1- Learning Theory Recap

Melih Kandemir

University of Southern Denmark
Department of Mathematics and Computer Science (IMADA)
kandemir@imada.sdu.dk

Fall 2022

Machine learning at large

Definition: *"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured in P, improves with experience E".* [Mitchell, 1997]

Purpose: Designing algorithms to solve T with maximum P and minimum

- time complexity
- space complexity
- sample complexity

Supervised learning

$T:$

- Feature vector $x \in \mathcal{X}$ in feature space $\mathcal{X}$
- Label $y \in \mathcal{Y}$ in label space $\mathcal{Y}$
- Concept $c : \mathcal{X} \to \mathcal{Y}, c \in \mathcal{C}$ where $\mathcal{C}$ is a concept class.
- Find a hypothesis $h \in \mathcal{H}$ such that $h(x) \approx c(x), \forall x \in \mathcal{X}$ for some hypothesis class $\mathcal{H}$

$E:$

- Sample (data set) $S = \{(x_1, c(x_1))), \ldots, (x_m, c(x_m)\}$ for $x_i \overset{i.i.d.}{\sim} \mathcal{D}$ *independent and identically distributed (i.i.d.)* sampled from unknown data distribution $\mathcal{D}$

$P:$

- Loss (risk) function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$

  $L(y, \widehat{y}) = 1_{y \neq \widehat{y}}$ for discrete $\mathcal{Y}$ (zero-one loss)

  $L(y, \widehat{y}) = (y - \widehat{y})^2$ for continuous $\mathcal{Y}$ (squared error)

  where $y \in \mathcal{Y}$ is observed label and $\widehat{y} \in \mathcal{Y}$ is a prediction.

PAC Learnability

- Generalization error (risk) is $R(h) = \mathbb{P}_{x \sim \mathcal{D}}[h(x) \neq c(x)]$
- Empirical error (risk) is $\widehat{R}_S(h) = \frac{1}{m} \sum_{i=1}^{m} 1_{h(x_i) \neq c(x_i)}$. According to the law of large numbers $\mathbb{E}[\widehat{R}_S(h)] = R(h)$.
- $\mathcal{C}$ is **PAC-learnable** if $\exists$ an algorithm $\mathcal{A}$ returning h_S and a polynomial function $poly(\cdot,\cdot,\cdot,\cdot)$ s.t. $\forall \epsilon > 0, \delta > 0, \mathcal{D}, c \in \mathcal{C}$ it holds for any $m \geq poly(1/\epsilon, 1/\delta, n, size(c))$ that

$$\mathbb{P}_{S \sim \mathcal{D}^m}[R(h_S) \leq \epsilon] \geq 1 - \delta$$

  where representing $x \in \mathcal{X}$ costs $O(n)$ and $c \in \mathcal{C}$ at most $size(c)$.
- PAC: Probably Approximately Correct
  Probably $\Rightarrow$ high probability $\Rightarrow \delta \approx 0,$
  Approximately Correct $\Rightarrow$ high confidence $\Rightarrow \epsilon \approx 0.$

A learning bound

- Let $\mathcal{A}$ return a **consistent** hypothesis h_S, i.e. $\widehat{R}_S(h_S) = 0$, then $\forall \epsilon > 0, \delta > 0$,

$$m \geq \frac{1}{\epsilon}\Big(\log |\mathcal{H}| + \log \frac{1}{\delta}\Big) \Rightarrow \mathbb{P}_{S \sim \mathcal{D}^m}[R(h_S) \leq \epsilon] \geq 1 - \delta$$

  Equivalently, $\forall \epsilon > 0$

$$\mathbb{P}_{S \sim \mathcal{D}^m}\Big[R(h_S) \leq \frac{1}{m}\Big(\log |\mathcal{H}| + \log \frac{1}{\delta}\Big)\Big] \geq 1 - \delta$$

  .

That is, the success of the learning algorithm depends on

- Sample size (the larger the better)
- Hypothesis set size (the smaller the better)

The stochastic output case

- Sample (data set) $S = \{(x_1, y_1)), \ldots, (x_m, y_m)\}$ for $(x_i, y_i) \overset{i.i.d.}{\sim} \mathcal{D}$ *independent and identically distributed (i.i.d.)* sampled from unknown data distribution $\mathcal{D}$
- Find h that minimizes

$$R(h) = \mathbb{P}_{(x,y)\sim\mathcal{D}}[L(h(x), y)]$$

- $\mathcal{A}$ is an **agnostic PAC learning** algorithm if $\exists$ an algorithm $\mathcal{A}$ returning h_S and a polynomial function $poly(\cdot,\cdot,\cdot,\cdot)$ s.t. $\forall \epsilon > 0, \delta > 0, \mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ it holds

$$\mathbb{P}_{S\sim\mathcal{D}^m}[R(h_S) - \min_{h\in\mathcal{H}} R(h) \leq \epsilon] \geq 1 - \delta$$

  for any $m \geq poly(1/\epsilon, 1/\delta, n, size(c))$.
- $\mathcal{A}$ is an **efficient agnostic PAC learning** algorithm if its time complexity is $poly(1/\epsilon, 1/\delta, n, size(c))$.

