STAT 615: Statistical Learning

Classification: Part 2

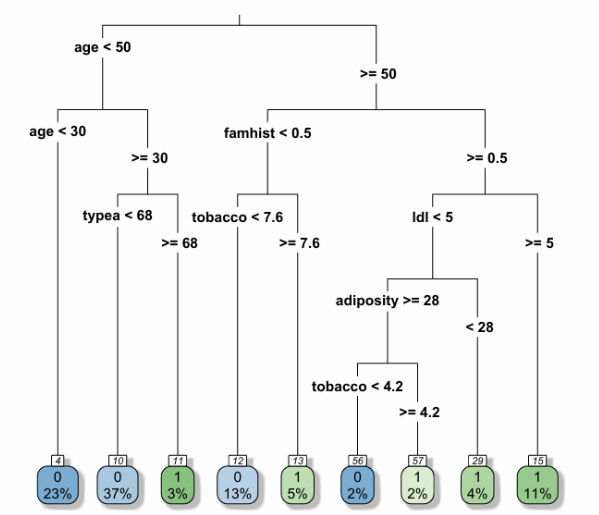
Yuling Yan

University of Wisconsin–Madison, Fall 2024

Tree-based methods

Classification tree

South African heart disease data: "0"="Yes, Disease", "1"="No"



Classification tree

Setup: $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{1, \ldots, K\}$, training data $(X_1, Y_1), \ldots, (X_n, Y_n)$

Idea: grow a tree to recursively partition the feature space into a set of rectangles, and do a simple majority vote in each rectangle

- Each node represents a rectangle in the feature space. The root node is the feature space $\mathcal{X} = \mathbb{R}^d$
- Each node is either a leaf (no children) or a parent (has two children)
- The left and right children comes from a partition of their parent node
- Suppose we have a collection of final partitioned regions associated with the leaves at the bottom of the tree, denoted by $R_1, \ldots, R_M$
- For any input x, suppose that $x \in R_j$, then this classification tree returns

$$\widehat{f}(x) = \arg\max_{k \in \mathcal{Y}} \sum_{X_i \in R_j} \mathbb{1}\{Y_i = k\}$$

i.e., the predicted label is the majority in the region R_j

How to grow a classification tree?

In order to grow a classification tree, we need to ask:

1. How to split each parent node?
2. How large should we grow the tree?

For the first question: **minimizing impurity**

- Suppose that the parent node is associated with a rectangle R
- Choose a covariate X_j and a split point t that minimizes the impurity
- Let the rectangles associated with its left and right children be

$$R_1(j,t) = \{X \in R : X_j \leq t\} \quad \text{and} \quad R_2(j,t) = \{X \in R : X_j > t\},$$

For the second question: **set some stopping criteria**.

- For example, we may fix some number n_0, and we might stop partition a node when its associated rectangle has fewer than n_0 training data points.

Impurity function

Let R be the node to be split into two regions. We choose

$$\arg\min_{j,t} \underbrace{\frac{|R_1(j,t)|}{|R|}\gamma(R_1(j,t)) + \frac{|R_2(j,t)|}{|R|}\gamma(R_2(j,t))}_{\text{impurity function}},$$

- Here $\gamma(R)$ measures the "variance" of the labels of data in R: we want

$$\{Y_i : X_i \in R\} \quad \text{to have low variability}$$

- For any given rectangle R, let

$$p_k = \frac{1}{|R|}\sum_{X_i \in R} \mathbb{1}\{Y_i = k\}, \quad 1 \le k \le K.$$

  Two common choice of the function $\gamma(\cdot)$:

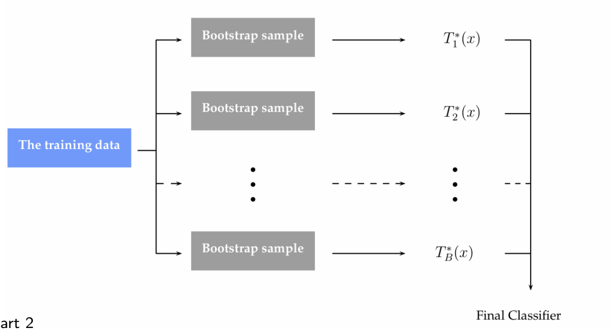
  - **Gini index**: $\gamma(R) = \sum_k p_k(1 - p_k)$
  - **Cross entropy**: $\gamma(R) = -\sum_k p_k \log p_k$

Insights

- **advantage:** the tree structure provides great interpretability
  - for example, it allows reasoning about the cause of diseases
- **disadvantage:** instability due to the use of *greedy* search:
  - splitting process is greedy
  - small changes in the training data can lead to significantly different tree structures
- **Solutions:**
  - Regularization: controlling tree growth parameters
  - Pruning: removing branches that do not provide significant predictive power
  - Ensemble Methods: use bagging to create a random forest

Bootstrap aggregating (Bagging)

