Bound the expected maximal deviation

$$\mathbb{E}[\sup_{C \in \mathcal{G}} |\hat{L}_n(C) - L(C)|]$$

by integrating the bound: if $Z \ge 0$ a.s., $\mathbb{E}[Z] = \int_{t>0} \mathbb{P}\{Z \ge t\} dt$

The finite situation

- **Lemma:** Let $Z_1, \ldots, Z_n$ be r.v.'s such that: $\forall s > 0$, $\mathbb{E}[\exp sY_i] \leq \exp s^2\sigma^2/2$. Then

$$\mathbb{E}[\max_{1\leq i\leq n} Z_i] \leq \sigma\sqrt{2\log n}$$

- **Application:**

$$\mathbb{E}[\sup_{C\in\mathcal{G}} |\hat{L}_n(C) - L(C)|] \leq \sqrt{\frac{\log(2N)}{2n}}$$

McDarmid's inequality - Bounded differences

- $g : A^n \to \mathbb{R}$ such that: $\forall i \in \{1, \ldots, n\}$, $\forall x \in A^n$, $\forall x_i' \in A$,

  $|g(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - g(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c_i$

- If $X_1, \ldots, X_n$ are independent and g has bounded differences, then: $\forall t > 0$,

$$\mathbb{P}\{g(X_1, \ldots, X_n) - \mathbb{E}[g(X_1, \ldots, X_n)] > t\} \leq e^{-\frac{2t^2}{\sum_{i=1}^n c_i^2}}$$

- **Application:** Whatever the class $\mathcal{G}$, we have:

$$\mathbb{P}\{|\sup_{C \in \mathcal{G}} |\hat{L}_n(C) - L(C)| - \mathbb{E}[|\sup_{C \in \mathcal{G}} |\hat{L}_n(C) - L(C)|] > \epsilon\} \leq 2e^{-2n\epsilon^2}$$

Measuring Complexity - Combinatorial Approach

A. Chervonenkis & V. Vapnik

Measuring Complexity - Combinatorial Approach

- Vapnik - Chervonenkis: **VC dimension** of a class $\mathcal{A}$ of subsets $A \subset \mathbb{R}^d$
- Let $x_1^n = (x_1, \ldots, x_n)$ be n points in $\mathbb{R}^d$. Define
  - **Trace:**
    $$Tr(\mathcal{A}, x_1^n) = \{A \cap x_1^n;\ A \in \mathcal{A}\}$$
  - **Shattering coefficient:**
    $$S_{\mathcal{A}}(n) = \max_{x_1^n} Card\, Tr(\mathcal{A}, x_1^n)$$
  - Ex: half-lines of $\mathbb{R}$: $S_{\mathcal{A}}(n) = n + 1$
- Other approaches: entropy metric, Rademacher chaos, *etc.*

Vapnik-Chervonenkis inequality

- $\mathcal{A}$ class of borelian subsets $A \subset \mathbb{R}^d$, μ probability measure on $\mathbb{R}^d$
- $X_i \overset{i.i.d.}{\sim} \mu(dx)$, empirical measure $\hat{\mu}_n = (1/n)\sum_{i\leq n} \delta_{X_i}$
- **Result:**

$$\mathbb{E}\left[\sup_{A\in\mathcal{A}} |\hat{\mu}_n(A) - \mu(A)|\right] \leq 2\sqrt{\frac{\log(2S_{\mathcal{A}}(n))}{n}}$$

- **Proof:** Ghost sample $X_i' \overset{i.i.d.}{\sim} \mu(dx)$ independent from $X_1, \ldots, X_n$
  $\hat{\mu}_n' = (1/n)\sum_{i\leq n} \delta_{X_i'}$

$$\begin{aligned}
\mathbb{E}\left[\sup_{A\in\mathcal{A}} |\hat{\mu}_n(A) - \mu(A)|\right] &= \mathbb{E}\left[\sup_{A\in\mathcal{A}} |\mathbb{E}[\hat{\mu}_n(A) - \hat{\mu}_n'(A) \mid X_1', \ldots, X_n']|\right] \\
&\leq \mathbb{E}\left[\sup_{A\in\mathcal{A}} \mathbb{E}[|\hat{\mu}_n(A) - \hat{\mu}_n'(A)| \mid X_1', \ldots, X_n']\right] \\
&\leq \mathbb{E}\left[\sup_{A\in\mathcal{A}} |\hat{\mu}_n(A) - \hat{\mu}_n'(A)|\right] \quad \text{symmetrization}
\end{aligned}$$

Vapnik-Chervonenkis inequality

randomization: consider a Rademacher chaos $\sigma_1, \ldots, \sigma_n$, i.i.d. and independent from the $(X_i, X_i')'s$, $\mathbb{P}\{\sigma_i = +1\} = \mathbb{P}\{\sigma_i = -1\} = 1/2$

$$\mathbb{E}\left[\sup_{A \in \mathcal{A}} |\hat{\mu}_n(A) - \mu(A)|\right] \leq \frac{1}{n} \times$$

$$\mathbb{E}\left[\mathbb{E}[\sup_{A \in \mathcal{A}} |\sum_{i=1}^{n} \sigma_i \left(\mathbb{I}\{X_i \in A\} - \mathbb{I}\{X_i' \in A\}\right)| \mid X_1, \ldots, X_n,\ X_1', \ldots, X_n']\right]$$

