Probably approximately correct learning

Maximilian Kasy

February 2021

These slides summarize Chapter 2-6 of the following textbook:

Shalev-Shwartz, S. and Ben-David, S. (2014).

Understanding machine learning: From theory to algorithms.
Cambridge University Press.

Setup and basic definitions

VC dimension and the Fundamental Theorem of statistical learning

Setup and notation

- Features (predictive covariates): x
- Labels (outcomes): $y \in \{0, 1\}$
- Training data (sample): $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{n}$
- Data generating process: (x_i, y_i) are i.i.d. draws from a distribution $\mathcal{D}$
- Prediction rules (hypotheses): $h: x \to \{0, 1\}$

Learning algorithms

- Risk (generalization error): Probability of misclassification

$$L_{\mathcal{D}}(h) = E_{(x,y)\sim\mathcal{D}}\left[\mathbf{1}(h(x) \neq y)\right].$$

- Empirical risk: Sample analog of risk,

$$L_{\mathcal{S}}(h) = \frac{1}{n}\sum_i \mathbf{1}(h(x) \neq y).$$

- Learning algorithms
  map samples $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n$
  into predictors $h_{\mathcal{S}}$.

Empirical risk minimization

- Optimal predictor:

$$h^*_{\mathcal{D}} = \underset{h}{\operatorname{argmin}}\ L_{\mathcal{D}}(h) = \mathbf{1}(E_{(x,y)\sim\mathcal{D}}[y|x] \geq 1/2).$$

- Hypothesis class for h: $\mathcal{H}$.
- Empirical risk minimization:

$$h^{ERM}_{\mathcal{S}} = \underset{h\in\mathcal{H}}{\operatorname{argmin}}\ L_{\mathcal{S}}(h).$$

- Special cases (for more general loss functions):
  Ordinary least squares, maximum likelihood,
  minimizing empirical risk over model parameters.

(Agnostic) PAC learnability

Definition 3.3

A hypothesis class $\mathcal{H}$ is agnostic probably approximately correct (PAC) learnable if

- there exists a learning algorithm $h_{\mathcal{S}}$
- such that for all $\epsilon, \delta \in (0, 1)$ there exists an $n < \infty$
- such that for all distributions $\mathcal{D}$

$$L_{\mathcal{D}}(h_{\mathcal{S}}) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$$

- with probability of at least $1 - \delta$
- over the draws of training samples

$$\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{n} \sim^{iid} \mathcal{D}.$$

Discussion

- Definition is not specific to $0/1$ prediction error loss.
- **Worst case** over **all possible distributions** $\mathcal{D}$.
- Requires small **regret**:
  The oracle-best predictor in $\mathcal{H}$ doesn't do much better.
- Comparison to the best predictor in the **hypothesis class** $\mathcal{H}$ rather than to the unconditional best predictor $h^*_{\mathcal{D}}$.
- $\Rightarrow$ The smaller the hypothesis class $\mathcal{H}$
  the easier it is to fulfill this definition.
- Definition requires small (relative) loss **with high probability**, not just in expectation.

Question: How does this relate to alternative performance criteria?

ϵ-representative samples

- *Definition 4.1*
  A training set $\mathcal{S}$ is called ϵ-representative if

$$\sup_{h \in \mathcal{H}} |L_{\mathcal{S}}(h) - L_{\mathcal{D}}(h)| \leq \epsilon.$$

- *Lemma 4.2*
  Suppose that $\mathcal{S}$ is $\epsilon/2$-representative.
  Then the empirical risk minimization predictor $h_{\mathcal{S}}^{ERM}$ satisfies

$$L_{\mathcal{D}}(h_{\mathcal{S}}^{ERM}) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon.$$

- *Proof:* if $\mathcal{S}$ is $\epsilon/2$-representative,
  then for all $h \in \mathcal{H}$

$$L_{\mathcal{D}}(h_{\mathcal{S}}^{ERM}) \leq L_{\mathcal{S}}(h_{\mathcal{S}}^{ERM}) + \epsilon/2 \leq L_{\mathcal{S}}(h) + \epsilon/2 \leq L_{\mathcal{D}}(h) + \epsilon.$$

