Overview of Learning Theory

Tong Zhang

Rutgers University

Standard Model for Supervised Learning

- Data (X, Y) are randomly drawn from an underlying distribution D.
  - Binary classification: $Y \in \{\pm 1\}$
- Assume training data are iid samples from D:

$$S_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$$

- Want to construct prediction function f from training data to minimize future loss over D

$$\mathbf{E}_{(X,Y)\sim D} I(f(X) \neq Y)$$

Learning Algorithm

- Learning algorithm $\mathcal{A}$
  - learn prediction rule $\hat{f} = \mathcal{A}(S_n)$ from training data $S_n = \{(X_i, Y_i)\}_{i=1,\ldots,n}$.
- Training error

$$\text{TRAINING ERROR}(f) = \frac{1}{n}\sum_{i=1}^{n} I(f(X_i) \neq Y_i)$$

- Evaluate performance of a learning algorithm on test data.

$$\text{TEST ERROR}(f) = \mathbf{E}_{(X,Y)\sim D} I(f(X_i) \neq Y_i)$$

Generalization Analysis

- How good is a learning algorithm and how to theoretically justify it?

Generalization Analysis

- How good is a learning algorithm and how to theoretically justify it?
- Given a learning algorithm, we want to know
  - how good is the learning algorithm compared to the best possible prediction rule in a class?
  - oracle inequality

Generalization Analysis

- How good is a learning algorithm and how to theoretically justify it?
- Given a learning algorithm, we want to know
  - how good is the learning algorithm compared to the best possible prediction rule in a class?
  - oracle inequality
- How to estimate test error from training error
  - we observe training error.
  - we want to minimize test error.
  - the goal is to estimate the difference of test error and training error

Generalization Analysis

- How good is a learning algorithm and how to theoretically justify it?
- Given a learning algorithm, we want to know
  - how good is the learning algorithm compared to the best possible prediction rule in a class?
  - oracle inequality
- How to estimate test error from training error
  - we observe training error.
  - we want to minimize test error.
  - the goal is to estimate the difference of test error and training error
- Other properties of learning algorithm
- Key concept: uniform convergence
  - we will discuss technical tools to prove uniform convergence

Example: Linear Classification I

- Linear classification rule: weight vector $w \in R^d$ and predict label as

$$f(x) = \mathrm{sgn}(w^T x)$$

- Classification error

$$I(f(x) \neq y) = I(f(x)y \leq 0) = I(w^T xy \leq 0).$$

- Training error:

$$\text{TRAINING ERROR}(w) = n^{-1} \sum_{i=1}^{n} I(w^T X_i Y_i \leq 0).$$

- Test error:

$$\text{TEST ERROR}(w) = \mathbf{E}_{(X,Y)} I(w^T XY \leq 0).$$

Example: Linear Classification II

- Learning algorithm: minimize training error

$$\hat{w} = \arg\min_{w} \sum_{i=1}^{n} I(w^T X_i Y_i \leq 0).$$

- What we are interested in: test error
- Our question: how good is this algorithm?

How Close is Test Error to Training Error

- We observe the training error of $\hat{w}$, how to estimate its test error?
- Statement: with probability $1 - \eta$ over randomly drawn training data, we have

  $$\text{TEST ERROR}(\hat{w}) \leq \text{TRAINING ERROR}(\hat{w}) + C\sqrt{(d + \ln(1/\eta))/n},$$

  wehre C is a constant.
  - why probability: due to the randomness of training data, there are chance that the trained classifier may not be good
  - it may not be representative of test data.

How Close is Test Error to Training Error

- We observe the training error of $\hat{w}$, how to estimate its test error?
- Statement: with probability $1 - \eta$ over randomly drawn training data, we have

$$\text{TEST ERROR}(\hat{w}) \leq \text{TRAINING ERROR}(\hat{w}) + C\sqrt{(d + \ln(1/\eta))/n},$$

  wehre C is a constant.
  - why probability: due to the randomness of training data, there are chance that the trained classifier may not be good
  - it may not be representative of test data.
- Proof is via uniform convergence.

Uniform Convergence

Uniform Convergence Statement:

- with probability $1 - \eta$ over training data, we have for all classifiers $\tilde{w}$ that may depend on training data

$$|\text{TEST ERROR}(\tilde{w}) - \text{TRAINING ERROR}(\tilde{w})| \leq C\sqrt{(d + \ln(1/\eta))/n},$$

Implication of Uniform Convergence:
since it applies to all training data dependent $\tilde{w}$, we may take $\tilde{w} = \hat{w}$ and estimate test error of $\hat{w}$ as

$$\text{TEST ERROR}(\hat{w}) \leq \text{TRAINING ERROR}(\hat{w}) + C\sqrt{d\ln(1/\eta)/n}.$$

How Good is Error Minimization Learner?