- Training data $Z_n = \{(X_i, Y_i), 1 \leq i \leq n\}$
- Bootstrap sample $Z^{(*b)} = \{(X_i^{(*b)}, Y_i^{(*b)}), 1 \leq i \leq n\}$: sample n data points randomly from Z_n with replacement
- Apply the learning algorithm to the bootstrap sample for B times, and produce outcomes $\widehat{f}_b$
- Majority vote: $\widehat{f}^{\text{bagging}}(x) = \arg\max_{k \in \mathcal{Y}} \sum_{b=1}^{B} \mathbb{1}\{\widehat{f}_b(x) = k\}$



Insights

- Trees generated in bagging are identically distributed (not independent!)
- Bias of bagged tress is the same as the individual tree
- **Pro**: Reduce the variance, so good for high-variance, low-bias procedures, like trees.
- **Heuristics**: Suppose we have B identically distributed random variables with variance σ^2 and positive pairwise correlation ρ, then their average has variance of

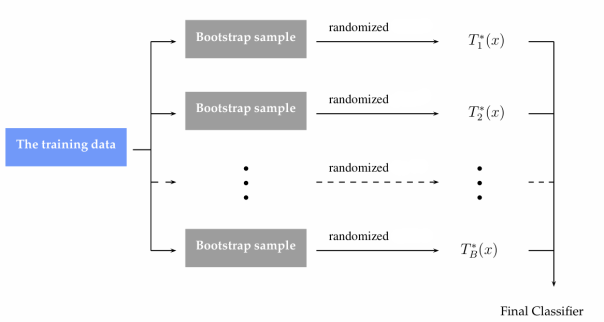
$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$$

- Increasing B does not reduce the first term

— Random Forest!

Random forests

- **Key idea**: use random dropout to decorrelate bootstrapped trees
- When growing a tree on a bootstrapped sample, before each split of the node, select $m \ll d$ variables at random as candidates to split
- Typical values for m is $\sqrt{d}$.
- Majority vote: $\widehat{f}^{\mathsf{RF}}(x) = \arg\max_{k \in \mathcal{Y}} \sum_{b=1}^{B} \mathbb{1}\{\widetilde{f}_b(x) = k\}$



How to remove bias: Boosting

- Setup: $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{\pm 1\}$
- Weak classifier: error rate only slightly better than random guess
- Key idea: sequentially apply weak classification algorithm to repeatedly modified versions of the data to produce a sequence of weak classifiers
    - assign unequal weights to training data points

      — *possible for trees*
    - sequentially find a committee of weak classifiers $\{\widehat{f}_m\}_{m=1}^M$
    - produce the final prediction through a weighted majority vote

$$\widehat{f}(x) := \mathrm{sign}(\sum_{m=1}^{M} \alpha_m \widehat{f}_m(x))$$

AdaBoost

Initialization: set the weights $w_i = 1/n$ for $1 \le i \le n$.
For $m = 1, \dots, M$:

- Fit a weak classifier $\widehat{f}_m(x)$ using training data with weights $\omega_1, \dots, \omega_n$
- Compute the weighted misclassification error:

$$\mathsf{err}^{(m)} = \frac{\sum_{i=1}^{n} w_i \, \mathbb{1}\{Y_i \neq \widehat{f}_m(X_i)\}}{\sum_{i=1}^{n} w_i}.$$

- Compute:

$$\alpha_m = \log\left(\frac{1 - \mathsf{err}^{(m)}}{\mathsf{err}^{(m)}}\right).$$

- Update the weights by:

$$w_i \leftarrow w_i \cdot \exp\left(\alpha_m \cdot \mathbb{1}\{Y_i \neq \widehat{f}_m(X_i)\}\right), \quad i = 1, 2, \dots, n.$$

Output: $\widehat{f}(x) = \mathsf{sign}\big(\sum_{m=1}^{M} \alpha_m \widehat{f}_m(x)\big)$.

AdaBoost: insights

Key idea: in the weight update step

$$w_i \leftarrow w_i \cdot \exp\left(\alpha_m \cdot \mathbb{1}\{Y_i \neq \widehat{f}_m(X_i)\}\right), \quad i = 1, 2, \ldots, n.$$

- For incorrectly classified data points, their weights get inflated by e^{α_m}
- Note that $\alpha_m > 0$ should always hold
- This re-weighting encourages the next classifier to focus more on the misclassified data points

Discussion: three main approaches to classification

Three main approaches

Bayes optimal classifier: for any $x \in \mathcal{X}$, output

$$f^\star(x) := \arg\max_{y \in \mathcal{Y}} \mathbb{P}(Y = y \mid X = x)$$

minimizes the Bayes risk $R(f) = \mathbb{P}(f(X) \neq Y)$

Three main approaches

Bayes optimal classifier: for any $x \in \mathcal{X}$, output

$$f^\star(x) \coloneqq \operatorname*{arg\,max}_{y \in \mathcal{Y}} \mathbb{P}(Y = y \mid X = x)$$

minimizes the Bayes risk $R(f) = \mathbb{P}(f(X) \neq Y)$

- Plug-in approach: model data distribution ρ, then estimate densities

$$\mathbb{P}(Y = y \mid X = x) = \frac{\mathbb{P}(X = x \mid Y = y)\,\mathbb{P}(Y = y)}{\sum_{y' \in \mathcal{Y}} \mathbb{P}(X = x \mid Y = y')\,\mathbb{P}(Y = y')}$$

