Topics in Learning Theory

Lecture 4: Binary Classification

Topics

- VC-dimension
  - empirical L_∞ generalization bound
  - empirical L_2-cover and Rademacher bound
- Margin bounds
  - Using L_∞ covering number
  - Simple data-dependent bounds
  - (Using Rademacher complexity: next time)

Binary-Classification Problem

- Predict binary label $\boldsymbol{y} \in \{\pm \mathbf{1}\}$.
- Classifier $f(x)$:
  - binary valued: $\boldsymbol{f}(\boldsymbol{x}) \in \{\pm \mathbf{1}\}$
- Classification error loss: $\boldsymbol{\phi}(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) = \boldsymbol{I}(\boldsymbol{f}(\boldsymbol{X}) \neq \boldsymbol{Y})$
  - I : indicator function.

Binary Linear Classifier

- Let $x \in R^d$, take $\mathcal{H} = \{f(X) = 2I(w^T X + b \geq 0) - 1 : w \in R^d, b \in R\}$
- Empirical risk minimization (minimize classification error)

$$[\hat{w}, \hat{b}] = \arg \min_{w \in R^d, b \in R} \sum_{i=1}^{n} I((w^T X_i + b) Y_i \leq 0).$$

- What is the performance of this algorithm?
- What is the performance of empirical risk minimization with general function class $\mathcal{H}$?

Covering number for binary functions

- Given a function family $\mathcal{H}$ of $f(x)$ that takes $\{0,1\}$ values, what is its empirical L_∞ covering number?

$$L_\infty(\mathcal{H}, 0, |S_n) = |\{[\phi(f(X_1), Y_1), \ldots, \phi(f(X_n), Y_n)] : f \in \mathcal{H}\}|.$$

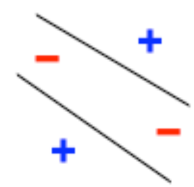
- Partial answer, a binary-valued family $\mathcal{H}$ has a unique number called VC-dimension $VC(\mathcal{H})$.
  - if this number is finite, then the covering number is polynomial in n
  - if this number is infinite, then there exists a distribution such that the empirical covering number is 2^n.

VC dimension

- Shattering: a function class $\mathcal{H}$ is said to shatter a set of data points $(X_1, X_2, \ldots, X_n)$ if, for all assignments of labels to those points $(Y_1, \ldots, Y_n)$, there exists a f such that the model f makes no errors when evaluating that set of data points: $f(X_i) = Y_i$ for all i.
  - Any label can be explained
  - Complete overfitting
- VC dimension $VC(\mathcal{H})$: the maximum n such that there exist data point of cardinality n that can be shattered.

Example: linear separator in 2d

- In 2d:
  - data $x \in R^2$
  - $\mathcal{H} = \{w^T x + b : w \in R^2, b \in R\}$
- There exists 3 points $[0, 0], [0, 1], [1, 0]$ that can be shattered by $\mathcal{H}$
- Any four points cannot be shattered:



- So VC dimension is 3
- More generally: d dimensional linear classifier has VC dimension $d + 1$

VC dimension and covering number

- If VC dimension is infinity, then for any n, there is a sample of size n such that one can fit any data — no generalization
- What about finite VC dimension $= VC(\mathcal{H})$?
- Sauer's Lemma $(n \geq d)$:

$$N_\infty(\mathcal{H}, 0|S_n) \leq \sum_{i=0}^{d} \binom{n}{i} \leq (en/d)^d.$$

- Empirical L_2 cover bound (can bound Rademacher complexity): there exists constant $C > 0$ such that

$$N_2(\mathcal{H}, \epsilon|S_n) \leq C\,(1/\epsilon)^d$$

Generalization Bound using VC dimension: Rademacher complexity bound

- Rademacher complexity (using L_2-covering number and chaining bound): exists constant C

$$R(\mathcal{H}|S_n) \leq C\sqrt{d/n}$$

- Generalization bound:

$$E_{X,Y}\phi(f(X),Y) \leq \frac{1}{n}\sum_{i=1}^{n}\phi(f(X_i),Y_i) + C\sqrt{d/n} + \sqrt{\frac{\ln(1/\eta)}{2n}}.$$

- Draw-back: does not give $O(1/n)$ rate

Generalization Bound using L_∞-cover bound

- Learning bound using empirical L_∞ covering number: let $Q(f)$ be a function depending on f (its complexity), we want to prove an inequality with probability $1-\eta$:

$$\sup_{f\in\mathcal{H}}[\mathbf{E}_{X,Y}\phi(f(X),Y) - \frac{1}{n}\sum_{i=1}^{n}\phi(f(X_i),Y_i) - Q(f)] < 0,$$

where given $\lambda > 0$, we take $Q(f)$ of the form

$$Q(f) = [2(e^{\lambda}-\lambda-1)E\phi(f(X),Y) + \frac{1}{n}\ln(N_\infty(\mathcal{H},0|2n)/\eta)]/\lambda$$