Uniform convergence

- *Definition 4.3*
  $\mathcal{H}$ has the uniform convergence property if
  - for all $\epsilon, \delta \in (0, 1)$ there exists an $n < \infty$
  - such that for all distributions $\mathcal{D}$
  - with probability of at least $1 - \delta$ over draws of training samples $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{n} \sim^{iid} \mathcal{D}$
  - it holds that $\mathcal{S}$ is ϵ-representative.

- *Corollary 4.4*
  If $\mathcal{H}$ has the uniform convergence property, then
  1. the class is agnostically PAC learnable, and
  2. $h_{\mathcal{S}}^{ERM}$ is a successful agnostic PAC learner for $\mathcal{H}$.

- *Proof:* From the definitions and Lemma 4.2.

Finite hypothesis classes

- *Corollary 4.6*
  Let $\mathcal{H}$ be a finite hypothesis class, and assume that loss is in $[0, 1]$.
  Then $\mathcal{H}$ enjoys the uniform convergence property, where we set

$$n = \left\lceil \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2} \right\rceil$$

  The class $\mathcal{H}$ is therefore agnostically PAC learnable.

- *Sketch of proof:* Union bound over $h \in \mathcal{H}$,
  plus Hoeffding's inequality,

$$P(|L_{\mathcal{S}}(h) - L_{\mathcal{D}}(h)| > \epsilon) \leq 2 \exp(-2n\epsilon^2).$$

No free lunch

Theorem 5.1

- Consider any learning algorithm $h_{\mathcal{S}}$ for binary classification with $0/1$ loss on some domain $\mathcal{X}$.
- Let $n < |\mathcal{X}|/2$ be the training set size.
- Then there exists a $\mathcal{D}$ on $\mathcal{X} \times \{0, 1\}$, such that $y = f(x)$ for some f with probability 1, and
- with probability of at least $1/7$ over the distribution of $\mathcal{S}$,

$$L_{\mathcal{D}}(h_{\mathcal{S}}) \geq 1/8.$$

- *Intuition of proof:*
  - Fix some set $\mathcal{C} \subset \mathcal{X}$ with $|\mathcal{C}| = 2n$,
  - consider $\mathcal{D}$ uniform on $\mathcal{C}$,
    and corresponding to arbitrary mappings $y = f(x)$.
  - Lower-bound worst case $L_{\mathcal{D}}(h_{\mathcal{S}})$
    by the average of $L_{\mathcal{D}}(h_{\mathcal{S}})$ over all possible choices of f.

- *Corollary 5.2*
  Let $\mathcal{X}$ be an infinite domain set
  and let $\mathcal{H}$ be the set of all functions from $\mathcal{X}$ to $\{0, 1\}$.
  Then $\mathcal{H}$ is not PAC learnable.

Error decomposition

$$
\begin{aligned}
L_{\mathcal{D}}(h_{\mathcal{S}}) &= \epsilon_{app} + \epsilon_{est} \\
\epsilon_{app} &= \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) \\
\epsilon_{est} &= L_{\mathcal{D}}(h_{\mathcal{S}}) - \min_{h \in \mathcal{H}} .
\end{aligned}
$$

- Approximation error: ϵ_{app}.
- Estimation error: ϵ_{est}.
- **Bias-complexity tradeoff**:
  Increasing $\mathcal{H}$ increases ϵ_{est}, but decreases ϵ_{app}.
- Learning theory provides bounds on ϵ_{est}.

Setup and basic definitions

VC dimension and the Fundamental Theorem of statistical learning

Shattering

From now on, restrict to $y \in \{0, 1\}$.

