COL866: Foundations of Data Science

Ragesh Jaiswal, IITD

Machine Learning: Generalization

Theorem

Let $\mathcal{H}$ be a hypothesis class and let $\varepsilon, \delta > 0$. If a training set S of size

$$n \geq \frac{1}{\varepsilon}(\ln|\mathcal{H}| + \ln 1/\delta),$$

is drawn from distribution D, then with probability at least $(1-\delta)$ every $h \in \mathcal{H}$ with true error $err_D(h) \geq \varepsilon$ has training error $err_S(h) > 0$. Equivalently, with probability at least $(1-\delta)$, every $h \in \mathcal{H}$ with training error 0 has true error at most ε.

Theorem

Let $\mathcal{H}$ be a hypothesis class and let $\varepsilon, \delta > 0$. If a training set S of size

$$n \geq \frac{1}{\varepsilon}(\ln |\mathcal{H}| + \ln 1/\delta),$$

is drawn from distribution D, then with probability at least $(1-\delta)$ every $h \in \mathcal{H}$ with true error $err_D(h) \geq \varepsilon$ has training error $err_S(h) > 0$. Equivalently, with probability at least $(1-\delta)$, every $h \in \mathcal{H}$ with training error 0 has true error at most ε.

- The above result is called the PAC-learning guarantee since it states that if we can find an $h \in \mathcal{H}$ consistent with the sample, then this h is *Probably Approximately Correct*.
- What if we manage to find a hypothesis with small disagreement on the sample? Can we say that the hypothesis will have small true error?

Theorem

Let $\mathcal{H}$ be a hypothesis class and let $\varepsilon, \delta > 0$. If a training set S of size

$$n \geq \frac{1}{\varepsilon}(\ln |\mathcal{H}| + \ln (1/\delta)),$$

is drawn from distribution D, then with probability at least $(1 - \delta)$ every $h \in \mathcal{H}$ with true error $err_D(h) \geq \varepsilon$ has training error $err_S(h) > 0$. Equivalently, with probability at least $(1 - \delta)$, every $h \in \mathcal{H}$ with training error 0 has true error at most ε.

Theorem (Uniform convergence)

Let $\mathcal{H}$ be a hypothesis class and let $\varepsilon, \delta > 0$. If a training set S of size

$$n \geq \frac{1}{2\varepsilon^2}(\ln |\mathcal{H}| + \ln (2/\delta)),$$

is drawn from distribution D, then with probability at least $(1 - \delta)$ every $h \in \mathcal{H}$ satisfies $|err_D(h) - err_S(h)| \leq \varepsilon$.

Theorem

Let $\mathcal{H}$ be a hypothesis class and let $\varepsilon, \delta > 0$. If a training set S of size

$$n \geq \frac{1}{\varepsilon}(\ln |\mathcal{H}| + \ln (1/\delta)),$$

is drawn from distribution D, then with probability at least $(1 - \delta)$ every $h \in \mathcal{H}$ with true error $err_D(h) \geq \varepsilon$ has training error $err_S(h) > 0$. Equivalently, with probability at least $(1 - \delta)$, every $h \in \mathcal{H}$ with training error 0 has true error at most ε.

Theorem (Uniform convergence)

Let $\mathcal{H}$ be a hypothesis class and let $\varepsilon, \delta > 0$. If a training set S of size

$$n \geq \frac{1}{2\varepsilon^2}(\ln |\mathcal{H}| + \ln (2/\delta)),$$

is drawn from distribution D, then with probability at least $(1 - \delta)$ every $h \in \mathcal{H}$ satisfies $|err_D(h) - err_S(h)| \leq \varepsilon$.

- The above theorem essentially means that conditioned on S being sufficiently large, good performance on S will translate to good performance on D.

Theorem (Uniform convergence)

Let $\mathcal{H}$ be a hypothesis class and let $\varepsilon, \delta > 0$. If a training set S of size

$$n \geq \frac{1}{2\varepsilon^2}(\ln|\mathcal{H}| + \ln(2/\delta)),$$

is drawn from distribution D, then with probability at least $(1-\delta)$ every $h \in \mathcal{H}$ satisfies $|err_D(h) - err_S(h)| \leq \varepsilon$.

- The above theorem follows from the following tail inequality.

