Lecture 17: PAC Learning

Madhavan Mukund

https://www.cmi.ac.in/~madhavan

Data Mining and Machine Learning
August–December 2020

Supervised learning

- Set of possible input instances X
- Categories C, say $\{0, 1\}$
- Build a classification model $M : X \to C$
- Restrict the types of models
  - Hypothesis space $\mathcal{H}$ — e.g., linear separators
  - Search for best $M \in \mathcal{H}$
- How do we find the best M?
  - Labelled training data
  - Choose M to minimize error (loss) with respect to this set
  - Why should M generalize well to arbitrary data?

No free lunch

- ML algorithms minimize training loss
- Goal is to minimize generalization loss

No Free Lunch Theorem [Wolpert, Macready 1997]

Averaged over all possible data distributions, every classification algorithm has the same error rate when classifying previously unobserved points.

- Is the situation hopeless?
- NFL theorem refers to prediction inputs coming from all possible distributions
- ML assumes training set is "representative" of overall data
  - Prediction instances follow roughly the same distribution as training set

A theoretical framework for ML

- X is the space of input instances
- $C \subseteq X$ is the target concept to be learned
  - e.g., X is all emails, C is the set of spam emails
- X is equipped with a probability distribution D
  - Any random sample from X is drawn using D
  - In particular, training set and test set are such random samples
- $\mathcal{H}$ is a set of hypotheses
  - Each $h \in \mathcal{H}$ identifies a subset of X
  - Choose the best $h \in \mathcal{H}$ as model

Training error and true error

- True error: Probability that h incorrectly classifies $x \in X$ drawn randomly according to D
  - $h \Delta C = (h \setminus C) \cup (C \setminus h)$
    - Symmetric difference, error region
  - $\mathrm{err}_D(h) = Pr_{x \sim D}(h \Delta C)$
- Training error: Given a training sample $S \subseteq X$
  - $\mathrm{err}_S(h) = |S \cap (h \Delta C)|/|S|$
- Can make $\mathrm{err}_S(h)$ arbitrarily small
  - Store and look up training data in a table — zero error
  - Poor generalization — overfitting
- Goal: minimizing $\mathrm{err}_S(h)$ should also minimize $\mathrm{err}_D(h)$

Generalization guarantees

- Overfitting Low training error but high true error
- Underfitting Cannot achieve low training/true error
- Related to the representational capacity of $\mathcal{H}$
  - How expressive is $\mathcal{H}$? How many different concepts can it capture?
  - Capacity too high — overfitting
  - Capacity too low — underfitting
- For now, assume that $\mathcal{H}$ is finite
  - Example: classify population based on age and income
  - Age and income are discrete values with lower and upper bounds
  - Assume classifier is of the form $(a_1 \leq \text{age} \leq a_2) \wedge (i_1 \leq \text{income} \leq i_2)$
    - Rectangle with corners (a_1, i_1), (a_2, i_2)
  - Only finite number of possibilities

Probably Approximately Correct (PAC) learning

- With high probability, the hypothesis h that fits the sample S also fits the concept approximately correctly

Theorem (PAC learning guarantee)

Let $\mathcal{H}$ be a hypothesis class, $\delta, \epsilon > 0$ and S a training set of size $n \geq \frac{1}{\epsilon}(\ln|\mathcal{H}| + \ln(1/\delta))$ drawn using D. With probability $\geq 1 - \delta$,

- Every $h \in \mathcal{H}$ with true error $\mathrm{err}_D > \epsilon$ has training error $\mathrm{err}_S > 0$.
- Equivalently, every $h \in \mathcal{H}$ with training error $\mathrm{err}_S = 0$ then true error $\mathrm{err}_D < \epsilon$.

