IFT 6085 - Lecture 2
(lecture title here)

Scribe(s): [your name(s)] **Instructor:** Ioannis Mitliagkas

Guidelines for scribing

This section gives some instructions for scribing and should not be present in the final notes. The rest of this document is sample context to demonstrate usage of this template. The sample content does not include proofs, but your notes should include the proofs we covered and key elements of the in-class discussion.

Scribing accounts for 10% of your final grade. Your notes will be evaluated on the following elements:

- Completeness: cover all major ideas discussed in class. In cases where we covered the proofs, include the proofs in the notes and try to explain all steps that use a property or result beyond basic algebra. Use latex's proof environment.

  ```
  \begin{proof}
  ...
  \end{proof}
  ```

- Clarity: use full sentences to clearly set up and describe the material. Include key elements of the discussion we has in class to help explain/unpack the material after the technical results are presented.
- Credit: use the .bib file to cite the sources where we draw our results from. See the assigned reading material on the website for the sources. Try to be precise, citing specific chapters or theorems from the sources.

1 Summary

In the previous lecture we [short summary of last lecture]

In this lecture we [short summary of this lecture]

2 [sample content] Binary Classification

In this section we introduce the basic elements of supervised classification.

Definition 1 (Observation)**.** *A d-dimensional vector*

$$x \in \mathcal{X} \subseteq \mathbb{R}^d$$

where $\mathcal{X}$ is a measurable space equipped with a σ-algebra.

Definition 2 (Class)**.** *Binary label assigned to measurement*

$$y \in \{-1, 1\}$$

Binary classification entails most of the challenges found in the multiclass problem.

Definition 3 (Classifier)**.** *Mapping from observations to labels*

$$g : \mathcal{X} \to \{-1, 1\}$$

For a pair (x, y)*,* $g(x) \neq y$ *is the* error *event.*

Assumption 4 (Probabilistic Setting)**.** *Let* (X, Y) *be a random pair. Distribution described by*

$$\mathbb{P}\{X \in A\}, \quad A \in \sigma\text{-algebra}$$

and the a posteriori probability

$$\eta(x) = \mathbb{P}\{Y = 1 | X = x\}.$$

Definition 5 (Probability of Error)**.**

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

General goal: Find g that minimizes the probability of error. If we are given $\eta(x)$ we can construct classifier with minimal probability of error.

Definition 6 (Bayes Classifier)**.**

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

then $L(g^*) \leq L(g)$ *for any* g*.*

$$L^* \triangleq L(g^*)$$

is called the Bayes risk *or* Bayes error*.*

Definition 7 (Sample)**.** *Sequence of i.i.d. random pairs*

$$(X_1, Y_1), \ldots, (X_n, Y_n) \sim (X, Y)$$

n*-length data set*

$$D_n = (X1, Y_1, \ldots, X_n, Y_n)$$

We construct a classifier based on the data

$$g_n(X) = g_n(X; X_1, Y_1, \ldots, X_n, Y_n)$$

Fixing the data, we measure the performance of g_n using the *conditional probability of error*

Definition 8 (Conditional Probability of Error)**.**

$$L(g_n) = \mathbb{P}\{g_n(X) \neq Y | D_n\}$$

Goal: Find g_n that minimizes probability of error.

3 Empirical Risk Minimization

We restrict to recovering classifiers for a predetermined set, i.e. $g \in \mathcal{C}$.

Use our data, D_n, to estimate $L(g)$ and select classifier.

Definition 9 (Empirical Error (or Risk))**.**

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

Let g_n^* denote the minimizer of empirical risk.

$$L_n(g_n^*) \leq L_n(g), \quad \forall g \in \mathcal{C}$$

Its probability of error (on unseen data), $L(g_n^*) = \mathbb{P}(g_n^*(X) \neq Y | D_n)$ satisfies,

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

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

Uniform deviation bound on RHS controls the probability of error of selected classifier g_n^*.

Bound can be quite loose (though sharp in minimax sense – we won't cover this). Improvements later in paper.

Definition 10 (Empirical Averages - Notation)**.** *For $X_1, \ldots, X_n$ iid random variables, with values in $\mathcal{X}$.*
Let $\mathcal{F}$ denote class of bounded functions $\mathcal{X} \to [-1, 1]$.
Expectation

$$Pf \triangleq \mathbb{E}[f(X_1)]$$

Empirical Average

$$P_n f = (1/n)\sum_{i=1}^{n} f(X_i)$$

Our goal is to get good bounds for the uniform deviation

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

Our first bound comes using bounded differences.

