PAC Learning

Grigoris Velegkas

ECE, NTUA

2/11/18

So far

- Domain Set X
- Label Set Y
- Training Data $S = ((x_1, y_1), ..., (x_m, y_m))$
- Learner's Output $h : X \to Y$ (predictor)
- Data Generation: D over X generates x_i, then $f\colon X \to Y$ labels it (we'll relax it later)
- Measure of success: $L_{D,f}(h) = P_{x \sim D}[h(x) \neq f(x)] = D(\{x : h(x) \neq f(x)\})$
- ERM: output h that minimizes $L_{D,f}$ over training data
- Overfitting: select H before seeing S
- Finite H (realizability + i.i.d.): $m \geq \frac{\log(|H|/\delta)}{\epsilon} \implies L_{D,f}(h_S) \leq \epsilon$ with probability at least $1 - \delta$

PAC

H is PAC learnable if $\exists m_H : (0,1)^2 \to \mathbb{N}$ and an algorithm A with the following property
$\forall \epsilon, \delta \in (0,1), \forall D$ over $X, \forall f : X \to \{0,1\}$ if the realizable assumption holds then when we run A on $m \geq m_H(\epsilon, \delta)$ i.i.d. samples generated by D and labeled by f, A returns $h \in H$ s.t. $P[L_{D,f}(h) \leq \epsilon] \geq 1 - \delta$

Remarks

- Two approximation parameters: confidence δ ("probably"), accuracy ϵ ("approximately")

Remarks

- Two approximation parameters: confidence δ ("probably"), accuracy ϵ ("approximately")
- Inevitable under that data access model

Remarks

- Two approximation parameters: confidence δ ("probably"), accuracy ϵ ("approximately")
- Inevitable under that data access model
- Small chance that sample is noninformative (e.g. same point) $\rightarrow \delta$

Remarks

- Two approximation parameters: confidence δ ("probably"), accuracy ϵ ("approximately")
- Inevitable under that data access model
- Small chance that sample is noninformative (e.g. same point) $\rightarrow \delta$
- Since sample is finite it might fail to reflect details of $D \rightarrow \epsilon$

Remarks

- Two approximation parameters: confidence δ (”probably”), accuracy ϵ (”approximately”)
- Inevitable under that data access model
- Small chance that sample is noninformative (e.g. same point) $\rightarrow \delta$
- Since sample is finite it might fail to reflect details of $D \rightarrow \epsilon$

Sample Complexity

$m_H : (0,1)^2 \to \mathbb{N}$ is the *sample complexity* of learning H

- Depends on δ, ϵ

Sample Complexity

$m_H : (0,1)^2 \to \mathbb{N}$ is the *sample complexity* of learning H

- Depends on δ, ϵ
- We take the "minimal function"

Sample Complexity

$m_H : (0,1)^2 \to \mathbb{N}$ is the *sample complexity* of learning H

- Depends on δ, ϵ
- We take the "minimal function"
- Finite H

$$m_H(\epsilon, \delta) \leq \left\lceil \frac{\log(|H|/\delta)}{\epsilon} \right\rceil$$

Generalizing the Model

- We have assumed that labels are provided by (a given) f, too strong
  We now assume that D is a distribution over $X \times Y$
  Two components D_x over unlabeled domain points, $D((x, y)|x)$ over the labels *given* a point
  We do not know anything about D!

Generalizing the Model

- We have assumed that labels are provided by (a given) f, too strong
  We now assume that D is a distribution over $X \times Y$
  Two components D_x over unlabeled domain points, $D((x, y)|x)$ over the labels *given* a point
  We do not know anything about D!
- We are interested in tasks beyond binary classification, Y can be a real-valued set or a finite set

Generalizing the Model

- True error (risk)

$$L_D(h) = \mathbb{P}_{(x,y)\sim D}[h(x) \neq y] = D(\{(x, y) : h(x) \neq y\})$$

Generalizing the Model

- True error (risk)

$$L_D(h) = \mathbb{P}_{(x,y)\sim D}[h(x) \neq y] = D(\{(x,y) : h(x) \neq y\})$$

- Empirical risk

$$L_S(h) = \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m}$$

Goal

- Ideally, we would like to predict an h that probably approximately minimizes the true error