Theorem (Chernoff-Hoeffding bound)

Let $x_1, ..., x_n$ be independent $\{0, 1\}$ random variables such that $\forall i, \mathbf{Pr}[x_i = 1] = p$. Let $s = \sum_{i=1}^{n} x_i$. For any $0 \leq \alpha \leq 1$,

$$\mathbf{Pr}[s/n > p + \alpha] \leq e^{-2n\alpha^2} \quad and \quad \mathbf{Pr}[s/n < p - \alpha] \leq e^{-2n\alpha^2}.$$

- Let us do a case study of *Learning Disjunctions*.
- Consider a binary classification context where the instance space $\mathcal{X} = \{0,1\}^d$.
- Suppose we believe that the target concept is a disjunction over a subset of features. For example, $c^\star = \{x : x_1 \vee x_{10} \vee x_{50}\}$.
- What is the size of the concept class $\mathcal{H}$?

- Let us do a case study of *Learning Disjunctions*.
- Consider a binary classification context where the instance space $\mathcal{X} = \{0, 1\}^d$.
- Suppose we believe that the target concept is a disjunction over a subset of features. For example, $c^\star = \{x : x_1 \lor x_{10} \lor x_{50}\}$.
- What is the size of the concept class $\mathcal{H}$? $|\mathcal{H}| = 2^d$
- So, if the sample size $|S| = \frac{1}{\varepsilon}(d \ln 2 + \ln(1/\delta))$ then good performance on the training set generalizes to the instance space.
- Question: Suppose the target concept is indeed a disjunction, then given any training set S is there an algorithm that can at least output a disjunction consistent with S.

- Occam's razor: William of Occam around 1320AD stated that one should prefer simpler explanations over more complicated ones.

- Occam's razor: William of Occam around 1320AD stated that one should prefer simpler explanations over more complicated ones.
- What do we mean by a rule being simple?
- Different people may have different description languages for describing rules.
- How many rules can be described using fewer than b bits?

- Occam's razor: William of Occam around 1320AD stated that one should prefer simpler explanations over more complicated ones.
- What do we mean by a rule being simple?
- Different people may have different description languages for describing rules.
- How many rules can be described using fewer than b bits? $< 2^b$

Theorem (Occam's razor)

Fix any description language, and consider a training sample S drawn from distribution D. With probability at least $(1-\delta)$ any rule h consistent with S that can be described in this language using fewer than b bits will have $err_D(h) \leq \varepsilon$ for $|S| = \frac{1}{\varepsilon}(b \ln 2 + \ln(1/\delta))$. Equivalently, with probability at least $(1-\delta)$ all rules that can be described in fewer than b bits will have $err_D(h) \leq \frac{b \ln(2) + \ln(1/\delta)}{|S|}$.

Theorem (Occam's razor)

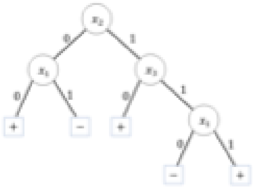
Fix any description language, and consider a training sample S drawn from distribution D. With probability at least $(1 - \delta)$ any rule h consistent with S that can be described in this language using fewer than b bits will have $err_D(h) \leq \varepsilon$ for $|S| = \frac{1}{\varepsilon}(b \ln 2 + \ln(1/\delta))$. Equivalently, with probability at least $(1 - \delta)$ all rules that can be described in fewer than b bits will have $err_D(h) \leq \frac{b \ln(2) + \ln(1/\delta)}{|S|}$.

- The theorem is valid irrespective of the description language.
- It does not say that complicated rules are bad.
- It suggests that Occam's rule is a good policy since simple rules are unlikely to fool us since there are not too many of them.

Theorem (Occam's razor)

Fix any description language, and consider a training sample S drawn from distribution D. With probability at least $(1 - \delta)$ any rule h consistent with S that can be described in this language using fewer than b bits will have $err_D(h) \leq \varepsilon$ for $|S| = \frac{1}{\varepsilon}(b \ln 2 + \ln(1/\delta))$. Equivalently, with probability at least $(1 - \delta)$ all rules that can be described in fewer than b bits will have $err_D(h) \leq \frac{b \ln(2) + \ln(1/\delta)}{|S|}$.

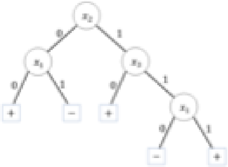
- Case study: Decision trees



Theorem (Occam's razor)

Fix any description language, and consider a training sample S drawn from distribution D. With probability at least $(1-\delta)$ any rule h consistent with S that can be described in this language using fewer than b bits will have $err_D(h) \leq \varepsilon$ for $|S| = \frac{1}{\varepsilon}(b \ln 2 + \ln(1/\delta))$. Equivalently, with probability at least $(1-\delta)$ all rules that can be described in fewer than b bits will have $err_D(h) \leq \frac{b \ln(2) + \ln(1/\delta)}{|S|}$.

- Case study: Decision trees
- What is the bit-complexity of describing a decision tree (in d variables) of size k?

