Foundations of machine learning

Probably approximately correct learning

Maximilian Kasy

Department of Economics, University of Oxford

Winter 2025

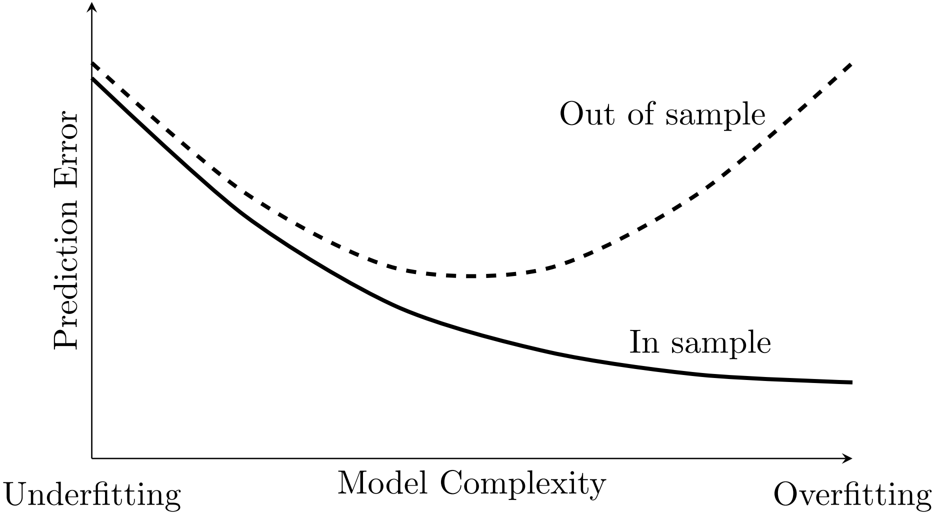
Outline

- Definitions:
  - Classification and prediction problems.
  - Empirical risk minimization.
  - PAC learnability.
- Proving the "Fundamental Theorem of statistical learning:"
  - ε-representative samples.
  - Uniform convergence.
  - No free lunch.
  - Shatterings.
  - VC dimension.

Takeaways for this part of class

- Classification and prediction is about out-of-sample prediction errors.
- These can be decomposed into an approximation error ("bias") and an estimation error ("variance").
- There is a trade-off between the two.
  Larger classes of predictors imply less approximation error (no "underfitting"), but more estimation error ("overfitting").
- The worst-case estimation error depends on the VC-dimension of the class of predictors considered.

Our goal: Understanding this figure



Setup and basic definitions

VC dimension and the Fundamental Theorem of statistical learning

References

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

$$L(h, \mathcal{D}) = E_{(X,Y)\sim\mathcal{D}}\left[1(h(X) \neq Y)\right].$$

- Empirical risk: Sample analog of risk,

$$L(h, \mathcal{S}) = \frac{1}{n}\sum_i 1(h(X) \neq Y).$$

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

- Notation:
  h corresponds to a in the decision theory slides,
  $\mathcal{D}$ corresponds to θ.

Chihuahua or muffin?

Empirical risk minimization

- Optimal predictor:

$$h^*_{\mathcal{D}} = \underset{h}{\operatorname{argmin}}\, L(h, \mathcal{D}) = 1(E_{(X,Y)\sim\mathcal{D}}[Y|X] \geq 1/2).$$

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

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

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

Practice problem

How does empirical risk minimization relate

1. to ordinary least squares, and

2. to maximum likelihood estimation?

(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 $\varepsilon, \delta \in (0,1)$ there exists an $n < \infty$
- such that for all distributions $\mathcal{D}$

$$L(h_{\mathcal{S}}, \mathcal{D}) \leq \inf_{h \in \mathcal{H}} L(h, \mathcal{D}) + \varepsilon$$

- 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.

Practice problem

How does PAC learnability relate to the performance criteria we discussed in the decision theory slides?

ε-representative samples

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

$$\sup_{h \in \mathcal{H}} |L(h, \mathcal{S}) - L(h, \mathcal{D})| \leq \varepsilon.$$

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

$$L(h_{\mathcal{S}}^{ERM}, \mathcal{D}) \leq \inf_{h \in \mathcal{H}} L(h, \mathcal{D}) + \varepsilon.$$

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

$$L(h_{\mathcal{S}}^{ERM}, \mathcal{D}) \leq L(h_{\mathcal{S}}^{ERM}, \mathcal{S}) + \varepsilon/2 \leq L(h, \mathcal{S}) + \varepsilon/2 \leq L(h, \mathcal{D}) + \varepsilon.$$

Uniform convergence

- *Definition 4.3*
  $\mathcal{H}$ has the uniform convergence property if
  - for all $\varepsilon, \delta \in (0,1)$ 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.

Probability theory intermission

Large deviations

- Suppose that
$$E[\exp(\lambda \cdot (Y - E[Y]))] \leq \exp(\psi(\lambda)).$$

- Let $\bar{Y}_n = \frac{1}{n}\sum_{1 \leq i \leq n} Y_i$ for i.i.d. Y_i.
  Then, by Markov's inequality and independence across t,

$$\begin{aligned}
P(\bar{Y}_n - E[Y] > \varepsilon) &\leq \frac{E[\exp(\lambda \cdot (\bar{Y}_n - E[Y]))]}{\exp(\lambda \cdot \varepsilon)} \\
&= \frac{\prod_{1 \leq i \leq n} E[\exp((\lambda/n) \cdot (Y_i - E[Y]))]}{\exp(\lambda \cdot \varepsilon)} \\
&\leq \exp(n\psi(\lambda/n) - \lambda \cdot \varepsilon).
\end{aligned}$$

Large deviations continued

- Define the Legendre-transformation of ψ as

$$\psi^*(\varepsilon) = \sup_{\lambda \geq 0} \left[\lambda \cdot \varepsilon - \psi(\lambda)\right].$$

- Taking the $\inf$ over λ in the previous slide implies

$$P(\bar{Y}_n - E[Y] > \varepsilon) \leq \exp(-n \cdot \psi^*(\varepsilon)).$$

- For distributions bounded by $[0,1]$:
  $\psi(\lambda) = \lambda^2/8$ and $\psi^*(\varepsilon) = 2\varepsilon^2$.

- This implies Hoeffding's inequality:

$$P(\bar{Y}_n - E[Y] > \varepsilon) \leq \exp(-2n\varepsilon^2).$$

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

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

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

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

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

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 $\mathcal{S}$,

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

- *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 $L(h_{\mathcal{S}}, \mathcal{D})$
    by the average of $L(h_{\mathcal{S}}, \mathcal{D})$ 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

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

- 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}.

Practice problem

Write out the approximation error and the (expected) estimation error for the case where

1. loss is given by the squared prediction error, and
2. $\mathcal{H}$ is given by the set of linear predictors.

Setup and basic definitions

VC dimension and the Fundamental Theorem of statistical learning

References

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: $h(X) = 1(X \leq c)$.
  $VCdim = 1$

- Intervals: $h(X) = 1(X \in [a, b])$.
  $VCdim = 2$

- Finite classes: $h \in \mathcal{H} = \{h_1, \ldots, h_n\}$.
  $VCdim \leq \log_2(n)$

- $VCdim$ is not always # of parameters: $h_\theta(X) = \lceil .5sin(\theta X) \rceil$, $\theta \in \mathbb{R}$.
  $VCdim = \infty$.

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

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

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.

References

Shalev-Shwartz, S. and Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. *Cambridge University Press, chapters 2-6.*