CSCI B609: “Foundations of Data Science”

Lecture 11/12: Introduction to Machine Learning Continued

Slides at http://grigory.us/data-science-class.html

Grigory Yaroslavtsev

http://grigory.us

Intro to ML

- Classification problem
  - Instance space X: $\{0,1\}^{\boldsymbol{d}}$ or $\mathbb{R}^{\boldsymbol{d}}$ (feature vectors)
  - Classification: come up with a mapping $X \to \{0,1\}$
- Formalization:
  - Assume there is a probability distribution D over X
  - $\boldsymbol{c}^*$= "target concept" (set $\boldsymbol{c}^* \subseteq X$ of positive instances)
  - Given labeled i.i.d. samples from D produce $\boldsymbol{h} \subseteq X$
  - **Goal:** have $\boldsymbol{h}$ agree with $\boldsymbol{c}^*$ over distribution D
  - Minimize: $err_D(\boldsymbol{h}) = \Pr_D[\boldsymbol{h} \,\Delta\, \boldsymbol{c}^*]$
  - $err_D(\boldsymbol{h})$ = "true" or "generalization" error

Intro to ML

- Training error
  - S = labeled sampled (pairs $(x, l), x \in X, l \in \{0,1\}$)
  - Training error: $err_S(\boldsymbol{h}) = \frac{|S \cap (\boldsymbol{h} \,\Delta\, \boldsymbol{c}^*)|}{|S|}$
- "Overfitting": low training error, high true error
- Hypothesis classes:
  - H: collection of subsets of X called hypotheses
    - If $X = \mathbb{R}$ could be all intervals $\{[a, b], a \leq b\}$
    - If $X = \mathbb{R}^d$ could be linear separators:
      $\left\{\{\boldsymbol{x} \in \mathbb{R}^d \middle| \boldsymbol{w} \cdot \boldsymbol{x} \geq w_0\} | \boldsymbol{w} \in \mathbb{R}^d, w_0 \in \mathbb{R}\right\}$
- If S is large enough (compared to some property of H) then overfitting doesn't occur

Overfitting and Uniform Convergence

- **PAC learning (agnostic)**: For $\epsilon, \delta > 0$ if

$$|S| \geq 1/2\epsilon^2(\ln|H| + \ln 2/\delta)$$

then with probability $1 - \delta$:

$$\forall \boldsymbol{h} \in \mathrm{H}: |err_S(\boldsymbol{h}) - err_D(\boldsymbol{h})| \leq \epsilon$$

- Size of the class of hypotheses can be very large
- Can also be infinite, how to give a bound then?
- We will see ways around this today

VC-dimension

- VC-dim$(H) \leq \ln|H|$
- Consider database age vs. salary
- Query: fraction of the overall population with ages 35—45 and salary \$$(50 - 70)$K
- How big a database can answer with $\pm\epsilon$ error
- 100 ages $\times$ 1000 salaries $\Rightarrow 10^{10}$ rectangles
- $1/2\epsilon^2(10 \ln 10 + \ln 2/\delta)$ samples suffice
- What if we don't want to discretize?

VC-dimension

- **Def.** Concept class H **shatters** a set S if $\forall A \subseteq S$ there is $\mathrm{h} \in H$ labeling A positive and $\mathrm{A} \setminus S$ negative
- **Def. VC-dim**(H) = size of the largest shattered set
- Example: axis-parallel rectangles on the plane
  – 4-point diamond is shattered
  – No 5-point set can be shattered
  – VC-dim(axis-parallel rectangles) = 4
- **Def.** $H[S] = \{h \cap S : h \in H\} =$ set of labelings of the points in S by functions in H
- **Def. Growth function** $H(n) = \max_{|S|=n} |H[S]|$
- Example: growth function of a-p. rectangles is $O(n^4)$

Growth function & uniform convergence

- **PAC learning via growth function**: For $\epsilon, \delta > 0$ if

$$|S| = n \geq 8/\epsilon^2 (\ln|2H(2n)| + \ln 1/\delta)$$

then with probability $1 - \delta$:

$$\forall \boldsymbol{h} \in \mathrm{H}: |err_S(\boldsymbol{h}) - err_D(\boldsymbol{h})| \leq \epsilon$$