Definition 6.3

- A hypothesis class $\mathcal{H}$
- shatters a finite set $C \subset \mathcal{X}$
- if the restriction of $\mathcal{H}$ to C (denoted $\mathcal{H}_C$)
- is the set of all functions from C to $\{0, 1\}$.
- In this case: $|\mathcal{H}_C| = 2^{|C|}$.

VC dimension

Definition 6.5

- The VC-dimension of a hypothesis class $\mathcal{H}$, $VCdim(\mathcal{H})$,
- is the maximal size of a set $C \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$.
- If $\mathcal{H}$ can shatter sets of arbitrarily large size
- we say that $\mathcal{H}$ has infinite VC-dimension.

Corollary of the no free lunch theorem:

- Let $\mathcal{H}$ be a class of infinite VC-dimension.
- Then $\mathcal{H}$ is not PAC learnable.

Examples

- Threshold functions: $h(x) = \mathbf{1}(x \le c)$.
  $VCdim = 1$
- Intervals: $h(x) = \mathbf{1}(x \in [a, b])$.
  $VCdim = 2$
- Finite classes: $h \in \mathcal{H} = \{h_1, \ldots, h_n\}$.
  $VCdim \le \log_2(n)$
- $VCdim$ is not always # of parameters: $h_\theta(x) = \lceil .5 sin(\theta x) \rceil$, $\theta \in \mathbb{R}$.
  $VCdim = \infty$.

The Fundamental Theorem of Statistical learning

Theorem 6.7

- Let $\mathcal{H}$ be a hypothesis class of functions
- from a domain $\mathcal{X}$ to $\{0, 1\}$,
- and let the loss function be the $0 - 1$ loss.

Then, the following are equivalent:

1. $\mathcal{H}$ has the uniform convergence property.
2. Any ERM rule is a successful agnostic PAC learner for $\mathcal{H}$.
3. $\mathcal{H}$ is agnostic PAC learnable.
4. $\mathcal{H}$ has a finite VC-dimension.

Proof

1. $\to$ 2.: Shown above (Corollary 4.4).
2. $\to$ 3.: Immediate.
3. $\to$ 4.: By the no free lunch theorem.
4. $\to$ 1.: That's the tricky part.
    - Sauer-Shelah-Perles's Lemma.
    - Uniform convergence for classes of small effective size.

Growth function

- The growth function of $\mathcal{H}$ is defined as

$$\tau_{\mathcal{H}}(n) := \max_{C \subset \mathcal{X}: |C| = n} |\mathcal{H}_C|.$$

- Suppose that $d = VCdim(\mathcal{H}) \leq \infty$.
  Then for $n \leq d$, $\tau_{\mathcal{H}}(n) = 2^n$ by definition.

Sauer-Shelah-Perles's Lemma

Lemma 6.10
For $d = VCdim(\mathcal{H}) \leq \infty$,

$$\tau_{\mathcal{H}}(b) \leq \max_{C \subset \mathcal{X} : |C| = n} |\{B \subseteq C : \mathcal{H} \text{ shatters } B\}|$$
$$\leq \sum_{i=0}^{d} \binom{n}{i} \leq \left(\frac{en}{d}\right)^{d}.$$

- First inequality is the interesting / difficult one.
- Proof by induction.

Uniform convergence for classes of small effective size

Theorem 6.11

- For all distributions $\mathcal{D}$ and every $\delta \in (0,1)$
- with probability of at least $1-\delta$ over draws of training samples $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n \sim^{iid} \mathcal{D}$,
- we have

$$\sup_{h \in \mathcal{H}} |L_{\mathcal{S}}(h) - L_{\mathcal{D}}(h)| \leq \frac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2n))}}{\delta\sqrt{2n}}.$$

Remark

- We already saw that uniform convergence holds for finite classes.
- This shows that uniform convergence holds for classes with polynomial growth of

$$\tau_{\mathcal{H}}(m) = \max_{C \subset \mathcal{X} : |C| = m} |\mathcal{H}_C|.$$

- These are exactly the classes with finite VC dimension, by the preceding lemma.