- How good is $\hat{w}$ compared to best linear classifier w^*, defined as

$$w^* = \arg\min_w \text{TEST ERROR}(w).$$

- Statement: with probability $1 - \eta$ over randomly drawn training data, we have

$$\text{TEST ERROR}(\hat{w}) \leq \text{TEST ERROR}(w^*) + 2C\sqrt{(d + \ln(1/\eta))/n}.$$

Proof based on Uniform Convergence

$$\text{TRAINING ERROR}(\hat{w}) \leq \text{TRAINING ERROR}(w^*)$$

From uniform convergence, we know

$$\text{TEST ERROR}(\hat{w}) \leq \text{TRAINING ERROR}(\hat{w}) + C\sqrt{(d + \ln(1/\eta))/n}$$

and

$$\text{TRAINING ERROR}(w^*) \leq \text{TEST ERROR}(w^*) + C\sqrt{(d + \ln(1/\eta))/n}$$

Add the above inequalities we obtain the desired bound.

Some Common Techniques for Uniform Convergence

- Exponential Probability Inequality (Chernoff Bound)
- VC dimension
- Covering numbers
- Rademacher Complexity

Exponential Inequality (Chernoff Bound)

- Let $X \in [0, 1]$ be a random variable, with mean μ.
- Let $X_1, \ldots, X_n$ are iid samples from the same distribution, and let

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

  be the empirical mean.
- We want to know: how good is $\bar{X}_n$ as an estimator of μ?
- Exponential tail inequality: $|\bar{X}_n - \mu| > \epsilon$ is exponentially small.

Exponential Inequality (Chernoff Bound)

- Let $X \in [0, 1]$ be a random variable, with mean μ.
- Let $X_1, \ldots, X_n$ are iid samples from the same distribution, and let

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

  be the empirical mean.
- We want to know: how good is $\bar{X}_n$ as an estimator of μ?
- Exponential tail inequality: $|\bar{X}_n - \mu| > \epsilon$ is exponentially small.
- Chernoff bound (Hoeffding's inequality):

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq 2e^{-2n\epsilon^2}.$$

  Alternatively, we have with probability $1 - \eta$:

$$|\bar{X}_n - \mu| \leq \sqrt{\ln(2/\eta)/2n}$$

  by setting $\eta = 2e^{-2n\epsilon^2}$ which implies $\epsilon = \sqrt{\ln(2/\eta)/2n}$.

Consequence of Exponential Inequality

- Test versus training errors for single classifier:
  - $X \in \{0,1\}$: classification error of a single classifier f
  - use Chernoff bound to estimate test performance from training error:

$$|\bar{X}_n - \mu| \leq \sqrt{\ln(2/\eta)/2n}.$$

- Test versus training error for M classifiers $f_1, \ldots, f_M$
  - training error $\text{TRAINING ERROR}(f_j)$
  - test error $\text{TEST ERROR}(f_j)$.
  - uniform convergence statement: with probability $1-\eta$, for all j

$$|\text{TRAINING ERROR}(f_j) - \text{TEST ERROR}(f_j)| \leq \sqrt{\ln(2M/\eta)/2n}.$$

Proof via Union Bound

- For each j, the probability of the following fails

$$|\text{TRAINING ERROR}(f_j) - \text{TEST ERROR}(f_j)| \leq \sqrt{\ln(2M/\eta)/2n}.$$

  is no more than η/M.
- The probability the above inequality fails for at least one j is no more than $M * (\eta/M) = \eta$.

Consequence: this means that the following inequality holds for all j with probability at least $1 - \eta$:

$$|\text{TRAINING ERROR}(f_j) - \text{TEST ERROR}(f_j)| \leq \sqrt{\ln(2M/\eta)/2n}.$$

or equivalently with probability $1 - \eta$:

$$\sup_j |\text{TRAINING ERROR}(f_j) - \text{TEST ERROR}(f_j)| \leq \sqrt{\ln(2M/\eta)/2n}.$$

Consequences: measure overfitting

Testing M classifiers on training and pick the best

$$\hat{j} = \arg\min_j \text{TRAINING ERROR}(f_j).$$

Uniform convergence: with probability $1 - \eta$,

$$\sup_j |\text{TRAINING ERROR}(f_j) - \text{TEST ERROR}(f_j)| \leq \sqrt{\ln(2M/\eta)/2n}.$$

Estimating test error from training error:

$$\text{TEST ERROR}(f_{\hat{j}}) \leq \text{TRAINING ERROR}(f_{\hat{j}}) + \sqrt{\ln(2M/\eta)/2n}.$$

- degree of overfitting: $\sqrt{\ln(2M/\eta)/2n}$.
- when n is large, overfitting is small when $\ln M/n = o(1)$.
- rule of thumb: one can tolerate exponential in n many models.

Empirical Processes

- Empirical process is an abstraction of estimating test error from training errors for multiple classifiers
- Problem set up
  - observations training data S_n
  - classifier: $f_\theta(x)$, parameterized by $\theta \in \Theta$
  - $\hat{\theta}$: estimated from training data, what is its test error?
  - Uniform convergence:

$$\sup_{\theta \in \Theta} \left| \text{TRAINING ERROR}(f_\theta) - \text{TEST ERROR}(f_\theta) \right|.$$

- If Θ is finite, we know how to obtain uniform convergence bound from Chernoff bound.