Goal

- Ideally, we would like to predict an h that probably approximately minimizes the true error
- Bayes Optimal Predictor: Given a D over $X \times \{0, 1\}$, the best label predicting function is

$$f_D(x) = \begin{cases} 1 & \mathbb{P}[y = 1|x] \geq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

Goal

- Ideally, we would like to predict an h that probably approximately minimizes the true error
- Bayes Optimal Predictor: Given a D over $X \times \{0,1\}$, the best label predicting function is

$$f_D(x) = \begin{cases} 1 & \mathbb{P}[y=1|x] \geq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

We do not know D! If we make no assumptions about D we cannot find a predictor which is as good as that

Agnostic PAC

H is agnostic PAC learnable if $\exists m_H : (0,1)^2 \to \mathbb{N}$ and an algorithm A with the following property
$\forall \epsilon, \delta \in (0,1), \forall D$ over $X \times Y$ when we run A on $m \geq m_H(\epsilon, \delta)$ i.i.d. samples generated by D, A returns $h \in H$ s.t.
$P[L_D(h) \leq \min_{h' \in H} L_D(h') + \epsilon] \geq 1 - \delta$

Scope of Learning Problems

- Multiclass Classification: X represents the features of the domain space, Y the different labels

Scope of Learning Problems

- Multiclass Classification: X represents the features of the domain space, Y the different labels
- Regression: We want to find a simple pattern in the data (e.g. linear function) to predict a value. Different measure of success

$$L_D(h) = \mathbb{E}_{(x,y)\sim D}(h(x) - y)^2$$

Scope of Learning Problems

- Multiclass Classification: X represents the features of the domain space, Y the different labels
- Regression: We want to find a simple pattern in the data (e.g. linear function) to predict a value. Different measure of success

$$L_D(h) = \mathbb{E}_{(x,y)\sim D}(h(x) - y)^2$$

- Different tasks require different loss functions

Generalized Loss Functions

- $l: H \times Z \to \mathbb{R}_+$, in prediction problems $Z = X \times Y$

Generalized Loss Functions

- $l : H \times Z \to \mathbb{R}_+$, in prediction problems $Z = X \times Y$
- $L_D(h) = \mathbb{E}_{z \sim D}[l(h, z)]$

Generalized Loss Functions

- $l : H \times Z \to \mathbb{R}_+$, in prediction problems $Z = X \times Y$
- $L_D(h) = \mathbb{E}_{z \sim D}[l(h, z)]$
- $L_S(h) = \frac{1}{m} \sum_{i=1}^{m} l(h, z_i)$

Generalized Loss Functions

- $l : H \times Z \to \mathbb{R}_+$, in prediction problems $Z = X \times Y$
- $L_D(h) = \mathbb{E}_{z \sim D}[l(h, z)]$
- $L_S(h) = \frac{1}{m} \sum_{i=1}^{m} l(h, z_i)$
- $0 - 1$ loss: z ranges over $X \times Y$

$$l_{0-1}(h, (x, y)) = \begin{cases} 0 & h(x) = y \\ 1 & \text{otherwise} \end{cases}$$

Generalized Loss Functions

- $l : H \times Z \to \mathbb{R}_+$, in prediction problems $Z = X \times Y$
- $L_D(h) = \mathbb{E}_{z \sim D}[l(h, z)]$
- $L_S(h) = \frac{1}{m} \sum_{i=1}^{m} l(h, z_i)$
- $0 - 1$ loss: z ranges over $X \times Y$

$$l_{0-1}(h, (x, y)) = \begin{cases} 0 & h(x) = y \\ 1 & \text{otherwise} \end{cases}$$

- Square loss: z ranges over $X \times Y$

$$l_{sq}(h, (x, y)) = (h(x) - y)^2$$

Agnostic PAC for general loss functions

H is agnostic PAC learnable with respect to a set Z and a loss function $l : H \times Z \to \mathbb{R}_+$, if $\exists m_H : (0,1)^2 \to \mathbb{N}$ and an algorithm A with the following property
$\forall \epsilon, \delta \in (0,1), \forall D$ over $X \times Y$ when we run A on $m \geq m_H(\epsilon, \delta)$ i.i.d. samples generated by D, A returns $h \in H$ s.t.
$P[L_D(h) \leq \min_{h' \in H} L_D(h') + \epsilon] \geq 1 - \delta$, where $L_D(h) = \mathbb{E}_{z \sim D}[l(h, z)]$