Observe that, for fixed $(X_1, X_1'), \ldots, (X_n, X_n')$, $\sup_{\mathcal{A}} = \max_{\widehat{\mathcal{A}}}$ with $\#\widehat{\mathcal{A}} \leq S_{\mathcal{A}}(2n)$ and

$$\mathbb{E}[\max_{\widehat{\mathcal{A}}} |\sum_{i=1}^{n} \sigma_i \left(\mathbb{I}\{x_i \in A\} - \mathbb{I}\{x_i' \in A\}\right)|] \leq \sqrt{2n \log(2 S_{\mathcal{A}}(2n))}$$

Notice finally that $S_{\mathcal{A}}(2n) \leq S_{\mathcal{A}}(n)^2$

Vapnik-Chervonenkis dimension

- $dim_{VC}\mathcal{A} = \max\{n \geq 1 : \ S_{\mathcal{A}}(n) = 2^n\}$
- **Application to ERM:** if $dim_{VC}\mathcal{G} = V_{\mathcal{G}}$,

$$\begin{aligned} \mathbb{E}[L(\hat{C}_n) - \inf_{C \in \mathcal{G}} L(C)] &\leq 4\sqrt{\frac{\log(2S_{\mathcal{G}}(n))}{n}} \\ &\leq 4\sqrt{\frac{V_{\mathcal{G}} \log(n+1) + \log 2}{n}} \end{aligned}$$

Shatter coefficients - Basic properties

- $S_{\mathcal{A}}(n+m) \leq S_{\mathcal{A}}(n) \times S_{\mathcal{A}}(m)$
- If $\mathcal{C} = \mathcal{A} \cup \mathcal{B}$, then $S_{\mathcal{C}}(n) \leq S_{\mathcal{A}}(n) + S_{\mathcal{B}}(n)$
- If $\mathcal{B} = \{A^c : \ A \in \mathcal{A}\}$, then $S_{\mathcal{A}}(n) = S_{\mathcal{B}}(n)$
- If $\mathcal{C} = \{A \cap B : \ A \in \mathcal{A}, B \in \mathcal{B}\}$, then $S_{\mathcal{C}}(n) \leq S_{\mathcal{A}}(n) \times S_{\mathcal{B}}(n)$
- If $\mathcal{C} = \{A \times B : \ A \in \mathcal{A}, B \in \mathcal{B}\}$, then $S_{\mathcal{C}}(n) \leq S_{\mathcal{A}}(n) \times S_{\mathcal{B}}(n)$
- **Sauer lemma** If $dim_{VC}\mathcal{A} = V < +\infty$, then

$$S_{\mathcal{A}}(n) \leq \sum_{i=0}^{V} \binom{n}{i}$$

- If $dim_{VC}\mathcal{A} = V < +\infty$, then, $\forall n$, we have $S_{\mathcal{A}}(n) \leq (n+1)^V$
- If $dim_{VC}\mathcal{A} = V < +\infty$, then, $\forall n \geq V$, we have $S_{\mathcal{A}}(n) \leq (ne/V)^V$

Examples

- If $\mathcal{A}$ is the class of all rectangles in $\mathbb{R}^d$, then $V = 2d$
- Let $\mathcal{G}$ be an m-dimensional vector space of functions $g : \mathbb{R}^d \to \mathbb{R}$. The class
$$\mathcal{A} = \{\{x : \ g(x) = 0\} : \ g \in \mathcal{G}\}$$
has VC dimension $V \leq m$
- **Consequences:**
    - the class of all linear halfspaces $\{\{x \in \mathbb{R}^d : \ A^t x \geq b\} : \ A \in \mathbb{R}^d, \ b \in \mathbb{R}\}$ has VC dimension $\leq d+1$
    - the class of all closed balls $\{\{x \in \mathbb{R}^d : \ ||x-c|| \geq b\} : \ c \in \mathbb{R}^d, \ b \in \mathbb{R}\}$ has VC dimension $\leq d+2$
    - the class of all ellipsoids $\{\{x \in \mathbb{R}^d : \ x^t \Gamma^{-1} x \leq 1\} : \ \Gamma$ symmetric positive definite$\}$ has VC dimension $\leq d(d+1)/2 + 1$

Application:

- VC theory provides **statistical guarantees** (generalization ability) for application of the ERM principle based on
  - binary decision trees with perpendicular/diagonal splits
  - general partitioning techniques with hypercubes
  - linear separators
  - *etc.*
- VC theory is **useless** for
  - nonlinear SVM
  - boosting
  - random forest
- but VC theory will return ... for the purpose of **model selection** (structural risk minimization)