Three main approaches

Bayes optimal classifier: for any $x \in \mathcal{X}$, output

$$f^\star(x) \coloneqq \operatorname*{arg\,max}_{y \in \mathcal{Y}} \mathbb{P}(Y = y \mid X = x)$$

minimizes the Bayes risk $R(f) = \mathbb{P}(f(X) \neq Y)$

- Plug-in approach: model data distribution ρ, then estimate densities

$$\mathbb{P}(Y = y \mid X = x) = \frac{\widehat{\mathbb{P}}(X = x \mid Y = y)\,\widehat{\mathbb{P}}(Y = y)}{\sum_{y' \in \mathcal{Y}} \widehat{\mathbb{P}}(X = x \mid Y = y')\,\widehat{\mathbb{P}}(Y = y')}$$

  Example: LDA, QDA, Kernel density classifier

Three main approaches

Bayes optimal classifier: for any $x \in \mathcal{X}$, output

$$f^\star(x) \coloneqq \operatorname*{arg\,max}_{y \in \mathcal{Y}} \mathbb{P}(Y = y \mid X = x)$$

minimizes the Bayes risk $R(f) = \mathbb{P}(f(X) \neq Y)$

- Plug-in approach
- Regression: modeling and estimating each

$$r_k(x) \coloneqq \mathbb{P}(Y = k \mid X = x) \quad \text{for} \quad k = 1, \ldots, K$$

Three main approaches

Bayes optimal classifier: for any $x \in \mathcal{X}$, output

$$f^\star(x) \coloneqq \operatorname*{arg\,max}_{y \in \mathcal{Y}} \mathbb{P}(Y = y \mid X = x)$$

minimizes the Bayes risk $R(f) = \mathbb{P}(f(X) \neq Y)$

- Plug-in approach
- Regression: modeling and estimating each

  $$r_k(x) \coloneqq \mathbb{P}(Y = k \mid X = x) \quad \text{for} \quad k = 1, \ldots, K$$

  Example: logistic regression

Three main approaches

Bayes optimal classifier: for any $x \in \mathcal{X}$, output

$$f^\star(x) \coloneqq \operatorname*{arg\,max}_{y \in \mathcal{Y}} \mathbb{P}(Y = y \mid X = x)$$

minimizes the Bayes risk $R(f) = \mathbb{P}(f(X) \neq Y)$

- Plug-in approach
- Regression
- Empirical risk minimization: choose a set of classifiers $\mathcal{F}$ and find $\widehat{f} \in \mathcal{F}$ that minimizes the “empirical risk”:

$$R_n(f) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{f(X_i) \neq Y_i\}$$

Intuition: when n is large, $R_n(f) \approx R(f)$ by LLN

Three main approaches

Bayes optimal classifier: for any $x \in \mathcal{X}$, output

$$f^\star(x) := \arg\max_{y \in \mathcal{Y}} \mathbb{P}(Y = y \mid X = x)$$

minimizes the Bayes risk $R(f) = \mathbb{P}(f(X) \neq Y)$

- Plug-in approach
- Regression
- Empirical risk minimization
- Other approaches: SVM, tree-based methods...

ERM: advantages

$$\widehat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{f(X_i) \neq Y_i\} =: R_n(f)$$

- a straightforward method based on simple heuristics

— *robustness!*

- can be easily generalized to other loss $\ell(\cdot, \cdot)$ by considering

$$R_n(f) := \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i)$$

  if the ultimate goal is to minimize $R_\ell(f) = \mathbb{E}[\ell(f(X), Y)]$. For example, in binary classification (i.e., $\mathcal{Y} = \{0, 1\}$)

  - Hinge loss $\ell(f(x), y) = \max\{0, 1 - yf(x)\}$
  - Logistic loss $\ell(f(x), y) = \log(1 + \exp(-yf(x)))$

— *Logistic regression can also be viewed as ERM!*

ERM: disadvantages

$$\widehat{f}_n = \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{f(X_i) \neq Y_i\} =: R_n(f)$$

- Not easy to compute (due to nonsmoothness of the indicator function)
- Solution: in binary classification (i.e., $\mathcal{Y} = \{0, 1\}$), consider using hinge loss or logistic loss $\ell(\cdot)$

  $$R_n(f) := \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i)$$

  and relax $f : \mathbb{R}^d \to \mathbb{R}$, and finally output sign$(2(f(x) - 1))$
- Here we will only focus on the standard ERM

ERM: error decomposition

$$\widehat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{f(X_i) \neq Y_i\} =: R_n(f)$$

- We want to control the excess risk

$$R(\widehat{f}_n) - R(f^\star) = \underbrace{R(\widehat{f}_n) - \min_{f \in \mathcal{F}} R(f)}_{\geq 0,\ \text{statistical error}} + \underbrace{\min_{f \in \mathcal{F}} R(f) - R(f^\star)}_{\geq 0,\ \text{approximation error}}$$

- approximation error: becomes smaller when choosing larger $\mathcal{F}$

  — *becomes 0 when $f^\star \in \mathcal{F}$*

- statistical error: becomes smaller when n becomes larger, and when choosing smaller $\mathcal{F}$ (why?)
- **trade-off between fit and complexity**
- In this course, we will focus on understanding statistical error with a given $\mathcal{F}$ that includes $f^\star$ (so that approximation error $= 0$)