Uniform Convergence

Idea: Uniformly over all hypotheses in H we want the empirical risk to be close to the true risk

Uniform Convergence

Idea: Uniformly over all hypotheses in H we want the empirical risk to be close to the true risk

• A training set S is called ϵ-representative if
$\forall h \in H, |L_S(h) - L_D(h)| \leq \epsilon$

Uniform Convergence

Idea: Uniformly over all hypotheses in H we want the empirical risk to be close to the true risk

- A training set S is called ϵ-representative if $\forall h \in H, |L_S(h) - L_D(h)| \leq \epsilon$
- Lemma: If S is $\epsilon/2$-representative then any output of $ERM_H(S)$ satisfies: $L_D(h_S) \leq \min_{h' \in H} L_D(h') + \epsilon$

Uniform Convergence

- Idea: In order to show that a class H is agnostic PAC learnable it suffices to show that with probability $1 - \delta$ the training set will be ϵ-representative

Uniform Convergence

- Idea: In order to show that a class H is agnostic PAC learnable it suffices to show that with probability $1-\delta$ the training set will be ϵ-representative
- Uniform Convergence: We say that a hypothesis class H has the uniform convergence property if $\exists m_H^{UC}(0,1)^2 \rightarrow \mathbb{N}$ s.t. $\forall \epsilon, \delta \in (0,1), \forall D$ if S is a sample of $m \geq m_H^{UC}(\epsilon, \delta)$ i.i.d. points drawn according to D, then with probability at least $1-\delta, S$ is $\epsilon-$representative

Uniform Convergence

- Idea: In order to show that a class H is agnostic PAC learnable it suffices to show that with probability $1-\delta$ the training set will be ϵ-representative
- Uniform Convergence: We say that a hypothesis class H has the uniform convergence property if $\exists m_H^{UC}(0,1)^2 \to \mathbb{N}$ s.t. $\forall \epsilon, \delta \in (0,1), \forall D$ if S is a sample of $m \geq m_H^{UC}(\epsilon,\delta)$ i.i.d. points drawn according to D, then with probability at least $1-\delta, S$ is $\epsilon-$representative
- Corollary: If H has the uniform convergence property then it is agnostically PAC learnable with sample complexity $m_H(\epsilon,\delta) \leq m_H^{UC}(\epsilon/2,\delta)$. Furthermore, ther ERM paradigm is a successful agnostic PAC learner.

Finite Classes are APAC learnable

- We have to show that
  $D^m(\{S : \forall h \in H, |L_S(h) - L_D(h)| \leq \epsilon\}) \geq 1 - \delta$

Finite Classes are APAC learnable

- We have to show that $D^m(\{S : \forall h \in H, |L_S(h) - L_D(h)| \leq \epsilon\}) \geq 1 - \delta$
- Equivalently $D^m(\{S : \exists h \in H, |L_S(h) - L_D(h)| > \epsilon\}) < \delta$

Finite Classes are APAC learnable

- We have to show that $D^m(\{S : \forall h \in H, |L_S(h) - L_D(h)| \leq \epsilon\}) \geq 1 - \delta$
- Equivalently $D^m(\{S : \exists h \in H, |L_S(h) - L_D(h)| > \epsilon\}) < \delta$
- $\{S : \exists h \in H, |L_S(h) - L_D(h)| > \epsilon\} = \cup_{h \in H}\{S : |L_S(h) - L_D(h)| > \epsilon\}$