Derivation (let $N_\infty(\mathcal{H},0|n) = \sup_{S_n} N_\infty(\mathcal{H},0|S_n)$):

$$
\begin{aligned}
&P[\sup_{f\in\mathcal{H}}[\mathbf{E}_{X,Y}\phi(f(X),Y) - \frac{1}{n}\sum_{i=1}^{n}\phi(f(X_i),Y_i) - Q(f)] \geq 0] \\
\leq & E_{S_n} \sup_{f\in\mathcal{H}} e^{\lambda n[\mathbf{E}_{X,Y}\phi(f(X),Y) - \frac{1}{n}\sum_{i=1}^{n}\phi(f(X_i),Y_i) - Q(f)]} \\
\leq & E_{S_n,S'_n} \sup_{f\in\mathcal{H}} e^{\lambda \sum_{i=1}^{n}[\phi(f(X'_i),Y'_i) - \phi(f(X_i),Y_i)] - \lambda n Q(f)} \\
\leq & N_\infty(\mathcal{H},0|2n) \sup_{f\in\mathcal{H}} E_{S_n,S'_n} e^{\lambda \sum_{i=1}^{n}[\phi(f(X'_i),Y'_i) - \phi(f(X_i),Y_i)] - \lambda n Q(f)} \\
\leq & N_\infty(\mathcal{H},0|2n) \sup_{f\in\mathcal{H}} e^{n(e^\lambda - \lambda - 1)Var(\phi(f(X'_1),Y'_1) - \phi(f(X_1),Y_1)) - \lambda n Q(f)} \\
\leq & N_\infty(\mathcal{H},0|2n) \sup_{f\in\mathcal{H}} e^{2n(e^\lambda - \lambda - 1)E\phi(f(X),Y) - \lambda n Q(f)} = \eta.
\end{aligned}
$$

Bounds for empirical risk minimization using VC dimension

- $\ln N_\infty(\mathcal{H}, 0|2n) \leq d \ln(2en/d)$ ($d = VC(\mathcal{H})$)
- Taking $\lambda \leq 0.5$, and note $(e^\lambda - \lambda - 1)/\lambda^2$ is increasing function, we have

$$(1 - 1.2\lambda)\mathbf{E}_{X,Y}\phi(\hat{f}(X), Y) \leq \frac{1}{n}\sum_{i=1}^{n}\phi(\hat{f}(X_i), Y_i) + \frac{1}{\lambda n}\ln(N_\infty(\mathcal{H}, 0|2n)/\eta)$$

thus $\mathbf{E}_{X,Y}\phi(f(X), Y) \leq \frac{1+3\lambda}{n}\sum_{i=1}^{n}\phi(\hat{f}(X_i), Y_i) + \frac{2.5}{\lambda n}\ln(N_\infty(\mathcal{H}, 0|2n)/\eta)$, which implies

$$\mathbf{E}_{X,Y}\phi(f(X), Y) \leq \frac{1 + 3\lambda}{n}\sum_{i=1}^{n}\phi(\hat{f}(X_i), Y_i) + \frac{2.5}{\lambda n}\ln(N_\infty(\mathcal{H}, 0|2n)/\eta).$$

Generalization Bounds for Binary Linear Classifier

- $x \in R^d$, linear classifier $2I(\hat{w}^T x + \hat{b} \geq 0) - 1$:

$$[\hat{w}, \hat{b}] = \arg \min_{w \in R^d, b \in R} \sum_{i=1}^{n} I((w^T X_i + b)Y_i \leq 0).$$

- VC dimension is $d + 1$, thus $\exists C > 0$:

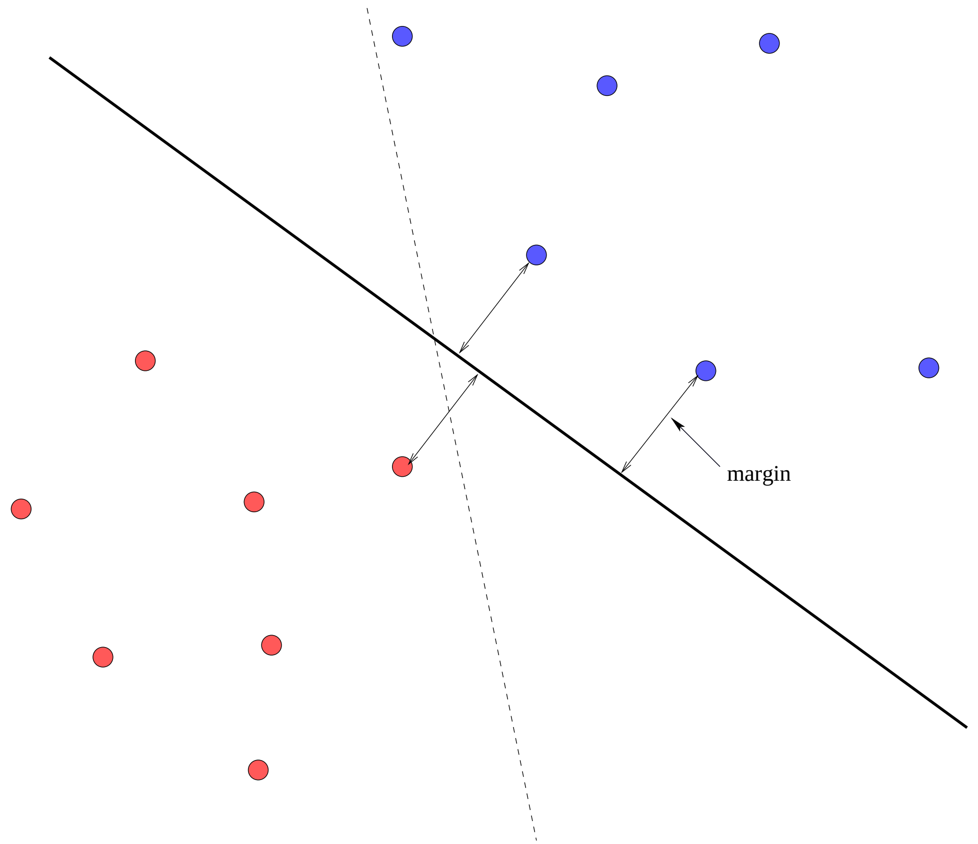
$$\mathbf{E}_{X,Y} I((w^T X + b)Y \leq 0) \leq \frac{2}{n} \sum_{i=1}^{n} I((w^T X_i + b)Y_i \leq 0) + \frac{C d \ln n}{n}.$$

$$\mathbf{E}_{X,Y} I((w^T X + b)Y \leq 0) \leq \frac{1}{n} \sum_{i=1}^{n} I((w^T X_i + b)Y_i \leq 0) + C\sqrt{d/n}.$$

Comments on VC dimension

- For d dimensional linear classifier, requires $n \geq \Omega(d)$ examples.
  - not suitable for large dimensional data where $n \ll d$.
- Characterizes worst case performance bounds:
  - performance can be much better in reality if the distribution is not worst-case
- How to characterize good distribution?
  - in particular, how to handle large dimension

Margin: are all linear separator equally good?



Margin Bound

- Let $f(x) \in \mathcal{H}$ be a real valued function
  - e.g. linear function: $f(x) = w^T x$ ($x \in R^d$)
- To bound $\mathbf{E}_{X,Y} I(f(X)Y \leq 0)$ in term of $\frac{1}{n}\sum_{i=1}^{n} I(f(X_i)Y_i \leq \gamma)$
  - $\gamma > 0$ is margin
- Want a bound of the form:

$$\mathbf{E}_{X,Y} I(\hat{f}(X)Y \leq 0) \leq \frac{1}{n}\sum_{i=1}^{n} I(\hat{f}(X_i)Y_i \leq \gamma) + Q(f).$$

Margin Bound using L_∞-cover bound

- Let $Q(f)$ be a function depending on f (its complexity). Given $\lambda > 0$ and $\alpha = 2(e^\lambda - \lambda - 1)/\lambda$, we want to prove an inequality with probability $1 - \eta$:

$$\sup_{f \in \mathcal{H}} [(1-\alpha)\mathbf{E}_{X,Y} I(\hat{f}(X)Y \le 0) - \frac{1}{n}\sum_{i=1}^{n} I(\hat{f}(X_i)Y_i \le \gamma) - Q(f)] < 0,$$

where we take $Q(f)$ of the form

$$Q(f) = \frac{1}{\lambda n} \ln(N_\infty(\mathcal{H}, \gamma/2|2n)/\eta).$$

Derivation (let $N_\infty(\mathcal{H}, \gamma/2|n) = \sup_{S_n} N_\infty(\mathcal{H}, \gamma/2|S_n)$):