- δ: probability of choosing a bad training set
- ϵ: how much error we can tolerate
- $|\mathcal{H}|$: model capacity

Probably Approximately Correct (PAC) learning

Theorem (PAC learning guarantee)

Let $\mathcal{H}$ be a hypothesis class, $\delta, \epsilon > 0$ and S a training set of size $n \geq \frac{1}{\epsilon}(\ln|\mathcal{H}| + \ln(1/\delta))$ drawn using D. With probability $\geq 1 - \delta$,

- Every $h \in \mathcal{H}$ with true error $\text{err}_D > \epsilon$ has training error $\text{err}_S > 0$.
- Equivalently, every $h \in \mathcal{H}$ with training error $\text{err}_S = 0$ then true error $\text{err}_D < \epsilon$.

Proof

- Let $h_1, h_2, \ldots, \in \mathcal{H}$ have $\text{err}_D \geq \epsilon$ but $\text{err}_S = 0$ — don't want output these
- Event A_i: h_i has $\text{err}_S = 0$ on random sample S
  - Every h_i has $\text{err}_D \geq \epsilon \Rightarrow$ probability that random input is correct is $\leq (1 - \epsilon)$
  - $|S| = n$, so $Pr(A_i) \leq (1 - \epsilon)^n$

Probably Approximately Correct (PAC) learning

Proof

- Let $h_1, h_2, \ldots, \in \mathcal{H}$ have $\text{err}_D \geq \epsilon$ but $\text{err}_S = 0$ — don't want output these
- Event A_i: h_i has $\text{err}_S = 0$ on random sample S
  - Every h_i has $\text{err}_D \geq \epsilon \Rightarrow$ probability that random input is correct is $\leq (1-\epsilon)$
  - $|S| = n$, so $Pr(A_i) \leq (1-\epsilon)^n$
- Probability that some h_i has $\text{err}_S = 0$: $Pr(\bigcup_i A_i) \leq |\mathcal{H}|(1-\epsilon)^n$ (Union Bound)
- Since $1 - \epsilon \leq e^{-\epsilon}$ (Taylor expansion of e^x), $Pr(\bigcup_i A_i) \leq |\mathcal{H}|e^{-\epsilon n}$
- We assumed $n \geq \frac{1}{\epsilon}(\ln|\mathcal{H}| + \ln(1/\delta))$, so $Pr(\bigcup_i A_i) \leq |\mathcal{H}|e^{-\ln|\mathcal{H}| - \ln(1/\delta)}$
- $|\mathcal{H}|e^{-\ln|\mathcal{H}| - \ln(1/\delta)} = |\mathcal{H}|e^{-\ln|\mathcal{H}|}e^{-\ln(1/\delta)} = |\mathcal{H}| \cdot (1/|\mathcal{H}|) \cdot \delta = \delta$
- Hence, probability that some $h \in \mathcal{H}$ has $err_D > \epsilon$ and $err_S = 0$ is $< \delta$
- Hence, with probability $\geq 1 - \delta$, every $h \in \mathcal{H}$ satisfies PAC learning guarantee

Uniform convergence

- PAC learning guarantee If h has $\text{err}_S = 0$ then h has $\text{err}_D \leq \epsilon$
- What if there is no h with $\text{err}_S = 0$
- Would like a statement like the following:

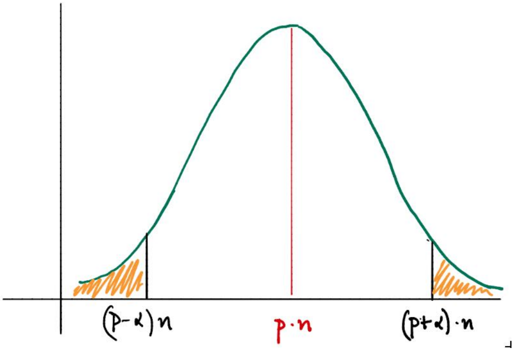
Uniform convergence

For a sufficiently large training set S, every hypothesis $h \in \mathcal{H}$ with high probability has training error within $\pm\epsilon$ of true error.