Bayes Error

- Bayes error: $R^* = \min_h R(h)$
- Bayes hypothesis: $R(h) = R^*$
- When $y = c(x)$, $R^* = 0$ as Bayes hypothesis can be chosen as c
- $\forall x \in X$, $h_{Bayes}(x) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \, \mathbb{P}[y|x]$

Measuring the capacity of infinite $\mathcal{H}$

Way 1: Rademacher complexity

- Assume $L : \mathcal{Y} \times \mathcal{Y} \mapsto [0, 1]$, then **Empirical Rademacher complexity**

$$\widehat{\mathfrak{R}}_S(\mathcal{H}) = \mathbb{E}_{\boldsymbol{\sigma}}\Big[\max_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i L(h(x_i), y_i)\Big]$$

  where $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_m)$ with σ_i (Rademacher variables) taking random values in $\{-1, +1\}$.
- **Rademacher complexity**: $\mathfrak{R}_m(\mathcal{H}) = \mathbb{E}_{S \sim \mathcal{D}^m}[\widehat{\mathfrak{R}}_S(\mathcal{H})]$.
- and its learning bound, with probability at least $1 - \delta$

$$\mathbb{E}[L(h(x), y))] \leq \frac{1}{m} \sum_{i=1}^{m} L(h(x_i, y) + 2\mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log(1/\delta)}{2m}}$$

$$\mathbb{E}[L(h(x), y))] \leq \frac{1}{m} \sum_{i=1}^{m} L(h(x_i, y) + 2\widehat{\mathfrak{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\log(2/\delta)}{2m}}$$

Measuring the capacity of infinite $\mathcal{H}$

Way 2: Vapnik-Chervonenkis Dimension

- **Growth function** $\Pi_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$ is

$$\forall m \in \mathbb{N}, \qquad \Pi_{\mathcal{H}}(m) = \max_{\{x_1,\ldots,x_m\} \subseteq \mathcal{X}} \Big| \{(h(x_1), \ldots, h(x_m))\} \Big|.$$

- Assume $\mathcal{Y} = \{-1, +1\}$ then

$$R(h) \leq \widehat{R}_S(h) + \sqrt{\frac{2 \log \Pi_{\mathcal{H}}(m)}{m}} + \sqrt{\frac{\log(1/\delta)}{2m}}$$

- **Vapnik-Chervonenkis (VC) Dimension:**

$$VC(\mathcal{H}) = \max\{m : \Pi_{\mathcal{H}}(m) = 2^m\}$$

If $\Pi_{\mathcal{H}}(m) = 2^m$, the set S is said to be **shattered** by $\mathcal{H}$.

- Assume $\mathcal{Y} = \{-1, +1\}$ then

$$R(h) \leq \widehat{R}_S(h) + \sqrt{\frac{2VC(\mathcal{H}) \log(em/VC(\mathcal{H}))}{m}} + \sqrt{\frac{\log(1/\delta)}{2m}}$$

Empirical Risk Minimization (ERM)

- **Model selection:** The choice of $\mathcal{H}$
- The estimation-approximation dilemma

$$\underbrace{R(h) - R^*}_{excess\ error} = \underbrace{\Big(R(h) - \inf_{h \in \mathcal{H}} R(h)\Big)}_{estimation\ error} + \underbrace{\Big(\inf_{h \in \mathcal{H}} R(h) - R^*\Big)}_{approximation\ error}$$

  Agnostic PAC-learning considers only estimation error.

- **Empirical Risk Minimization:**

$$\mathcal{A}_{ERM}(S, \mathcal{H}) = \Big\{ h_S^{ERM} \Big| \operatorname*{argmin}_{h \in \mathcal{H}} \widehat{R}_S(h) \Big\}$$

  the performance of which can be bounded as

$$\mathbb{P}\Big[R(h_S^{ERM}) - \inf_{h \in \mathcal{H}} R(h) > \epsilon\Big] \\ \leq \mathbb{P}\Big[\sup_{h \in \mathcal{H}} |R(h) - \widehat{R}_S(h)| > \frac{\epsilon}{2}\Big] \leq 2e^{-2m(\epsilon - \mathfrak{R}_m(\mathcal{H}))}$$

Structural Risk Minimization (SRM)

- ERM disregards the complexity of $\mathcal{H}$ and often performs poorly because of the estimation-approximation dilemma.
- Choose large $\mathcal{H} = \bigcup_{k>1} \mathcal{H}_k$ such that $\mathcal{H}_k \subset \mathcal{H}_{k+1}, \forall k \geq 1$
- SRM hinges on the bound below

$$R(h) \leq \widehat{R}_S(h) + \mathfrak{R}_m(\mathcal{H}_{k(h)}) + \sqrt{\frac{\log k}{m}} + \sqrt{\frac{\log(2/\delta)}{2m}}$$

and performs

$$\mathcal{A}_{SRM}(S, \mathcal{H}) = \left\{ h_S^{SRM} \middle| \underset{k\geq 1, h \in \mathcal{H}_k}{\operatorname{argmin}} \ \widehat{R}_S(h) + \mathfrak{R}_m(\mathcal{H}_k) + \sqrt{\frac{\log k}{m}} \right\}$$

with bound

$$R(h_S^{SRM}) \leq \inf_{h \in \mathcal{H}} \left(R(h) + 2\mathfrak{R}_m(\mathcal{H}_{k(h)}) + \sqrt{\frac{\log k(h)}{m}} \right) + \sqrt{\frac{2\log(3/\delta)}{m}}$$

Remember this plot?

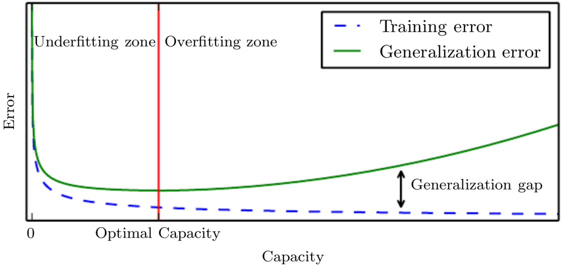

Figure: Goodfellow et al., Deep Learning, MIT Press, 2016