Excess risk via uniform deviations

$$\widehat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{f(X_i) \neq Y_i\} =: R_n(f)$$

Theorem 3.1

The excess risk is upper bounded by

$$R(\widehat{f}_n) - R(f^\star) \leq 2 \sup_{f \in \mathcal{F}} |R_n(f) - R(f)|$$

Implications:

- For a given f, we know that $R_n(f) \to R(f)$ at a rate $O(1/\sqrt{n})$ by CLT

$$\sqrt{n}\big(R_n(f) - R(f)\big) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \mathsf{var}(\mathbb{1}\{f(X) \neq Y\}))$$

- But what about the uniform convergence of $\sup_{f \in \mathcal{F}} |R_n(f) - R(f)|$?

Concentration inequalities and uniform convergence

Why concentration inequalities?

Consider i.i.d. variables $X_1, \ldots, X_n$ with $\mathbb{E}[X_i] = \mu$ and var$(X_i) = \sigma^2$

- Central limit theorem (CLT):

$$\sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\Big) \xrightarrow{\mathsf{d}} \mathcal{N}(0, \sigma^2)$$

  tells us that the sample average concentrates around μ, and the deviation scales like $\sigma/\sqrt{n}$ as $n \to \infty$

- But this does not say anything useful when n is finite

- We want some non-asymptotic statement like:

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\Big| \geq \varepsilon(n, \delta)\Big) \leq \delta$$

  holds for any $\delta > 0$, where $\varepsilon(n, \delta) > 0$ is some quantity that depends on the sample size n and the exceptional probability δ

A simple case with i.i.d. Gaussian

Suppose that $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, then we have

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \sim \mathcal{N}\Big(0, \frac{\sigma^2}{n}\Big)$$

Theorem 3.2

For $G \sim \mathcal{N}(0,1)$ and any $t > 0$, we have

$$\Big(\frac{1}{t} - \frac{1}{t^3}\Big)\frac{1}{\sqrt{2\pi}}e^{-t^2/2} \leq \mathbb{P}(G \geq t) \leq \frac{1}{t}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}.$$

As a result,

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\Big| \geq t\Big) \leq \frac{2\sigma}{\sqrt{n}t}\frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{nt^2}{2\sigma^2}\Big)$$

A simple case with i.i.d. Gaussian

Suppose that $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, then we have

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \sim \mathcal{N}\Big(0, \frac{\sigma^2}{n}\Big)$$

Theorem 3.2

For $G \sim \mathcal{N}(0,1)$ and any $t > 0$, we have

$$\Big(\frac{1}{t} - \frac{1}{t^3}\Big)\frac{1}{\sqrt{2\pi}}e^{-t^2/2} \le \mathbb{P}(G \ge t) \le \frac{1}{t}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}.$$

As a result,

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\Big| \ge t\Big) \le \frac{2\sigma}{\sqrt{n}t}\frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{nt^2}{2\sigma^2}\Big)$$

Question: how can we extend these to more general distributions?

From Gaussian to sub-Gaussian

- Question: can we generalize these results to other random variables?
- Idea: consider other random variables with similar tail probability
- From Theorem 3.2, we know that for $G \sim \mathcal{N}(0, \sigma^2)$,

$$\mathbb{P}(|G| \geq t) \lesssim e^{-t^2/\sigma^2} \quad \text{for all } t \geq 0$$

- We may consider random variables satisfy this type of tail properties

— *sub-Gaussian*

Sub-Gaussian properties

Let X be a random variable, then the following properties are equivalent:

1. The tails of X satisfy
$$\mathbb{P}(|X| \geq t) \leq 2\exp\left(-t^2/K_1^2\right) \quad \text{for all } t \geq 0$$
2. The moments of X satisfy
$$\|X\|_{L^p} \coloneqq (\mathbb{E}[|X|^p])^{1/p} \leq K_2\sqrt{p} \quad \text{for all } p \geq 1$$
3. The moment generating function (MGF) of X^2 satisfies
$$\mathbb{E}\big[\exp(\lambda^2 X^2)\big] \leq \exp(K_3^2\lambda^2) \quad \text{for all}\, \lambda \text{ such that } |\lambda| \leq 1/K_3$$
4. The MGF of X^2 is bounded at some point, namely
$$\mathbb{E}\big[\exp\left(X^2/K_4^2\right)\big] \leq 2.$$
5. If $\mathbb{E}X = 0$, then the MGF of X satisfies
$$\mathbb{E}\big[\exp(\lambda X)\big] \leq \exp(K_5^2\lambda^2) \quad \text{for all } \lambda \in \mathbb{R}.$$

where $K_1, \ldots, K_5 > 0$ may differ by at most a multiplicative constant factor

Sub-Gaussian distributions: definition

- If X satisfies one of properties 1-4, it is a *sub-Gaussian random variable*.
- The *sub-Gaussian norm* of X, denoted $\|X\|_{\psi_2}$, is defined to be the smallest K_4 in property 4. In other words, we define

$$\|X\|_{\psi_2} = \inf\left\{t > 0 : \mathbb{E}\exp\left(X^2/t^2\right) \leq 2\right\}.$$