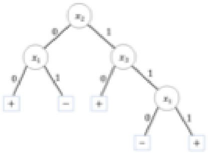


Machine Learning
Generalization bounds

Theorem (Occam's razor)

Fix any description language, and consider a training sample S drawn from distribution D. With probability at least $(1-\delta)$ any rule h consistent with S that can be described in this language using fewer than b bits will have $err_D(h) \leq \varepsilon$ for $|S| = \frac{1}{\varepsilon}(b \ln 2 + \ln(1/\delta))$. Equivalently, with probability at least $(1-\delta)$ all rules that can be described in fewer than b bits will have $err_D(h) \leq \frac{b \ln(2) + \ln(1/\delta)}{|S|}$.

- Case study: Decision trees
- What is the bit-complexity of describing a decision tree (in d variables) of size k? $O(k \log d)$
- So, the true error is low if we can produce a consistent tree with fewer than $\frac{\varepsilon |S|}{\log d}$ nodes.



- We have seen that for good generalization, the size of the training set should depend on $\log_2(\mathcal{H})$ that in some sense captures the complexity of the hypothesis class.
- Let us try to understand this using a simple example. Consider the age-versus-salary data.
  - There are 100 possible ages and 1000 different salaries. This makes the instance space $\mathcal{X}$ of size 10^5.
  - The hypothesis class consists of axis-parallel rectangles. What is the size of $\mathcal{H}$?

- We have seen that for good generalization, the size of the training set should depend on $\log_2(\mathcal{H})$ that in some sense captures the complexity of the hypothesis class.
- Let us try to understand this using a simple example. Consider the age-versus-salary data.
  - There are 100 possible ages and 1000 different salaries. This makes the instance space $\mathcal{X}$ of size 10^5.
  - The hypothesis class consists of axis-parallel rectangles. What is the size of $\mathcal{H}$? $|\mathcal{H}| = 10^{10}$
  - Suppose there are only $N = 100$ employed people for which we know the data. Then for the purpose of generalization, we may use $|\mathcal{H}| \leq N^4$.
- Question: Is there is a tighter measure of complexity of a hypothesis class with respect to generalization?
  - Independent of the size of the support of the distribution D.

- Question: Is there is a tighter measure of complexity of a hypothesis class with respect to generalization?
  - Independent of the size of the support of the distribution D.

Definition (Shattering)

Given a set S of examples and a concept class $\mathcal{H}$, we say that S is shattered by $\mathcal{H}$ if for every $A \subseteq S$ there exists some $h \in \mathcal{H}$ that labels all examples in A as positive and all examples in $S \setminus A$ as negative.

Definition (VC Dimension)

The VC-dimension of $\mathcal{H}$ is the size of the largest set shattered by $\mathcal{H}$.

Definition (Shattering)

Given a set S of examples and a concept class $\mathcal{H}$, we say that S is shattered by $\mathcal{H}$ if for every $A \subseteq S$ there exists some $h \in \mathcal{H}$ that labels all examples in A as positive and all examples in $S \setminus A$ as negative.

Definition (VC Dimension)

The VC-dimension of $\mathcal{H}$ is the size of the largest set shattered by $\mathcal{H}$.

- Example: Consider the hypothesis class $\mathcal{H}$ of axis-parallel rectangles.
- Question: What is the VC-dimension of $\mathcal{H}$?
  - Question: Does there exist a set of 4 points that $\mathcal{H}$ can shatter?

Definition (Shattering)

Given a set S of examples and a concept class $\mathcal{H}$, we say that S is shattered by $\mathcal{H}$ if for every $A \subseteq S$ there exists some $h \in \mathcal{H}$ that labels all examples in A as positive and all examples in $S \setminus A$ as negative.

Definition (VC Dimension)

The VC-dimension of $\mathcal{H}$ is the size of the largest set shattered by $\mathcal{H}$.

- Example: Consider the hypothesis class $\mathcal{H}$ of axis-parallel rectangles.
- Question: What is the VC-dimension of $\mathcal{H}$?
    - Question: Does there exist a set of 4 points that $\mathcal{H}$ can shatter?
      Yes
    - Question: Does there exist a set of 5 points that $\mathcal{H}$ can shatter?

Definition (Shattering)

Given a set S of examples and a concept class $\mathcal{H}$, we say that S is shattered by $\mathcal{H}$ if for every $A \subseteq S$ there exists some $h \in \mathcal{H}$ that labels all examples in A as positive and all examples in $S \setminus A$ as negative.

Definition (VC Dimension)

The VC-dimension of $\mathcal{H}$ is the size of the largest set shattered by $\mathcal{H}$.

