Foundations of Machine Learning

Introduction to Learning Theory

Oct 2021

Vineeth N Balasubramanian

आई आई टी हैदराबाद
IIT Hyderabad

Today

- Bayes Error
- Introduction to Learning Theory

Optimality of Bayes Decision Rule

- Let X be a *random variable* over the space Ω
- Two category decision problem:

$$H_1: X \in \omega_1$$

$$H_2: X \in \omega_2$$

- Optimal decision is:

$$\text{Choose } H_1 \text{ when } p(x|\omega_1)P(\omega_1) \geq p(x|\omega_2)P(\omega_2)$$

$$\text{Choose } H_2 \text{ when } p(x|\omega_1)P(\omega_1) < p(x|\omega_2)P(\omega_2)$$

Bayes Decision Rule

Slide Credit: slecture of Aaron Michaux, Purdue University

Optimality of Bayes Decision Rule

- Consider an arbitrary decision rule:
  - Partition Ω into two disjoint regions: $\mathcal{R}_1$ and $\mathcal{R}_2$

    Choose H_1 if $X \in \mathcal{R}_1$

    Choose H_2 if $X \in \mathcal{R}_2$
  - This decision rule is possibly non-optimal.

Slide Credit: slecture of Aaron Michaux, Purdue University

Optimality of Bayes Decision Rule

- Consider the Bayes decision rule:
  - Partition Ω into two disjoint regions: Ω_1 and Ω_2

    Choose H_1 if $X \in \Omega_1$

    Choose H_2 if $X \in \Omega_2$
  - Where:

$$\Omega_1 = \{x \in \Omega: p(x|\omega_1)P(\omega_1) \geq p(x|\omega_2)P(\omega_2)\}$$

$$\Omega_2 = \{x \in \Omega: p(x|\omega_1)P(\omega_1) < p(x|\omega_2)P(\omega_2)\}$$

Slide Credit: slecture of Aaron Michaux, Purdue University

Optimality of Bayes Decision Rule

- Two types of error: one for each of the decision rules
- For the arbitrary decision rule:

$$P(error) = P(X \in \mathcal{R}_1, \omega_2) + P(X \in \mathcal{R}_2, \omega_1)$$

$$= P(X \in \mathcal{R}_1 | \omega_2) P(\omega_2) + P(X \in \mathcal{R}_2 | \omega_1) P(\omega_1)$$

$$= \int_{\mathcal{R}_1} p(x|\omega_2) P(\omega_2) dx + \int_{\mathcal{R}_2} p(x|\omega_1) P(\omega_1) dx$$

- For the Bayesian decision rule

$$P(error_{Bayes}) = \int_{\Omega_1} p(x|\omega_2) P(\omega_2) dx + \int_{\Omega_2} p(x|\omega_1) P(\omega_1) dx$$

Slide Credit: slecture of Aaron Michaux, Purdue University

Optimality of Bayes Decision Rule

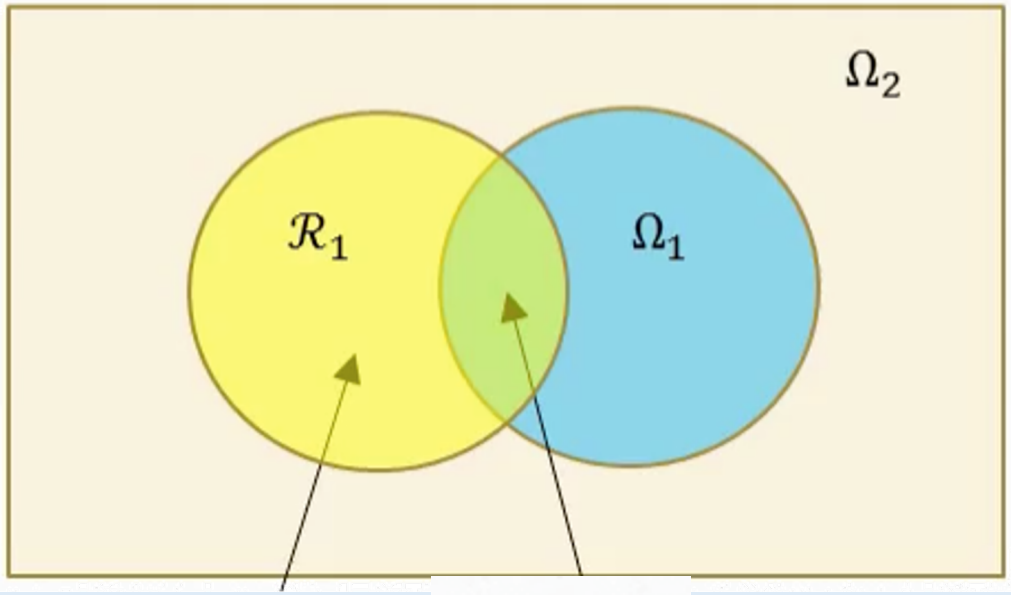
- Let $\Delta(error) = P(error) - P(error_{Bayes})$
- First recall that both $\mathcal{R}_1, \mathcal{R}_2$ and Ω_1, Ω_2 are partitions of Ω