— *can also be defined using* K_1, K_2 *or* K_3

- Properties: there exists some absolute constants $c, C > 0$ such that
  - $P(|X| \geq t) \leq 2\exp\left(-ct^2/\|X\|_{\psi_2}^2\right)$
  - $\|X\|_{L^p} \leq C\|X\|_{\psi_2}\sqrt{p}$
  - $\mathbb{E}\exp\left(X^2/\|X\|_{\psi_2}^2\right) \leq 2$
  - if $\mathbb{E}[X] = 0$, then $\mathbb{E}\exp(\lambda X) \leq \exp(C\lambda^2\|X\|_{\psi_2}^2)$

Sub-Gaussian distributions: examples

- **Gaussian:** if $X \sim \mathcal{N}(0, \sigma^2)$, then X is sub-Gaussian with

  $$\|X\|_{\psi_2} \leq C\sigma$$

  for some universal constant $C = 2\sqrt{2/3}$.

- **Bounded:** any bounded random variable X is sub-Gaussian with

  $$\|X\|_{\psi_2} \leq C\|X\|_\infty$$

  for some universal constant $C = 1/\sqrt{\log 2}$.

Sub-Gaussian norm can be viewed as a characterization of “magnitude” for light tail distributions.

Centering and independent sums

Theorem 3.3

- *If X is sub-Gaussian, then $X - \mathbb{E}[X]$ is sub-Gaussian with*
$$\|X - \mathbb{E}[X]\|_{\psi_2} \leq C\|X\|_{\psi_2}$$
where C is an absolute constant.
- *Let $X_1, \ldots, X_N$ be independent, mean zero, sub-Gaussian random variables. Then the sum $S_N = \sum_{i=1}^N X_i$ is also sub-Gaussian, and its sub-Gaussian norm satisfies*
$$\|S_N\|_{\psi_2}^2 \leq C \sum_{i=1}^N \|X_i\|_{\psi_2}^2,$$
where C is an absolute constant.

Analog:

- If $X_1, \ldots, X_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$, then $S_N \sim \mathcal{N}(0, N\sigma^2)$
- If $X_1, \ldots, X_n$ are independent with $\|X_i\|_{\psi_2} \leq \sigma$, then $\|S_N\|_{\psi_2} \lesssim \sqrt{N}\sigma$

Hoeffding's inequality

Theorem 3.4 (Hoeffding's Inequality)

Let $X_1, \dots, X_N$ be independent, mean-zero, sub-Gaussian random variables. Then, for any $t \geq 0$, we have:

$$\mathbb{P}\left(\left|\sum_{i=1}^{N} X_i\right| \geq t\right) \leq 2\exp\left(-\frac{ct^2}{\sum_{i=1}^{N} \|X_i\|_{\psi_2}^2}\right),$$

where c is an absolute constant.

Implications

- **General Hoeffding:** under the setup of Theorem 3.4, consider any vector $\boldsymbol{a} = (a_1, \dots, a_n) \in \mathbb{R}^n$, we have

$$\mathbb{P}\left(\left|\sum_{i=1}^{N} a_i X_i\right| \geq t\right) \leq 2\exp\left(-\frac{ct^2}{K^2\|\boldsymbol{a}\|_2^2}\right),$$

  where $K \coloneqq \max \|X_i\|_{\psi_2}$.

- **Example:** suppose that $X_i \sim$ Bernoulli(p_i) for $1 \leq i \leq n$, then

$$\mathbb{P}\left(\left|\sum_{i=1}^{N} (X_i - p_i)\right| \geq t\right) \leq 2\exp\left(-\frac{ct^2}{N}\right),$$

  A sharper result for binomial concentration: Chernoff's inequality (HW)

Back to ERM: finite $\mathcal{F}$

$$\widehat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{f(X_i) \neq Y_i\} =: R_n(f)$$

Theorem 3.5

Suppose that $\mathcal{F}$ is a finite set. Then with probability exceeding $1-\delta$, the excess risk of ERM is upper bounded by

$$R(\widehat{f}_n) - R(f^\star) \leq C\sqrt{\frac{\log(|\mathcal{F}|/\delta)}{n}}.$$

for some universal constant $C > 0$.

- Key proof idea: **union bound argument**
- What if $\mathcal{F}$ is not finite (e.g., the set of linear classifiers)?

— *use VC dimension!*

Back to ERM: finite $\mathcal{F}$

$$\widehat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{f(X_i) \neq Y_i\} =: R_n(f)$$

Theorem 3.5

Suppose that $\mathcal{F}$ is a finite set. Then with probability exceeding $1 - \delta$, the excess risk of ERM is upper bounded by

$$R(\widehat{f}_n) - R(f^\star) \leq C \sqrt{\frac{\log(|\mathcal{F}|/\delta)}{n}}.$$

for some universal constant $C > 0$.

- Key proof idea: **union bound argument**
- What if $\mathcal{F}$ is not finite (e.g., the set of linear classifiers)?