- Example: Consider the hypothesis class $\mathcal{H}$ of axis-parallel rectangles.
- Question: What is the VC-dimension of $\mathcal{H}$? $\textit{VC-dim}(\mathcal{H}) = 4$
  - Question: Does there exist a set of 4 points that $\mathcal{H}$ can shatter? Yes
  - Question: Does there exist a set of 5 points that $\mathcal{H}$ can shatter? No

Definition (Shattering)

Given a set S of examples and a concept class $\mathcal{H}$, we say that S is shattered by $\mathcal{H}$ if for every $A \subseteq S$ there exists some $h \in \mathcal{H}$ that labels all examples in A as positive and all examples in $S \setminus A$ as negative.

Definition (VC Dimension)

The VC-dimension of $\mathcal{H}$ is the size of the largest set shattered by $\mathcal{H}$.

Definition (Growth function)

Given a set S of examples and a concept class $\mathcal{H}$, let $\mathcal{H}[S] = \{h \cap S : h \in \mathcal{H}\}$. That is, $\mathcal{H}[S]$ is the concept class $\mathcal{H}$ restricted to the set of points S. For integer n and class $\mathcal{H}$, let $\mathcal{H}[n] = \max_{|S|=n} |\mathcal{H}[S]|$; this is called the growth function of $\mathcal{H}$.

Definition (Shattering)

Given a set S of examples and a concept class $\mathcal{H}$, we say that S is shattered by $\mathcal{H}$ if for every $A \subseteq S$ there exists some $h \in \mathcal{H}$ that labels all examples in A as positive and all examples in $S \setminus A$ as negative.

Definition (VC Dimension)

The VC-dimension of $\mathcal{H}$ is the size of the largest set shattered by $\mathcal{H}$.

Definition (Growth function)

Given a set S of examples and a concept class $\mathcal{H}$, let $\mathcal{H}[S] = \{h \cap S : h \in \mathcal{H}\}$. That is, $\mathcal{H}[S]$ is the concept class $\mathcal{H}$ restricted to the set of points S. For integer n and class $\mathcal{H}$, let $\mathcal{H}[n] = \max_{|S|=n} |\mathcal{H}[S]|$; this is called the growth function of $\mathcal{H}$.

- The growth function of a class is also called shatter function or shatter coefficient.

Definition (Shattering)

Given a set S of examples and a concept class $\mathcal{H}$, we say that S is shattered by $\mathcal{H}$ if for every $A \subseteq S$ there exists some $h \in \mathcal{H}$ that labels all examples in A as positive and all examples in $S \setminus A$ as negative.

Definition (VC Dimension)

The VC-dimension of $\mathcal{H}$ is the size of the largest set shattered by $\mathcal{H}$.

Definition (Growth function)

Given a set S of examples and a concept class $\mathcal{H}$, let $\mathcal{H}[S] = \{h \cap S : h \in \mathcal{H}\}$. That is, $\mathcal{H}[S]$ is the concept class $\mathcal{H}$ restricted to the set of points S. For integer n and class $\mathcal{H}$, let $\mathcal{H}[n] = \max_{|S|=n} |\mathcal{H}[S]|$; this is called the growth function of $\mathcal{H}$.

- Fill in the blanks:
  - S is shattered by $\mathcal{H}$ iff $|\mathcal{H}[S]| =$ ____?
  - The VC-dimension of $\mathcal{H}$ is the largest n such that $\mathcal{H}[n] =$ ____?
  - For the case of axis-parallel rectangles, $\mathcal{H}[n] =$ ____?
  - For linear separators in 2 dimensions, $VCdim(\mathcal{H}) =$ ____?
  - For linear separators in 2 dimensions, $\mathcal{H}[n] =$ ____?
  - For any $\mathcal{H}$, $VCdim(\mathcal{H}) \leq$ ____?

Definition (Shattering)

Given a set S of examples and a concept class $\mathcal{H}$, we say that S is shattered by $\mathcal{H}$ if for every $A \subseteq S$ there exists some $h \in \mathcal{H}$ that labels all examples in A as positive and all examples in $S \setminus A$ as negative.

Definition (VC Dimension)

The VC-dimension of $\mathcal{H}$ is the size of the largest set shattered by $\mathcal{H}$.

Definition (Growth function)

Given a set S of examples and a concept class $\mathcal{H}$, let $\mathcal{H}[S] = \{h \cap S : h \in \mathcal{H}\}$. That is, $\mathcal{H}[S]$ is the concept class $\mathcal{H}$ restricted to the set of points S. For integer n and class $\mathcal{H}$, let $\mathcal{H}[n] = \max_{|S|=n} |\mathcal{H}[S]|$; this is called the growth function of $\mathcal{H}$.