$$\mathcal{R}_1 = (\mathcal{R}_1 \cap \Omega_1) \cup (\mathcal{R}_1 \cap \Omega_2)$$

$$\mathcal{R}_2 = (\mathcal{R}_2 \cap \Omega_1) \cup (\mathcal{R}_2 \cap \Omega_2)$$

$$\Omega_1 = (\Omega_1 \cap \mathcal{R}_1) \cup (\Omega_1 \cap \mathcal{R}_2)$$

$$\Omega_2 = (\Omega_2 \cap \mathcal{R}_1) \cup (\Omega_2 \cap \mathcal{R}_2)$$



Slide Credit: slecture of Aaron Michaux, Purdue University

Optimality of Bayes Decision Rule

$$\Delta(error) = \int_{\mathcal{R}_1} p(x|\omega_2)P(\omega_2)dx + \int_{\mathcal{R}_2} p(x|\omega_1)P(\omega_1)dx - \int_{\Omega_1} p(x|\omega_2)P(\omega_2)dx - \int_{\Omega_2} p(x|\omega_1)P(\omega_1)dx$$

$$= P(\omega_2)\left[\int_{\mathcal{R}_1} p(x|\omega_2)dx - \int_{\Omega_1} p(x|\omega_2)dx\right] + P(\omega_1)\left[\int_{\mathcal{R}_2} p(x|\omega_1)dx - \int_{\Omega_2} p(x|\omega_1)dx\right]$$

$$= P(\omega_2)\left[\int_{\mathcal{R}_1 \cap \Omega_2} p(x|\omega_2)dx - \int_{\Omega_1 \cap \mathcal{R}_2} p(x|\omega_2)dx\right] + P(\omega_1)\left[\int_{\mathcal{R}_2 \cap \Omega_1} p(x|\omega_1)dx - \int_{\Omega_2 \cap \mathcal{R}_1} p(x|\omega_1)dx\right]$$

Slide Credit: slecture of Aaron Michaux, Purdue University

Optimality of Bayes Decision Rule

$$= \int_{\mathcal{R}_1 \cap \Omega_2} [p(x|\omega_2)P(\omega_2) - p(x|\omega_1)P(\omega_1)]dx + \int_{\Omega_1 \cap \mathcal{R}_2} [p(x|\omega_1)P(\omega_1) - p(x|\omega_2)P(\omega_2)]dx$$

Recall,

$$\Omega_1 = \{x \in \Omega : p(x|\omega_1)P(\omega_1) \geq p(x|\omega_2)P(\omega_2)\} \Rightarrow \int_{\mathcal{R}_2 \cap \Omega_1} [p(x|\omega_1)P(\omega_1) - p(x|\omega_2)P(\omega_2)]dx \geq 0$$

$$\Omega_2 = \{x \in \Omega : p(x|\omega_2)P(\omega_2) > p(x|\omega_1)P(\omega_1)\} \Rightarrow \int_{\mathcal{R}_1 \cap \Omega_2} [p(x|\omega_2)P(\omega_2) - p(x|\omega_1)P(\omega_1)]dx \geq 0$$

$$\Delta(error) \geq 0$$

Slide Credit: slecture of Aaron Michaux, Purdue University

Optimality of Bayes Decision Rule

$$\Delta(error) = P(error) - P(error_{Bayes}) \geq 0$$

$$\Leftrightarrow P(error) \geq P(error_{Bayes})$$

Recall:

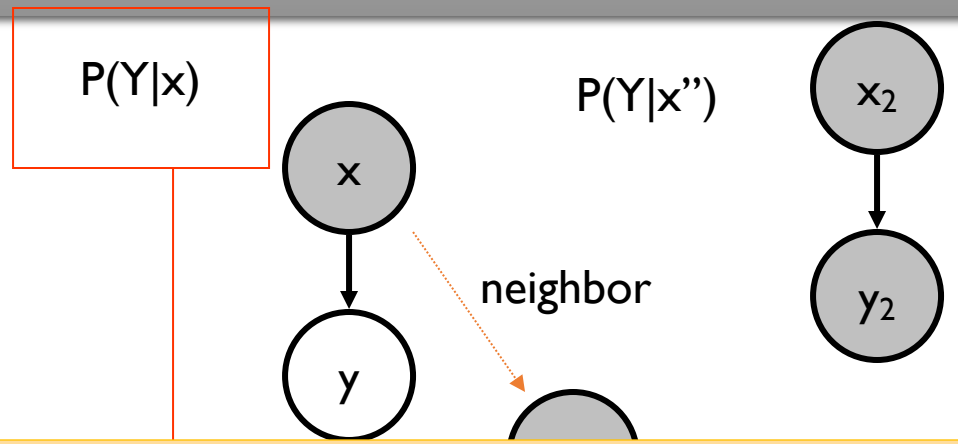
$$\text{Choose } H_1 \text{ when } p(x|\omega_1)P(\omega_1) \geq p(x|\omega_2)P(\omega_2)$$

$$\text{Choose } H_2 \text{ when } p(x|\omega_1)P(\omega_1) < p(x|\omega_2)P(\omega_2)$$

Slide Credit: slecture of Aaron Michaux, Purdue University

Recall: Convergence of 1-NN

$$P(\text{knnError})$$
$$= 1 - \Pr(y = y_1)$$
$$= 1 - \sum_{y'} \Pr(Y = y' | x)^2$$



Possible to show that: as the size of training data set approaches infinity, the one nearest neighbor classifier guarantees an error rate of no worse than twice the Bayes error rate (the minimum achievable error rate given the distribution of the data).

Today

- Bayes Error
- Introduction to Learning Theory

Towards formalizing 'learning'

- What does it mean to learn a concept?
  - Gain knowledge or experience of the concept.

- The basic process of learning
  - Observe a phenomenon
  - Construct a model from observations
  - Use that model to make decisions/predictions

How can we make this more precise?

Source: Nakul Verma, Columbia Univ

A statistical machinery for learning

- Phenomenon of interest:
  - Input space: X; Output space: Y
  - There is an unknown distribution D over (X,Y)
  - The learner observes m examples $(x_1,y_1),\ldots(x_m,y_m)$ drawn from D
- Construct a model:
  - Let F be a collection of models, where each f: X -> Y predicts y given x
  - From m observations, select a model f_m in F which predicts well.

Machine Learning

$$\mathrm{err}(f) := \mathbb{P}_{(x,y)\sim\mathcal{D}}\left[f(x) \neq y\right]$$ (generalization error of *f*)

- We can say that we have ***learned*** the phenomenon if

$$\mathrm{err}(f_m) - \mathrm{err}(f^*) \leq \epsilon \qquad f^* := \mathrm{argmin}_{f\in\mathcal{F}}\,\mathrm{err}(f)$$

for any tolerance level $\epsilon > 0$ of our choice.