— *use VC dimension!*

- But before going into that, let's first warm up with something simpler

ℓ_2 norm of a sub-Gaussian random vector

- Consider a random vector $\boldsymbol{x} = (X_1, \ldots, X_d)$, where $X_1, \ldots, X_d$ are independent random variables with $\mathbb{E}[X_i] = 0$ and $\|X_i\|_{\psi_2} \leq \sigma$
- Can we establish a non-asymptotic upper bound for $\|\boldsymbol{x}\|_2$?

ℓ_2 norm of a sub-Gaussian random vector

- Consider a random vector $\boldsymbol{x} = (X_1, \ldots, X_d)$, where $X_1, \ldots, X_d$ are independent random variables with $\mathbb{E}[X_i] = 0$ and $\|X_i\|_{\psi_2} \leq \sigma$
- Can we establish a non-asymptotic upper bound for $\|\boldsymbol{x}\|_2$?
- Solution 1: entrywise concentration and union bound

$$\mathbb{P}\big(\|\boldsymbol{x}\|_2 \leq C\sigma\sqrt{d\log(d/\delta)}\big) \geq 1 - \delta$$

for some universal constant $C > 0$

ℓ_2 norm of a sub-Gaussian random vector

- Consider a random vector $\boldsymbol{x} = (X_1, \ldots, X_d)$, where $X_1, \ldots, X_d$ are independent random variables with $\mathbb{E}[X_i] = 0$ and $\|X_i\|_{\psi_2} \le \sigma$
- Can we establish a non-asymptotic upper bound for $\|\boldsymbol{x}\|_2$?
- Solution 1: entrywise concentration and union bound

$$\mathbb{P}\big(\|\boldsymbol{x}\|_2 \le C\sigma\sqrt{d\log(d/\delta)}\big) \ge 1 - \delta$$

  for some universal constant $C > 0$

- Solution 2: uniform concentration using

$$\|\boldsymbol{x}\|_2 = \sup_{\boldsymbol{a} \in \mathcal{S}^{d-1}} \boldsymbol{a}^\top \boldsymbol{x}$$

  where $\mathcal{S}^{d-1} := \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 = 1\}$ is the unit sphere in $\mathbb{R}^d$

  — *could this provide a better concentration bound?*

Operator norm of sub-Gaussian matrix

- Consider a random matrix $\boldsymbol{X} = (X_{i,j})_{1 \le i,j \le d}$ with independent entries that satisfies $\mathbb{E}[X_{i,j}] = 0$ and $\|X_{i,j}\|_{\psi_2} \le \sigma$
- Can we establish a non-asymptotic upper bound for $\|\boldsymbol{X}\|$?
- Operator norm:

$$\|\boldsymbol{X}\| = \sup_{\boldsymbol{a} \in \mathcal{S}^{d-1}} \|\boldsymbol{X}\boldsymbol{a}\|_2 = \sup_{\boldsymbol{a},\boldsymbol{b} \in \mathcal{S}^{d-1}} \boldsymbol{a}^\top \boldsymbol{X} \boldsymbol{b}$$

  where $\mathcal{S}^{d-1} := \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 = 1\}$ is the unit sphere in $\mathbb{R}^d$

A framework for uniform concentration

- **Goal:** upper bounding $\sup_{\boldsymbol{a}\in\mathcal{S}^{d-1}} \boldsymbol{a}^\top \boldsymbol{x}$
- **Step 1: pointwise concentration.** For any fixed $\boldsymbol{a}\in\mathcal{S}^{d-1}$, we can use Hoeffding's inequality to get

$$\mathbb{P}\big(|\boldsymbol{a}^\top\boldsymbol{x}| \le C\sigma\sqrt{\log(1/\delta)}\big) \ge 1-\delta$$

  for some universal constant $C>0$
- **Difficulty:** the unit sphere $\mathcal{S}^{d-1}$ is not a finite set, union bound argument does not work
- **Idea:** find a finite subset $\mathcal{N}$ of $\mathcal{S}^{d-1}$ that is *fine* enough, such that

$$\sup_{\boldsymbol{a}\in\mathcal{S}^{d-1}} \boldsymbol{a}^\top\boldsymbol{x} \overset{?}{\lesssim} \sup_{\boldsymbol{a}\in\mathcal{N}} \boldsymbol{a}^\top\boldsymbol{x} \le C\sigma\sqrt{\log(|\mathcal{N}|/\delta)}$$

  with probability at least $1-\delta$

Epsilon net

- Let (T, d) be a metric space. Consider a subset $K \subset T$ and let $\varepsilon > 0$.