Theorem 11 (Bounded Differences)**.** *For function $g : \mathcal{X}^n \to \mathbb{R}$ with bounded differences parameters $(c_i)_{i=1}^n$,*

$$\mathbb{P}(|Z - \mathbb{E}Z| > t) \leq 2\exp -2t^2/C$$

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

Proof. proof goes here □

Note that Z satisfies bounded differences with $c_i = 2/n$. Then with probability at least $1 - \delta$,

$$\sup_{f\in\mathcal{F}} |Pf - P_n f| \leq \mathbb{E}\sup_{f\in\mathcal{F}} |Pf - P_n f| + \sqrt{\frac{2\log 1/\delta}{n}}.$$

Now we just need to focus on the expectation term. We'll use **symmetrization**.

Let sample $X_1', \ldots, X_n'$ be independent of original sample and distributed identically. Then,

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

Now, introduce independent Rademacher random variables, $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = 1/2$.
Then,

$$\mathbb{E}\sup_{f\in\mathcal{F}} |P_n' f - P_n f| \leq 2\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i f(X_i)\right|\right].$$

Definition 12 (Rademacher Average). *Let $A \in \mathbb{R}^n$ be bounded set of vectors $a = (a_1, \ldots, a_n)$. Then*

$$R_n(A) \triangleq \mathbb{E} \sup_{a \in A} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i a_i \right|.$$

is called the Rademacher Average of A.

Definition 13. $\mathcal{F}(x_1^n)$

Given sequence $x_1, \ldots, x_n \in \mathcal{X}$, let $\mathcal{F}(x_1^n)$ denote the class of vectors $(f(x_1), \ldots, f(x_n))$ with $f \in \mathcal{F}$.

Theorem 14. *With probability at least $1 - \delta$,*

$$\sup_{f \in \mathcal{F}} |Pf - P_n f| \leq 2\mathbb{E}R_n(\mathcal{F}(X_1^n)) + \sqrt{\frac{2 \log 1/\delta}{n}}$$

and

$$\sup_{f \in \mathcal{F}} |Pf - P_n f| \leq 2R_n(\mathcal{F}(X_1^n)) + \sqrt{\frac{2 \log 2/\delta}{n}}$$

For second statement, use the fact that $R_n(\mathcal{F}(X_1^n))$ satisfies bounded differences. Relevant results in Bartlett et al. [1].

Second bound is **data dependent**.

Theorem 15 (Properties of Rademacher Averages). *If $A = \{a^{(1)}, \ldots, a^{(N)}\} \in \mathbb{R}^n$ is a finite set,*

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

Contraction Principle

If $\phi : \mathbb{R} \to \mathbb{R}$, with $\phi(0) = 0$ and L_ϕ-lipschitz and $\phi \circ A$ is the set of all $(\phi(a_1), \ldots, \phi(a_n)) \in \mathbb{R}^n$ with $a \in A$,

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

We will use Hoeffding's inequality to prove the first property.

We often want to get simple upper bounds on Rademacher Averages.

Example 16 (Indicator Functions). *From our classification example*

$$\mathcal{F} = \left\{ \mathbb{I}_{\{(x,y): g(x) \neq y\}} : g \in \mathcal{C} \right\}$$

Definition 17 (VC Shatter Coefficient). *Let $\mathcal{F}$ be the class of indicator functions.*

*For any collection of points $x_1^n = (x_1, \ldots, x_n)$, $\mathcal{F}(x_1^n)$ is a finite set and its **cardinality** is denoted by*

$$\mathbb{S}_{\mathcal{F}}(x_1^n) \triangleq |\mathcal{F}(x_1^n)| \leq 2^n.$$

The shatter coefficient is a measure of richness of function class $\mathcal{F}$.

Using its properties, we can upper bound the Rademacher Average using the shatter coefficient.

$$R_n(\mathcal{F}(x_1^n) \leq \sqrt{\frac{2 \log \mathbb{S}_{\mathcal{F}}(x_1^n)}{n}}.$$

Then,

$$\mathbb{E}\sup_{f\in\mathcal{F}}|Pf - P_n f| \leq 2\mathbb{E}\sqrt{\frac{2\log \mathbb{S}_{\mathcal{F}}(X_1^n)}{n}}.$$

Definition 18 (VC Dimension)**.**

"cardinality of the largest set of points a classifier can shatter"

If $A \subset \{-1,1\}^n$, then the size V of the largest set of indices $\{i_1,\dots,i_V\} \subset \{1,\dots,n\}$ such that:

for each binary V-vector, $b \in \{-1,1\}^V$, there exists $a = (a_1,\dots,a_n) \in A$ such that $(a_{i_1},\dots,a_{i_V}) = b$.