- Intuition: consider actual concept C and hypothesis h as binary strings
- Suppose C and h differ in 10% of positions (true error)
- If we take a sufficiently large subset of positions, within that subset we expect close to 10% discrepancy (training error)

Hoeffding bound

- Flip a coin n times, with $Pr(\text{heads}) = p$
- Expect to see $p \cdot n$ heads
- Let s be the actual number of heads
- What is the probability that s is far away from $p \cdot n$?



Hoeffding bound

- $Pr(s/n > p + \alpha) \leq e^{-2n\alpha^2}$
- $Pr(s/n < p - \alpha) \leq e^{-2n\alpha^2}$

Uniform convergence

Uniform convergence

Let $\mathcal{H}$ be a hypothesis class, $\delta, \epsilon > 0$. If a training set S of size $n \geq \frac{1}{2\epsilon^2}(\ln|\mathcal{H}| + \ln(2/\delta))$ is drawn using D, then with probability $\geq 1 - \delta$, every $h \in \mathcal{H}$ satisfies $|\mathrm{err}_S(h) - \mathrm{err}_D(h)| \leq \epsilon$.

Proof

- Fix $h \in \mathcal{H}$, $S = \{d_1, d_2, \ldots, d_n\}$.
- Boolean variables $\{x_1, x_2, \ldots, x_n\}$: $x_j = 1$ iff h makes a mistake on d_j.
- Actual training error $\mathrm{err}_S(h)$ is $\frac{\sum_{j=1}^{n} x_j}{n}$.
- Expected value of training error is $n \cdot \mathrm{err}_D(h)$.

Uniform convergence

- Fix $h \in \mathcal{H}$, $S = \{d_1, d_2, \ldots, d_n\}$, $x_j = 1$ iff h makes a mistake on d_j.
- Actual training error $\text{err}_S(h)$ is $\dfrac{\sum_{j=1}^{n} x_j}{n}$, expected value is $n \cdot \text{err}_D(h)$.
- Let A_h be the event that h is a bad hypothesis: $|\text{err}_S(h) - \text{err}_D(h)| > \epsilon$
- By Hoeffding bounds:
  - $Pr(\text{err}_S(h) > \text{err}_D(h) + \epsilon) < e^{-n\epsilon^2}$
  - $Pr(\text{err}_S(h) < \text{err}_D(h) - \epsilon) < e^{-n\epsilon^2}$
  - $Pr(A_h) = Pr(|\text{err}_S(h) - \text{err}_D(h)| > \epsilon) < 2e^{-n\epsilon^2}$
- Probability that some h is bad: $Pr(\bigcup_h A_h) \leq |\mathcal{H}| \cdot 2e^{-n\epsilon^2}$ (Union Bound)
- Substitute $n \geq \dfrac{1}{2\epsilon^2}(\ln|\mathcal{H}| + \ln(2/\delta))$ to get $Pr(\bigcup_h A_h) \leq \delta$.

Models with bounded description length

- Assume model is described using at most b bits
- $|\mathcal{H}| \leq 2^b$, so $\ln|\mathcal{H}| \leq b \ln 2$
- Applying PAC learning guarantee:
  - With probability at least $1-\delta$, any model with $\text{err}_S(h) = 0$ will have $\text{err}_D(h) < \dfrac{b\ln 2 + \ln(1/\delta)}{|S|}$
- Decision trees: k nodes, d columns/features
  - $\log d$ bits to write down question for each node
  - $k \log d$ bits for the whole tree
  - If $n \geq \dfrac{1}{\epsilon}(\ln(2^{k\log d}) + \ln(1/\delta))$, PAC learning guarantee holds
  - Solve for k, $k \leq (n\epsilon - \ln(1/\delta))/\log d$
  - If we find a small tree of size k with zero training error, it will generalize well

Summary

- How do we justify that a model optimized for training data generalizes well?
- PAC learning guarantee — training set size determined by parameters δ, ϵ, $|\mathcal{H}|$
- Extend to uniform convergence
- Apply to get bounds for models with bounded description length
- How do we compute representational capacity if $\mathcal{H}$ is infinite?