Source: Nakul Verma, Columbia Univ

PAC Learning

For all tolerance levels $\epsilon > 0$, and all confidence levels $\delta > 0$, if there exists some model selection algorithm $\mathcal{A}$ that selects $f_m^{\mathcal{A}} \in \mathcal{F}$ from *m* observations ie, $\mathcal{A} : (x_i, y_i)_{i=1}^m \mapsto f_m^{\mathcal{A}}$, and has the property:

with probability at least $1 - \delta$ over the draw of the sample,

$$\text{err}(f_m^{\mathcal{A}}) - \text{err}(f^*) \leq \epsilon$$

We call

- The model class $\mathcal{F}$ is PAC-learnable.
- If the m is polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$, then $\mathcal{F}$ is **efficiently** PAC-learnable

A popular algorithm:

Empirical risk minimizer (ERM) algorithm

$$f_m^{\text{ERM}} := \text{argmin}_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}\{f(x_i) \neq y_i\}$$

Source: Nakul Verma, Columbia Univ

PAC Learning Simple Model Classes

Theorem (finite size $\mathcal{F}$):

Pick any tolerance level $\epsilon > 0$, and any confidence level $\delta > 0$

let $(x_1, y_1), \ldots, (x_m, y_m)$ be m examples drawn from an unknown $\mathcal{D}$

if $m \geq C \cdot \frac{1}{\epsilon^2} \ln \frac{|\mathcal{F}|}{\delta}$, then with probability at least $1 - \delta$

$$\mathrm{err}(f_m^{\mathrm{ERM}}) - \mathrm{err}(f^*) \leq \epsilon$$

F is efficiently PAC-learnable

Source: Nakul Verma, Columbia Univ

Proof Sketch

Define:

$$\mathrm{err}(f) := \mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[\mathbf{1}\{f(x) \neq y\}\Big] \qquad \mathrm{err}_m(f) := \frac{1}{m}\sum_{i=1}^{m}\Big[\mathbf{1}\{f(x_i) \neq y_i\}\Big]$$

(generalization error of f) (sample error of f)

Fix any $f \in \mathcal{F}$ and a sample (x_i, y_i), define random variable

$$\mathbf{Z}_i^f := \mathbf{1}\{f(x_i) \neq y_i\}$$

$$\mathbb{E}\big[\mathbf{Z}_1^f\big] \qquad \frac{1}{m}\sum_{i=1}^{m}\big[\mathbf{Z}_i^f\big]$$

(generalization error of f) (sample error of f)

Source: Nakul Verma, Columbia Univ

Proof Sketch

Lemma (Chernoff-Hoeffding bound '63):

Let $\mathbf{Z_1}, \ldots, \mathbf{Z_m}$ be *m* Bernoulli r.v. drawn independently from **B(*p*)**.

for any tolerance level $\epsilon > 0$

$$\mathbb{P}_{\mathbf{z}_i}\left[\left|\frac{1}{m}\sum_{i=1}^{m}[\mathbf{Z_i}] - \mathbb{E}[\mathbf{Z_1}]\right| > \epsilon\right] \leq 2e^{-2\epsilon^2 m}.$$

A classical result in concentration of measure; Please see http://www-cs-students.stanford.edu/~blynn/pr/markov.html or https://www.probabilitycourse.com/chapter6/6_2_0_probability_bounds.php for more information

Source: Nakul Verma, Columbia Univ

Proof Sketch

Need to analyze

$$\mathbb{P}_{(x_i,y_i)}\left[\text{exists } f \in \mathcal{F},\ \left|\frac{1}{m}\sum_{i=1}^{m}[\mathbf{Z}_i^f] - \mathbb{E}[\mathbf{Z}_1^f]\right| > \epsilon\right]$$

$$\leq \sum_{f\in\mathcal{F}} \mathbb{P}_{(x_i,y_i)}\left[\left|\frac{1}{m}\sum_{i=1}^{m}[\mathbf{Z}_i^f] - \mathbb{E}[\mathbf{Z}_1^f]\right| > \epsilon\right]$$

$$\leq 2|\mathcal{F}|e^{-2\epsilon^2 m} \leq \delta$$

Source: Nakul Verma, Columbia Univ

Proof Sketch

Equivalently, by choosing $m \geq \frac{1}{2\epsilon^2} \ln \frac{2|\mathcal{F}|}{\delta}$ with probability at least $1 - \delta$, for ***all*** $f \in \mathcal{F}$