Theorem 19 (Sauer's Lemma)**.** *For any set $A \subset \{-1,1\}^n$,*

$$|A| \leq \sum_{i=0}^{V}\binom{n}{i} \leq (n+1)^V$$

where V is the VC-dimension of A.

Then,

$$\log \mathbb{S}_{\mathcal{F}}(x_1^n) \leq V(x_1^n)\log(n+1)$$

where $V(x_1^n)$ is the VC-dimension of $\mathcal{F}(x_1^n)$, and

$$\mathbb{E}\sup_{f\in\mathcal{F}}|Pf - P_n f| \leq 2\mathbb{E}\sqrt{\frac{2V(X_1^n)\log(n+1)}{n}}.$$

To make this result distribution-agnostic, let

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

Theorem 20 (Vapnik-Chervonenkis Inequality)**.** *For all distributions,*

$$\mathbb{E}\sup_{f\in\mathcal{F}}(Pf - P_n f) \leq 2\sqrt{\frac{2V\log(n+1)}{n}}.$$

4. Minimizing Cost Functions

When VC-dimension smaller than n, empirical risk minimization is guaranteed to work well.

Restricted classes $\mathcal{C}$ imply small VC-dimension, hence good generalization guarantees.

However, approximation error $\inf_{g\in\mathcal{C}} L(g) - L^*$ becomes an issue.

Furthermore, minimizing the empirical $L_n(g)$ is computationally hard (even simple cases are NP-hard). For example, $\mathcal{X} \in \mathbb{R}^d$ and $\mathcal{C}$ is the class of hyperplanes.

We'll soften things up.

Definition 21 (New parametrization)**.** *For functions, $f : \mathcal{X} \to \mathbb{R}$, we consider classifiers of the form,*

$$g_f(x) = \begin{cases} 1, & \text{if } f(x) \geq 0 \\ -1, & \text{otherwise} \end{cases}.$$

Then the probability of error of g is written as

$$L(f) \triangleq L(g_f) = \mathbb{P}(\operatorname{sgn}(f(X)) \neq Y) \leq \mathbb{E}\mathbb{I}_{f(X)Y<0}$$

Definition 22 (Cost Function)**.** *Let $\phi : \mathbb{R} \to \mathbb{R}_+$ be a non-negative function such that*

$$\phi(x) \geq \mathbb{I}_{x>0}$$

Examples include, $\phi(x) = \exp(x)$, $\phi(x) = \log_2(1 + \exp(x))$ and $\phi(x) = (1 + x)_+$.

Definition 23 (Cost Functional)**.**

$$A(f) = \mathbb{E}\phi(-f(X)Y)$$

and the empirical cost functional

$$A_n(f) = \frac{1}{n}\sum_{i=1}^{n} \phi(-f(X_i)Y_i).$$

By Definition 22,

$$L(f) \leq A(f)$$

and

$$L_n(f) \leq A_n(f).$$

Theorem 24 (Probability of Error is Close to Empirical Cost)**.** *For f_n chosen from class $\mathcal{F}j$ based on data D_n. Let B denote a uniform upper bound on $\phi(-f(x)y)$ and let L_ϕ be the Lipschitz constant of ϕ. Then with probability at least $1 - \delta$,*

$$L(f_n) \leq A_n(f_n) + 2L_\phi \mathbb{E}R_n(\mathcal{F}(X_1^n)) + B\sqrt{\frac{2\log(1/\delta)}{n}}.$$

4.1.1 Weighted Voting Schemes

In boosting and bagging simple classifiers combined make powerful ensembles.

Definition 25 (Weighted Voting Scheme)**.**

$$\mathcal{F}_\lambda = \left\{ f(x) = \sum_{j=1}^{N} c_j g_j(x) : N \in \mathbb{N}, \sum_{j=1}^{N} |c_j| \leq \lambda, g_1, \ldots, g_N \in \mathcal{C} \right\}$$

*$\mathcal{C}$ the class of **base classifiers**, $g : \mathcal{X} \to \{-1, 1\}$.*

Using the fact that the Rademacher average of the absolute convex hull of A is the same as that of A [2], and our VC-dimension upper bounds, we get

$$R_n(\mathcal{F}_\lambda(X_1^n)) \leq \lambda R_n(\mathcal{C}(X_1^n)) \leq \lambda\sqrt{\frac{2V_\mathcal{C}\log(n+1)}{n}}.$$

This implies the following guarantee on the probability of error of the learned classifier:

$$L(f_n) \leq A_n(f_n) + 2L_\phi\lambda\sqrt{\frac{2V_\mathcal{C}\log(n+1)}{n}} + B\sqrt{\frac{2\log(1/\delta)}{n}}.$$

The great feature of this bound, is that it only depends on the VC-dimension of the base class, $V_\mathcal{C}$.