  — *e.g., consider $T = \mathbb{R}^d$, $d(\cdot, \cdot)$ is Euclidean distance, $K = \mathcal{S}^{d-1}$*

- A subset $N \subseteq K$ is called an ε-net of K if every point in K is within distance ε of some point of N, i.e.,

$$\forall\, x \in K, \quad \exists\, x_0 \in N \quad \text{s.t.} \quad d(x, x_0) \le \varepsilon.$$

Theorem 3.6

Let $\mathcal{N}_\varepsilon$ be an ε-net of $\mathcal{S}^{d-1}$. If $\varepsilon < 1$, then for any $\boldsymbol{x} \in \mathbb{R}^d$,

$$\sup_{\boldsymbol{a} \in \mathcal{N}_\varepsilon} \boldsymbol{a}^\top \boldsymbol{x} \le \sup_{\boldsymbol{a} \in \mathcal{S}^{d-1}} \boldsymbol{a}^\top \boldsymbol{x} \le \frac{1}{1-\varepsilon} \sup_{\boldsymbol{a} \in \mathcal{N}_\varepsilon} \boldsymbol{a}^\top \boldsymbol{x},$$

and if $\varepsilon < 1/2$, then for any $\boldsymbol{X} \in \mathbb{R}^{d \times d}$,

$$\sup_{\boldsymbol{a}, \boldsymbol{b} \in \mathcal{N}_\varepsilon} \boldsymbol{a}^\top \boldsymbol{X} \boldsymbol{b} \le \sup_{\boldsymbol{a}, \boldsymbol{b} \in \mathcal{S}^{d-1}} \boldsymbol{a}^\top \boldsymbol{X} \boldsymbol{b} \le \frac{1}{1-2\varepsilon} \sup_{\boldsymbol{a}, \boldsymbol{b} \in \mathcal{N}_\varepsilon} \boldsymbol{a}^\top \boldsymbol{X} \boldsymbol{b}.$$

The covering number

Covering number: the smallest possible cardinality of an ε-net of K, denoted by $\mathcal{N}(K, \varepsilon)$

Theorem 3.7

The covering number of $\mathcal{S}^{d-1}$ is upper bounded by

$$\mathcal{N}(\mathcal{S}^{d-1}, \varepsilon) \leq \left(\frac{2}{\varepsilon} + 1 \right)^d$$

ℓ_2 norm of sub-Gaussian random vector

- **Goal:** upper bounding $\sup_{\boldsymbol{a}\in\mathcal{S}^{d-1}} \boldsymbol{a}^\top \boldsymbol{x}$
- **Step 1: pointwise concentration.** For any fixed $\boldsymbol{a}\in\mathcal{S}^{d-1}$, we can use Hoeffding's inequality to get

$$\mathbb{P}\big(|\boldsymbol{a}^\top \boldsymbol{x}| \le C_1\sigma\sqrt{\log(1/\delta)}\big) \ge 1-\delta$$

  for some universal constant $C_1 > 0$

ℓ_2 norm of sub-Gaussian random vector

- **Goal:** upper bounding $\sup_{\boldsymbol{a}\in\mathcal{S}^{d-1}} \boldsymbol{a}^\top \boldsymbol{x}$
- **Step 1: pointwise concentration.** For any fixed $\boldsymbol{a} \in \mathcal{S}^{d-1}$, we can use Hoeffding's inequality to get

$$\mathbb{P}\big(|\boldsymbol{a}^\top \boldsymbol{x}| \leq C_1 \sigma \sqrt{\log(1/\delta)}\big) \geq 1-\delta$$

  for some universal constant $C_1 > 0$

- **Step 2: uniform concentration over an $1/2$-net.** Let $\mathcal{N}_{1/2}$ be the smallest $1/2$-net of $\mathcal{S}^{d-1}$. By union bound argument and Theorem 3.7,

$$\mathbb{P}\big(\sup_{\boldsymbol{a}\in\mathcal{N}_{1/2}} |\boldsymbol{a}^\top \boldsymbol{x}| \leq C_2 \sigma \sqrt{d\log(1/\delta)}\big) \geq 1-\delta$$

  for some universal constant $C_2 > 0$

ℓ_2 norm of sub-Gaussian random vector

- **Goal:** upper bounding $\sup_{\boldsymbol{a}\in\mathcal{S}^{d-1}} \boldsymbol{a}^\top \boldsymbol{x}$
- **Step 1: pointwise concentration.** For any fixed $\boldsymbol{a}\in\mathcal{S}^{d-1}$, we can use Hoeffding's inequality to get

  $$\mathbb{P}\big(|\boldsymbol{a}^\top \boldsymbol{x}| \le C_1\sigma\sqrt{\log(1/\delta)}\big) \ge 1-\delta$$

  for some universal constant $C_1 > 0$
- **Step 2: uniform concentration over an $1/2$-net.** Let $\mathcal{N}_{1/2}$ be the smallest $1/2$-net of $\mathcal{S}^{d-1}$. By union bound argument and Theorem 3.7,

  $$\mathbb{P}\big(\sup_{\boldsymbol{a}\in\mathcal{N}_{1/2}} |\boldsymbol{a}^\top \boldsymbol{x}| \le C_2\sigma\sqrt{d\log(1/\delta)}\big) \ge 1-\delta$$

  for some universal constant $C_2 > 0$
- **Step 3: approximation.** By Theorem 3.6,

  $$\mathbb{P}\big(\|\boldsymbol{x}\|_2 \le C_3\sigma\sqrt{d\log(1/\delta)}\big) \ge 1-\delta$$