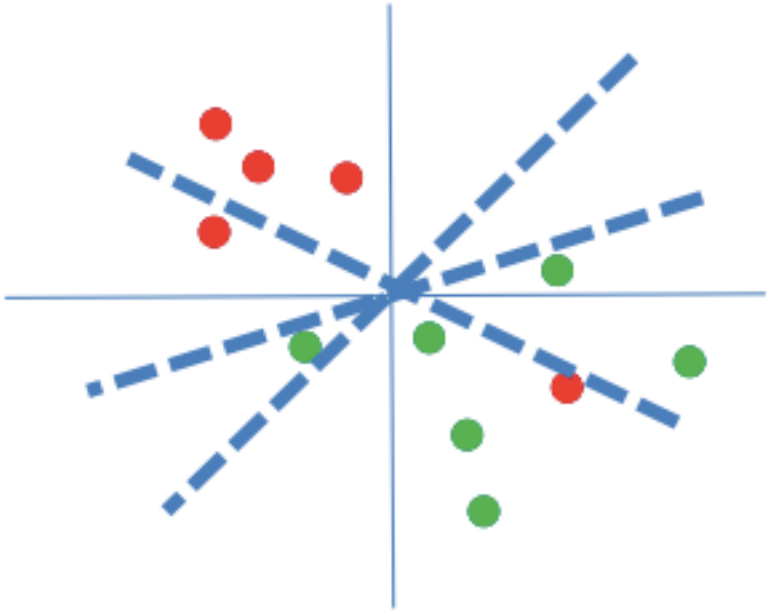
$$\left| \frac{1}{m} \sum_{i=1}^{m} [\mathbf{Z}_i^f] - \mathbb{E}[\mathbf{Z}_1^f] \right| = \left| \text{err}_m(f) - \text{err}(f) \right| \leq \epsilon$$

■

Source: Nakul Verma, Columbia Univ

Learning General Concepts

- Consider linear classification



$$\mathcal{F} = \left\{ \right.$$  $$\left. \right\}$$

$$|\mathcal{F}| = \infty$$

Previous theorem bound is ineffective

Source: Nakul Verma, Columbia Univ

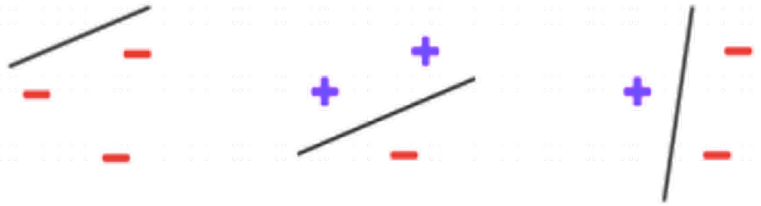
VC Theory

Definition (Vapnik-Chervonenkis or VC dimension):

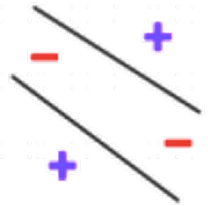
We say that a model class $\mathcal{F}$ as VC dimension *d*, if *d* is the largest set of points $x_1, \ldots, x_d \subset X$ such that for all possible labelings of $x_1, \ldots, x_d$ there exists some $f \in \mathcal{F}$ that achieves that labelling.