Finite Classes are APAC learnable

- We have to show that $D^m(\{S : \forall h \in H, |L_S(h) - L_D(h)| \leq \epsilon\}) \geq 1 - \delta$
- Equivalently $D^m(\{S : \exists h \in H, |L_S(h) - L_D(h)| > \epsilon\}) < \delta$
- $\{S : \exists h \in H, |L_S(h) - L_D(h)| > \epsilon\} = \cup_{h \in H}\{S : |L_S(h) - L_D(h)| > \epsilon\}$
- Union bound: $D^m(\{S : \exists h \in H, |L_S(h) - L_D(h)| > \epsilon\}) \leq \sum_{h \in H} D^m(\{S : |L_S(h) - L_D(h)| > \epsilon\})$

Finite Classes are APAC learnable

- We have to show that $D^m(\{S : \forall h \in H, |L_S(h) - L_D(h)| \leq \epsilon\}) \geq 1 - \delta$
- Equivalently $D^m(\{S : \exists h \in H, |L_S(h) - L_D(h)| > \epsilon\}) < \delta$
- $\{S : \exists h \in H, |L_S(h) - L_D(h)| > \epsilon\} = \cup_{h \in H}\{S : |L_S(h) - L_D(h)| > \epsilon\}$
- Union bound: $D^m(\{S : \exists h \in H, |L_S(h) - L_D(h)| > \epsilon\}) \leq \sum_{h \in H} D^m(\{S : |L_S(h) - L_D(h)| > \epsilon\})$
- Idea: We will show that each summand is small

Finite Classes are APAC learnable

- Recall that $L_D(h) = \mathbb{E}_{z \sim D}[l(h, z)], L_S(h) = \frac{1}{m} \sum_{i=1}^{m} l(h, z_i)$

Finite Classes are APAC learnable

- Recall that $L_D(h) = \mathbb{E}_{z \sim D}[l(h, z)], L_S(h) = \frac{1}{m} \sum_{i=1}^{m} l(h, z_i)$
- Since each z_i is sampled i.i.d. from D we have that $\mathbb{E}_{z_i \sim D}[l(h, z_i)] = L_D(h), \forall i \in [m]$

Finite Classes are APAC learnable

- Recall that $L_D(h) = \mathbb{E}_{z \sim D}[l(h, z)], L_S(h) = \frac{1}{m}\sum_{i=1}^{m} l(h, z_i)$
- Since each z_i is sampled i.i.d. from D we have that $\mathbb{E}_{z_i \sim D}[l(h, z_i)] = L_D(h), \forall i \in [m]$
- By linearity of expectation $L_D(h) = \mathbb{E}_{S \sim D^m}[L_S(h)]$, hence $|L_S(h) - L_D(h)|$ is the deviation of $L_S(h)$ from its expectation

Finite Classes are APAC learnable

- Hoeffding's Inequality: $\theta_1, ..., \theta_m$ i.i.d., $\mathbb{E}[\theta_i] = \mu, \mathbb{P}[a \leq \theta_i \leq b] = 1$
  $\mathbb{P}[|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu| > \epsilon] \leq 2\exp(-2m\epsilon^2/(b-a)^2)$

Finite Classes are APAC learnable

- Hoeffding's Inequality: $\theta_1, ..., \theta_m$ i.i.d., $\mathbb{E}[\theta_i] = \mu, \mathbb{P}[a \leq \theta_i \leq b] = 1$
  $\mathbb{P}[|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu| > \epsilon] \leq 2\exp(-2m\epsilon^2/(b-a)^2)$
- $\theta_i = l(h, z_i)$, since h is fixed and $z_1, ..., z_m$ are i.i.d. it follows that $\theta_1, ..., \theta_m$ are also i.i.d.

Finite Classes are APAC learnable

- Hoeffding's Inequality: $\theta_1, ..., \theta_m$ i.i.d., $\mathbb{E}[\theta_i] = \mu, \mathbb{P}[a \leq \theta_i \leq b] = 1$
  $\mathbb{P}[|\frac{1}{m}\sum_{i=1}^{m} \theta_i - \mu| > \epsilon] \leq 2\exp(-2m\epsilon^2/(b-a)^2)$
- $\theta_i = l(h, z_i)$, since h is fixed and $z_1, ..., z_m$ are i.i.d. it follows that $\theta_1, ..., \theta_m$ are also i.i.d.
- $L_S(h) = \frac{1}{m}\sum_{i=1}^{m} \theta_i, L_D(h) = \mu$, we assume that the range of l is $[0, 1]$, thus $\theta_i \in [0, 1]$