  for some universal constant $C_3 > 0$

Operator norm of sub-Gaussian random matrix

- **Goal:** upper bounding $\sup_{\boldsymbol{a},\boldsymbol{b}\in\mathcal{S}^{d-1}} \boldsymbol{a}^\top \boldsymbol{X}\boldsymbol{b}$
- **Step 1: pointwise concentration.** For any fixed $\boldsymbol{a},\boldsymbol{b}\in\mathcal{S}^{d-1}$, we can use Hoeffding's inequality to get

  $$\mathbb{P}\big(|\boldsymbol{a}^\top \boldsymbol{X}\boldsymbol{b}| \le C_1\sigma\sqrt{\log(1/\delta)}\big) \ge 1-\delta$$

  for some universal constant $C_1 > 0$

Operator norm of sub-Gaussian random matrix

- **Goal:** upper bounding $\sup_{\boldsymbol{a},\boldsymbol{b}\in\mathcal{S}^{d-1}} \boldsymbol{a}^\top \boldsymbol{X}\boldsymbol{b}$
- **Step 1: pointwise concentration.** For any fixed $\boldsymbol{a},\boldsymbol{b}\in\mathcal{S}^{d-1}$, we can use Hoeffding's inequality to get

  $$\mathbb{P}\big(|\boldsymbol{a}^\top \boldsymbol{X}\boldsymbol{b}| \le C_1\sigma\sqrt{\log(1/\delta)}\big) \ge 1-\delta$$

  for some universal constant $C_1 > 0$
- **Step 2: uniform concentration over an $1/4$-net.** Let $\mathcal{N}_{1/4}$ be the smallest $1/4$-net of $\mathcal{S}^{d-1}$. By union bound argument and Theorem 3.7,

  $$\mathbb{P}\big(\sup_{\boldsymbol{a},\boldsymbol{b}\in\mathcal{N}_{1/4}} |\boldsymbol{a}^\top \boldsymbol{X}\boldsymbol{b}| \le C_2\sigma\sqrt{d\log(1/\delta)}\big) \ge 1-\delta$$

  for some universal constant $C_2 > 0$

Operator norm of sub-Gaussian random matrix

- **Goal:** upper bounding $\sup_{\boldsymbol{a},\boldsymbol{b}\in\mathcal{S}^{d-1}} \boldsymbol{a}^\top \boldsymbol{X}\boldsymbol{b}$
- **Step 1: pointwise concentration.** For any fixed $\boldsymbol{a},\boldsymbol{b}\in\mathcal{S}^{d-1}$, we can use Hoeffding's inequality to get

  $$\mathbb{P}\big(|\boldsymbol{a}^\top \boldsymbol{X}\boldsymbol{b}| \le C_1\sigma\sqrt{\log(1/\delta)}\big) \ge 1-\delta$$

  for some universal constant $C_1 > 0$
- **Step 2: uniform concentration over an $1/4$-net.** Let $\mathcal{N}_{1/4}$ be the smallest $1/4$-net of $\mathcal{S}^{d-1}$. By union bound argument and Theorem 3.7,

  $$\mathbb{P}\big(\sup_{\boldsymbol{a},\boldsymbol{b}\in\mathcal{N}_{1/4}} |\boldsymbol{a}^\top \boldsymbol{X}\boldsymbol{b}| \le C_2\sigma\sqrt{d\log(1/\delta)}\big) \ge 1-\delta$$

  for some universal constant $C_2 > 0$
- **Step 3: approximation.** By Theorem 3.6,

  $$\mathbb{P}\big(\|\boldsymbol{X}\|_2 \le 2C_2\sigma\sqrt{d\log(1/\delta)}\big) \ge 1-\delta$$

  for some universal constant $C_3 > 0$

VC dimension

- Let $\mathcal{F}$ be a class of binary functions on the domain $\mathcal{X}$.
- **Shattering:** a set of points $\{x_1, \ldots, x_k\} \subseteq \mathcal{X}$ is shattered by $\mathcal{F}$ if for every possible labeling $\{0,1\}^k$, there exists a function $f \in \mathcal{F}$ that realizes the labeling.
- The **VC dimension** of $\mathcal{F}$, denoted VC$(\mathcal{F})$, is the largest integer k such that there exists a set of k points in $\mathcal{X}$ that can be *shattered* by $\mathcal{F}$.
- Examples:
  - When $\mathcal{X} = \mathbb{R}^2$, $\mathcal{F} =$ linear classifiers, we have vc$(\mathcal{F}) = 3$
  - In general, when $\mathcal{X} = \mathbb{R}^d$, $\mathcal{F} =$ linear classifiers, then vc$(\mathcal{F}) = d + 1$

Bounding excess risk via VC dimension

$$\widehat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{f(X_i) \neq Y_i\} =: R_n(f)$$

Theorem 3.8

Suppose that $\mathcal{F}$ is a class of Boolean function with $\mathsf{vc}(\mathcal{F}) < \infty$. Then with probability exceeding $1 - \delta$,

$$R(\widehat{f}_n) - R(f^\star) \le C \sqrt{\frac{\mathsf{vc}(\mathcal{F}) \log(1/\delta)}{n}}$$

for some universal constant $C > 0$.

Implications:

- For $\mathcal{F} =$ linear classifiers in $\mathbb{R}^d$, the excess risk is $O(\sqrt{d/n})$.