Skipping margin error, strictly convex cost functions and kernel methods in favour of improved bounds due to better use of concentration tools.

5 Tighter Bounds

Example 26 (Motivating Example)**.**

Consider fixed function $f : \mathcal{X} \to \{0, 1\}$.

Then $P_n f$ is the average of n independent Bernoullis with parameter Pf.

Bounded Differences

$$Pf - P_n f \leq \sqrt{\frac{2\log(1/\delta)}{n}}$$

Bernstein

$$Pf - P_n f \leq \sqrt{\frac{2Var(f)\log(1/\delta)}{n}} + \frac{2\log(1/\delta)}{3n}$$

Since f takes values in $\{0, 1\}$,

$$Var(f) = Pf(1 - Pf) \leq Pf,$$

and the second bound is tighter.

Now we apply this intuition to empirical risk minimization.

We saw that we get a significant improvement using information on the variance.

Going for a uniform bound on $\sup_{f\in\mathcal{F}}(Pf - P_n f)$ would force us to use the worst case variance over the whole class $\mathcal{F}$.

Instead, we will scale each individual difference by $\sqrt{Pf}$ to account for variability in the variance each f induces. Our quantity of interest becomes,

$$\sup_{f\in\mathcal{F}} \frac{Pf - P_n f}{\sqrt{Pf}}.$$

By symmetrization on the tail probabilities,

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}} \frac{Pf - P_n f}{\sqrt{Pf}} \geq t\right\} \leq 2\mathbb{P}\left\{\sup_{f\in\mathcal{F}} \frac{P'_n f - P_n f}{\sqrt{(P_n f + P'_n f)/2}} \geq t\right\}$$

And by introducing Rademacher random variables,

$$2\mathbb{P}\left\{\sup_{f\in\mathcal{F}} \frac{P'_n f - P_n f}{\sqrt{(P_n f + P'_n f)/2}} \geq t\right\} = 2\mathbb{E}\left[\mathbb{P}_\sigma\left\{\sup_{f\in\mathcal{F}} \frac{\frac{1}{n}\sum_{i=1}^n \sigma_i(f(X'_i) - f(X_i))}{\sqrt{(P_n f + P'_n f)/2}} \geq t\right\}\right]$$

This leads to the following result.

Theorem 27. *Let $\mathcal{F}$ be a class of functions, $f : \mathcal{X} \to \{0, 1\}$. With probability at least $1 - \delta$, all $f \in \mathcal{F}$ statisty*

$$\frac{Pf - P_n f}{\sqrt{Pf}} \leq 2\sqrt{\frac{\log \mathbb{S}_\mathcal{F}(X_1^{2n}) + \log(4/\delta)}{n}}$$

and

$$\frac{P_n f - Pf}{\sqrt{P_n f}} \leq 2\sqrt{\frac{\log \mathbb{S}_\mathcal{F}(X_1^{2n}) + \log(4/\delta)}{n}}$$

Applied to our empirical risk minimization problem, the above result implies.

Theorem 28. *Let g_n^* be the empirical risk minimizer in class $\mathcal{C}$ of VC dimension V. Then, with probability at least $1-\delta$,*

$$L(g_n^*) \leq L_n(g_n^*) + 2\sqrt{L_n(g_n^*)\frac{2V\log(n+1)+\log(4/\delta)}{n}} + 4\frac{2V\log(n+1)+\log(4/\delta)}{n}$$

In the extreme case when there exists a classifier in $\mathcal{C}$ that classifies without error. Then,

$$L_n(g_n^*) = 0$$

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

$$L(g_n^*) - \inf_{g\in\mathcal{C}} L(g) \leq 4\frac{2V\log(n+1)+\log(4/\delta)}{n}$$

Significantly improved rate.

More generally, if $L(g') = \inf_{g\in\mathcal{C}} L(g)$, we have

$$L_n(g_n^*) \leq L_n(g') = L_n(g') - L(g') + L(g')$$

and using Bernstein, we get w.p. at least $1-\delta$,

$$L_n(g_n^*) - L(g') \leq \sqrt{\frac{2L(g')\log(1/\delta)}{n}} + \frac{2\log(1/\delta)}{3n}$$

which along with the last Theorem, gives us the following result.

Theorem 29. *There exists a constant C s.t. with probability at least $1-\delta$,*

$$L(g_n^*) \inf_{g\in\mathcal{C}} L(g) \leq C\left(\sqrt{\inf_{g\in\mathcal{C}} L(g)\frac{V\log n+\log(1/\delta)}{n}} + \frac{V\log n+\log(1/\delta)}{n}\right)$$

References

[1] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *Annals of Statistics*, pages 1497–1537, 2005.

[2] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.