$$
\begin{aligned}
&P[\sup_{f\in\mathcal{H}}[(1-\alpha)\mathbf{E}_{X,Y}I(\hat{f}(X)Y\le 0)-\frac{1}{n}\sum_{i=1}^{n}I(\hat{f}(X_i)Y_i\le\gamma)-Q(f)]\ge 0]\\
\le& E_{S_n}\sup_{f\in\mathcal{H}} e^{\lambda n[(1-\alpha)\mathbf{E}_{X,Y}I(\hat{f}(X)Y\le 0)-\frac{1}{n}\sum_{i=1}^{n}I(\hat{f}(X_i)Y_i\le\gamma)-Q(f)]}\\
\le& E_{S_n,S'_n}\sup_{f\in\mathcal{H}} e^{\lambda\sum_{i=1}^{n}[(1-\alpha)I(f(X'_i)Y'_i\le 0)-I(f(X_i)Y_i\le\gamma)]-\lambda nQ(f)}\\
\le& N_\infty(\mathcal{H},\gamma/2|2n)\sup_{f\in\mathcal{H}} E_{S_n,S'_n}e^{\lambda\sum_{i=1}^{n}[(1-\alpha)I(f(X'_i)Y'_i\le\gamma/2)-I(f(X_i)Y_i\le\gamma/2)]-\lambda nQ(f)}\\
\le& N_\infty(\mathcal{H},\gamma/2|2n)\sup_{f\in\mathcal{H}} e^{\lambda n(\alpha(Var(I(f(X)Y\le\gamma/2))-EI(f(X)Y\le\gamma/2))-Q(f))}\\
\le& \eta.
\end{aligned}
$$

Margin Bounds for empirical risk minimization

Given any fixed λ and γ, with probability $1-\eta$, we have the following bound for all $f \in \mathcal{H}$:

$$\mathbf{E}_{X,Y} I(f(X)Y \leq 0) \leq \frac{1}{(1-\alpha)n} \sum_{i=1}^{n} I(f(X_i)Y_i \leq \gamma) + \frac{\ln(N_\infty(\mathcal{H}, \gamma/2|2n)/\eta)}{\lambda(1-\alpha)n},$$

where $\alpha = 2(e^\lambda - \lambda - 1)/\lambda$.

- Problem: margin needs to be known a priori
- Solution: sample dependent bound (adaptive to margin)

Data dependent bound

Let $j = 1, \ldots$ and a sequence of $\gamma_1 \geq \gamma_2 \cdots$ (for example, $\gamma_j = 1/2^j$), then the following margin bound holds with probability $1 - \eta_j$ where $\eta_j = \eta/j(j+1)$:

$$\mathbf{E}_{X,Y} I(f(X)Y \leq 0) \leq \frac{1}{(1-\alpha)n} \sum_{i=1}^{n} I(f(X_i)Y_i \leq \gamma_j) + \frac{\ln(N_\infty(\mathcal{H}, \gamma_j/2|2n)/\eta_j)}{\lambda(1-\alpha)n},$$

where $\alpha = 2(\exp(\lambda) - \lambda - 1)/\lambda$.

Take union bound over j. We have a unified statement that holds over all j. Given any fixed λ with probability $1 - \eta$, we have the following bound for all $f \in \mathcal{H}$ and all $j = 1, \ldots$:

$$\mathbf{E}_{X,Y} I(f(X)Y \leq 0) \leq \frac{1}{(1-\alpha)n} \sum_{i=1}^{n} I(f(X_i)Y_i \leq \gamma_j) + \frac{\ln(N_\infty(\mathcal{H}, \gamma_j/2|2n)j(j+1)/\eta)}{\lambda(1-\alpha)n}$$

We can take $\gamma_j = A/2^{j-1}$.

Now given $\gamma \in (0, A]$, we take $j = \lfloor \log_2(A/\gamma) \rfloor + 1$, then $\gamma_j \in [\gamma/2, \gamma]$. The above inequality holds for γ_j implies that with probability $1 - \eta$, the following holds for all $\gamma \in (0, A]$ and $f \in \mathcal{H}$:

$$\begin{aligned}\mathbf{E}_{X,Y} I(f(X)Y \leq 0) \leq & \frac{1}{(1-\alpha)n} \sum_{i=1}^{n} I(f(X_i)Y_i \leq \gamma) \\ & + \frac{\ln(N(\mathcal{H}, \gamma/2|2n)/\eta) + 2\ln(\lfloor \log_2(A/\gamma) \rfloor + 2)}{\lambda(1-\alpha)n}.\end{aligned}$$

We may adapt to λ in a similar matter by taking $\lambda_j = j/n$.

Data-dependent Rademacher Complexity Bound

If $\phi \in [0,1]$, then McDiarmid implies that Rademacher complexity concentrates: with probability $1-\eta$

$$E_{S_n} R(\phi(\mathcal{H})|S_n) \leq R(\phi(\mathcal{H})|S_n) + \sqrt{\ln(1/\eta)/(2n)}$$

Combine with Rademacher complexity bound, we obtain the following data dependent learning bound: with probability $1-\eta$:

$$E_{X,Y}\phi(\hat{f}(X),Y) \leq \frac{1}{n}\sum_{i=1}^{n}\phi(\hat{f}(X_i),Y_i) + 2R(\phi(\mathcal{H})|S_n) + 3\sqrt{\ln(2/\eta)/(2n)}.$$