- **Thm (Sauer's lemma).** If VC-dim(H)$=d$ then:

$$H(n) \leq \sum_{i=0}^{d} \binom{n}{i} \leq \left(\frac{en}{d}\right)^d$$

- For half-planes, VC-dim $= 3$, $H(n) = O(n^2)$

Sauer's Lemma Proof

- Let $d = VC$-dim(H) we'll show that if $|S| = n$:

$$|H[S]| \leq \binom{n}{\leq d} = \sum_{i=0}^{d} \binom{n}{i}$$

- $\binom{n}{\leq d} = \binom{n-1}{\leq d} + \binom{n-1}{\leq d-1}$

Proof (induction by set size):

- $S \setminus \{x\}$: by induction $|H[S \setminus \{x\}]| \leq \binom{n-1}{\leq d}$
- $|H[S]| - |H[S \setminus \{x\}]| \leq \binom{n-1}{\leq d-1}$?

$$|H[S]| - |H[S \setminus \{x\}]| \le \binom{n-1}{\le d-1}$$

- If $H[S] > H[S \setminus \{x\}]$ then it is because of the sets that differ only on x so let's pair them up
- For $h \in H[S]$ containing x let $\boldsymbol{twin(h)} = h \setminus \{x\}$
  $T = \{h \in H[S]: x \in h \;\; and \; \boldsymbol{twin(h)} \in H[S]\}$
- Note: $|H[S]| - |H[S \setminus \{x\}]| = |T|$
- What is the VC-dimension of T?
  - If VC-dim$(T) = d'$ then $\boldsymbol{R} \subseteq S \setminus \{x\}$ of d' is shattered
  - All $2^{d'}$ subsets of $\boldsymbol{R}$ are 0/1 extendable on x
  - $d \ge d' + 1 \Rightarrow$ VC-dim$(T) \le d \; - 1 \Rightarrow$ apply induction

Examples

- Intervals of the reals:
  - Shatter 2 points, don't shatter 3 ⇒ VC-dim $= 2$
- Pairs of intervals of the reals:
  - Shatter 4 points, don't shatter 5 ⇒ VC-dim $= 4$
- Convex polygons
  - Shatter any n points on a circle ⇒ VC-dim $= \infty$
- Linear separators in d dimensions:
  - Shatter $d + 1$ points (unit vectors + origin)
  - Take subset S and set $w_i = 0$ if $i \in S$:

    separator $w^T x \leq 0$

VC-dimension of linear separators

No set of $d + 2$ points can be shattered

- **Thm (Radon).** Any set $S \subseteq \mathbb{R}^d$ with $|S| = d + 2$

can be partitioned into two subsets A, B s.t.:

$$\text{Convex}(A) \cap \text{Convex}(B) \neq \emptyset$$

- Form $d \times (d + 2)$ matrix A, columns = points in S
- Add extra all-1 row $\Rightarrow$ matrix B
- $\boldsymbol{x} = (x_1, x_2, \dots, x_{d+2})$, non-zero vector: $Bx = 0$
- Reordering: $x_1, x_2, \dots, x_s \geq 0, x_{s+1}, \dots, x_{d+2} < 0$
- Normalize: $\sum_{i=1}^{s} |x_i| = 1$

Randon's Theorem (cont.)

- $\boldsymbol{b_i}, \boldsymbol{a_i} =$ i-th columns of B and A
- $\sum_{i=1}^{s} |x_i|\, \boldsymbol{b_i} = \sum_{i=s+1}^{d+2} |x_i|\, \boldsymbol{b_i}$
  - $\sum_{i=1}^{s} |x_i|\, \boldsymbol{a_i} = \sum_{i=s+1}^{d+2} |x_i|\, \boldsymbol{a_i}$
  - $\sum_{i=1}^{s} |x_i| = \sum_{i=s+1}^{d+2} |x_i| = 1$
- Convex combinations of two subsets intersect