Example: $\mathcal{F}$ = linear classifiers in $\mathbf{R}^2$

linear classifiers can realize all possible labellings of 3 points



linear classifiers CANNOT realize all labellings of 4 points



$$\mathrm{VC}(\mathcal{F}) = 3$$

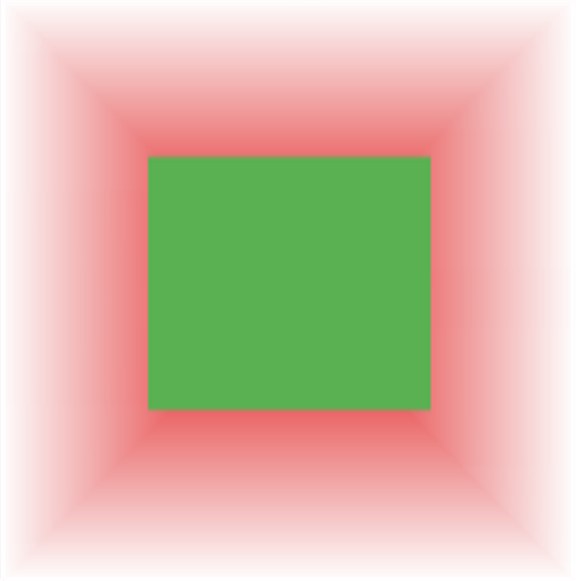
Source: Nakul Verma, Columbia Univ

VC Theory

- What about the VC-dimension of the family of rectangles?

$\mathcal{F}$ = Rectangles in $\mathbf{R}^2$

$$\mathrm{VC}(\mathcal{F}) = 4$$



The class of rectangles cannot realize this labelling

Source: Nakul Verma, Columbia Univ

VC Theory

Theorem (Vapnik-Chervonenkis '71):

Pick any tolerance level $\epsilon > 0$, and any confidence level $\delta > 0$

let $(x_1, y_1), \ldots, (x_m, y_m)$ be m examples drawn from an unknown $\mathcal{D}$

if $m \geq C \cdot \dfrac{\mathrm{VC}(\mathcal{F}) \ln(1/\delta)}{\epsilon^2}$, then with probability at least $1 - \delta$

$$\mathrm{err}(f_m^{\mathrm{ERM}}) - \mathrm{err}(f^*) \leq \epsilon$$

F is efficiently PAC-learnable

Source: Nakul Verma, Columbia Univ

Tightness of VC Bound

Theorem (VC lower bound):

Let $\mathcal{A}$ be any model selection algorithm that given *m* samples, returns a model from $\mathcal{F}$, that is, $\mathcal{A} : (x_i, y_i)_{i=1}^m \mapsto f_m^{\mathcal{A}}$

For all tolerance levels $0 < \epsilon < 1$, and all confidence levels $0 < \delta < 1/4$,

there exists a distribution $\mathcal{D}$ such that if $m \leq C \cdot \frac{\mathrm{VC}(\mathcal{F})}{\epsilon^2}$

$$\mathbb{P}_{(x_i,y_i)}\left[\left|\mathrm{err}(f_m^{\mathcal{A}}) - \mathrm{err}(f^*)\right| > \epsilon\right] > \delta$$

Source: Nakul Verma, Columbia Univ

VC Theory

- VC dimension:
  - A combinatorial concept to capture the true richness of F
  - Often (but not always!) proportional to the degrees-of-freedom or the number of independent parameters in F
- Other Observations
  - VC dimension of a model class fully characterizes its learning ability!
  - Results are agnostic to the underlying distribution.

Source: Nakul Verma, Columbia Univ

ERM

- From our discussion it may seem that ERM algorithm is universally consistent. Not really though!

Theorem (no free lunch, Devroye '82):

Pick any sample size *m*, any algorithm $\mathcal{A}$ and any $\epsilon > 0$

There exists a distribution $\mathcal{D}$ such that

$$\mathrm{err}(f_m^{\mathcal{A}}) > 1/2 - \epsilon$$

while the Bayes optimal error, $\min_f \mathrm{err}(f) = 0$

Source: Nakul Verma, Columbia Univ

Further

- How to do model class selection? Structural risk results.
- Dealing with kernels – Fat margin theory
- Incorporating priors over the models – PAC-Bayes theory
- Is it possible to get distribution dependent bound? Rademacher complexity
- How about regression? Can derive similar results for nonparametric regression.

Source: Nakul Verma, Columbia Univ

Readings

- http://ciml.info/dl/v0_99/ciml-v0_99-ch12.pdf
- VC Dimension and its applications in ML
- Machine Learning by Tom Mitchell, Chapter 7