Finite Classes are APAC learnable

- Hoeffding's Inequality: $\theta_1, ..., \theta_m$ i.i.d., $\mathbb{E}[\theta_i] = \mu, \mathbb{P}[a \leq \theta_i \leq b] = 1$
  $\mathbb{P}[|\frac{1}{m}\sum_{i=1}^m \theta_i - \mu| > \epsilon] \leq 2\exp(-2m\epsilon^2/(b-a)^2)$
- $\theta_i = l(h, z_i)$, since h is fixed and $z_1, ..., z_m$ are i.i.d. it follows that $\theta_1, ..., \theta_m$ are also i.i.d.
- $L_S(h) = \frac{1}{m}\sum_{i=1}^m \theta_i, L_D(h) = \mu$, we assume that the range of l is $[0, 1]$, thus $\theta_i \in [0, 1]$
- $D^m(\{S : |L_S(h) - L_D(h)| > \epsilon\}) = \mathbb{P}[|\frac{1}{m}\sum_{i=1}^m \theta_i - \mu| > \epsilon] \leq 2\exp(-2m\epsilon^2)$

Finite Classes are APAC learnable

- Hoeffding's Inequality: $\theta_1, ..., \theta_m$ i.i.d., $\mathbb{E}[\theta_i] = \mu, \mathbb{P}[a \leq \theta_i \leq b] = 1$ $\mathbb{P}[|\frac{1}{m}\sum_{i=1}^m \theta_i - \mu| > \epsilon] \leq 2\exp(-2m\epsilon^2/(b-a)^2)$
- $\theta_i = l(h, z_i)$, since h is fixed and $z_1, ..., z_m$ are i.i.d. it follows that $\theta_1, ..., \theta_m$ are also i.i.d.
- $L_S(h) = \frac{1}{m}\sum_{i=1}^m \theta_i, L_D(h) = \mu$, we assume that the range of l is $[0, 1]$, thus $\theta_i \in [0, 1]$
- $D^m(\{S : |L_S(h) - L_D(h)| > \epsilon\}) = \mathbb{P}[|\frac{1}{m}\sum_{i=1}^m \theta_i - \mu| > \epsilon] \leq 2\exp(-2m\epsilon^2)$
- $D^m(\{S : \exists h \in H, |L_S(h) - L_D(h)| > \epsilon\}) \leq \sum_{h \in H} 2\exp(-2m\epsilon^2) = 2|H|\exp(-2m\epsilon^2)$

Finite Classes are APAC learnable

- Hoeffding's Inequality: $\theta_1, ..., \theta_m$ i.i.d., $\mathbb{E}[\theta_i] = \mu, \mathbb{P}[a \leq \theta_i \leq b] = 1$
  $\mathbb{P}[|\frac{1}{m}\sum_{i=1}^{m} \theta_i - \mu| > \epsilon] \leq 2\exp(-2m\epsilon^2/(b-a)^2)$
- $\theta_i = l(h, z_i)$, since h is fixed and $z_1, ..., z_m$ are i.i.d. it follows that $\theta_1, ..., \theta_m$ are also i.i.d.
- $L_S(h) = \frac{1}{m}\sum_{i=1}^{m} \theta_i, L_D(h) = \mu$, we assume that the range of l is $[0, 1]$, thus $\theta_i \in [0, 1]$
- $D^m(\{S : |L_S(h) - L_D(h)| > \epsilon\}) = \mathbb{P}[|\frac{1}{m}\sum_{i=1}^{m} \theta_i - \mu| > \epsilon] \leq 2\exp(-2m\epsilon^2)$
- $D^m(\{S : \exists h \in H, |L_S(h) - L_D(h)| > \epsilon\}) \leq \sum_{h \in H} 2\exp(-2m\epsilon^2) = 2|H|\exp(-2m\epsilon^2)$
- $m \geq \frac{\log(2|H|/\delta)}{2\epsilon^2} \implies D^m(\{S : \exists h \in H, |L_S(h) - L_D(h)| > \epsilon\}) \leq \delta$