Empirical Processes

- Empirical process is an abstraction of estimating test error from training errors for multiple classifiers
- Problem set up
  - observations training data S_n
  - classifier: $f_\theta(x)$, parameterized by $\theta \in \Theta$
  - $\hat{\theta}$: estimated from training data, what is its test error?
  - Uniform convergence:

$$\sup_{\theta \in \Theta} \left| \text{TRAINING ERROR}(f_\theta) - \text{TEST ERROR}(f_\theta) \right|.$$

- If Θ is finite, we know how to obtain uniform convergence bound from Chernoff bound.
- What if Θ is infinity?

Infinite Θ: Covering Number

Many versions of covering numbers: we consider one definition

- Given classifiers $f_\theta(x)$ with $\theta \in \Theta$ that takes $\{0, 1\}$ values, we may define its empirical L_∞ covering number. Let $\mathcal{H} = \{f_\theta(x) : \theta \in \Theta\}$, and define

$$L_\infty(\mathcal{H}|S_n) = |\{[I(f_\theta(X_1) \neq Y_1), \ldots, I(f(X_n) \neq Y_n)]\} : \theta \in \Theta\}|.$$

  empirical covering number is the number of the functions f_θ can attain at finite number of training points.

How to estimate covering number?

- partial answer: VC-dimension $VC(\mathcal{H})$ for binary-valued functions
- there are other methods

VC Dimension

Definition (Shattering)

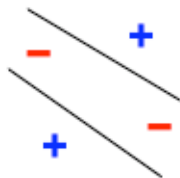
A function class $\mathcal{H}$ is said to shatter a set of data points $(X_1, X_2, \ldots, X_n)$ if, for every assignment of labels to those points $(Y_1, \ldots, Y_n)$, there exists a function $f \in \mathcal{H}$ such that f makes no errors when evaluating that set of data points: $f(X_i) = Y_i$ for all i.

- any possible labeling can be explained
- complete overfitting

VC dimension $VC(\mathcal{H})$: the maximum n such that there exist data points of cardinality n that can be shattered.

Example: linear separator in 2d

- In 2d:
  - data $x \in R^2$
  - $\mathcal{H} = \{\text{sign}(w^T x + b) : w \in R^2, b \in R\}$
- 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 general: d dimensional linear classifier has VC dimension $d + 1$

VC dimension and covering number

- Covering number bound: Sauer's Lemma ($n \geq d$)

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

  consequence:
  - uniform convergence: similar to M classifiers with $M = (en/d)^d$
  - can be shown using a technique called symmetrization
- Uniform convergence: with probability $1 - \eta$,

$$\sup_\theta |\text{TRAINING ERROR}(f_\theta) - \text{TEST ERROR}(f_\theta)|$$
$$\leq O(\sqrt{\ln(M/\eta)/n}) = O\left(\sqrt{\frac{d\ln(en/d) + \ln(1/\eta)}{n}}\right).$$

  more refined analysis can remove $\ln n$.

Rademacher Complexity

We define Rademacher Complexity of $\mathcal{H}$ as

$$R(\mathcal{H}) = E_{S_n} \, E_\sigma \sup_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i I(f(X_i) Y_i \le 0),$$

where $\sigma = \{\sigma_1, \ldots, \sigma_n\}$: each σ_i randomly takes values in $\{\pm 1\}$.

- Covering number bounds imply Rademacher complexity bounds
- There are other ways to bound Rademacher complexity.

Uniform convergence bound:

$$\begin{aligned} &\sup_{f \in \mathcal{H}} \left| \text{TEST ERROR}(f) - \text{TRAINING ERROR}(f) \right| \\ \le & 2R(\mathcal{H}) + \sqrt{\frac{\ln(2/\eta)}{2n}}. \end{aligned}$$

This is called concentration inequality (McDiarmid Inequality).

VC Dimension and Rademacher Complexity

- Rademacher complexity

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

- Uniform convergence bound:

$$\begin{aligned}&\sup_{f\in\mathcal{H}} |\text{TEST ERROR}(f) - \text{TRAINING ERROR}(f)|\\ \leq& 2C\sqrt{\mathrm{vc}(\mathcal{H})/n} + \sqrt{\frac{\ln(2/\eta)}{2n}}.\end{aligned}$$

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:

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

We can estimate $Q(\hat{f})$ using L_1 norm as:

$$Q_\gamma(\hat{f}) = \sqrt{\frac{\|\hat{w}\|_1^2 \sup_i \|X_i\|_\infty^2}{\gamma^2 n}}$$

The quantity depends on margin instead of dimension.

Summary

- The main goal of learning theory is to understand the effectiveness of learning algorithms
- The main results are oracle inequalities or inequalities relating test error and training error (generalization error bound)
- The main techniques are uniform convergence and empirical processes
- The main techniques for uniform convergence are exponential inequality, covering numbers, and Rademacher complexity
- Results are scattered in the literature. Further readings:
  - `http://www.cs.berkeley.edu/~bartlett/courses/281b-sp08`
  - `http://www.cs.huji.ac.il/~shais/AdvancedML.html`
  - `http://www.cc.gatech.edu/~ninamf/ML11/index.html`