- The growth function of a class is also called shatter function or shatter coefficient.
- Fill in the blanks:
  - S is shattered by $\mathcal{H}$ iff $|\mathcal{H}[S]| = \underline{2^{|S|}}$.
  - The VC-dimension of $\mathcal{H}$ is the largest n such that $\mathcal{H}[n] = \underline{2^n}$.
  - For the case of axis-parallel rectangles, $\mathcal{H}[n] = \underline{O(n^4)}$.
  - For linear separators in 2 dimensions, $VCdim(\mathcal{H}) = \underline{3}$.
  - For linear separators in 2 dimensions, $\mathcal{H}[n] = \underline{O(n^2)}$.
  - For any $\mathcal{H}$, $VCdim(\mathcal{H}) \leq \underline{\log_2(|\mathcal{H}|)}$.

Definition (Shattering)

Given a set S of examples and a concept class $\mathcal{H}$, we say that S is shattered by $\mathcal{H}$ if for every $A \subseteq S$ there exists some $h \in \mathcal{H}$ that labels all examples in A as positive and all examples in $S \setminus A$ as negative.

Definition (VC Dimension)

The VC-dimension of $\mathcal{H}$ is the size of the largest set shattered by $\mathcal{H}$.

Definition (Growth function)

Given a set S of examples and a concept class $\mathcal{H}$, let $\mathcal{H}[S] = \{h \cap S : h \in \mathcal{H}\}$. That is, $\mathcal{H}[S]$ is the concept class $\mathcal{H}$ restricted to the set of points S. For integer n and class $\mathcal{H}$, let $\mathcal{H}[n] = \max_{|S|=n} |\mathcal{H}[S]|$; this is called the growth function of $\mathcal{H}$.

- We can now discuss generalization bounds just in terms of growth function and VC dimension (instead of in terms of $|\mathcal{H}|$).

Theorem

For any hypothesis class $\mathcal{H}$ and distribution D, if a training sample S is drawn from D of size

$$n \geq \frac{2}{\varepsilon}\left[\log_2\left(2\mathcal{H}[2n]\right) + \log_2\left(1/\delta\right)\right].$$

then with probability at least $(1-\delta)$, every $h \in \mathcal{H}$ with error $err_D(h) \geq \varepsilon$ has $err_S(h) > 0$. Equivalently, every $h \in \mathcal{H}$ with $err_S(h) = 0$ has $err_D(h) < \varepsilon$.

Theorem

For any hypothesis class $\mathcal{H}$ and distribution D, if a training sample S is drawn from D of size

$$n \geq \frac{8}{\varepsilon^2}\left[\log_2\left(2\mathcal{H}[2n]\right) + \log_2\left(2/\delta\right)\right].$$

then with probability at least $(1-\delta)$, every $h \in \mathcal{H}$ will have $|err_D(h) - err_S(h)| \leq \varepsilon$.

Theorem

For any hypothesis class $\mathcal{H}$ and distribution D, if a training sample S is drawn from D of size $n \geq \frac{2}{\varepsilon}\left[\log_2(2\mathcal{H}[2n]) + \log_2(1/\delta)\right]$. then with probability at least $(1-\delta)$, every $h \in \mathcal{H}$ with error $err_D(h) \geq \varepsilon$ has $err_S(h) > 0$. Equivalently, every $h \in \mathcal{H}$ with $err_S(h) = 0$ has $err_D(h) < \varepsilon$.

Theorem

For any hypothesis class $\mathcal{H}$ and distribution D, if a training sample S is drawn from D of size $n \geq \frac{8}{\varepsilon^2}\left[\log_2(2\mathcal{H}[2n]) + \log_2(2/\delta)\right]$. then with probability at least $(1-\delta)$, every $h \in \mathcal{H}$ will have $|err_D(h) - err_S(h)| \leq \varepsilon$.

Theorem (Sauer's Lemma)

If $VCdim(\mathcal{H}) = d$, then $\mathcal{H}[n] \leq \sum_{i=0}^{d} \binom{n}{i} \leq \left(\frac{en}{d}\right)^d$.

Theorem

For any hypothesis class $\mathcal{H}$ and distribution D, a training sample S of size

$$O\left(\frac{1}{\varepsilon}\left[VCdim(\mathcal{H})\log(1/\varepsilon) + \log 1/\delta\right]\right)$$

is sufficient to ensure that with probability at least $(1-\delta)$, every $h \in \mathcal{H}$ with $err_D(h) \geq \varepsilon$ has $err_S(h) > 0$.

End