Computational and Statistical Learning Theory

TTIC 31120

Prof. Nati Srebro

Lecture 2:

PAC Learning and VC Theory I

From Adversarial Online to Statistical

- Three reasons to move from worst-case deterministic to stochastic:
  - Deal with errors. What if data not *exactly* realized by $\mathcal{H}$?
  - Avoid non-learnability due to very specific, adversarial, order of examples
  - Training on dedicated training set, then predict on "population"

The Statistical Learning Model

- **Unknown source distribution $\mathcal{D}$ over $(\boldsymbol{x}, \boldsymbol{y})$**
  - Describes "reality". What we want to classify, and what should it be classified as.
  - E.g. joint distribution over ( , b)
  - Can think of $\mathcal{D}$ as: distribution over x and $y|x = f(x)$
    - Distribution over images we expect to see (we don't expect to see uniformly distributed images: ), and what character each image represents
  - Or, as: distribution over y and over $x|y$
    - Distribution over characters ('e' more likely then '&'), and for each character, over possible images of that character

- **Goal: find predictor $\boldsymbol{h}$ with small expected error:**

$$L_{\mathcal{D}}(h) = \mathbb{P}_{(x,y)\sim\mathcal{D}}[h(x) \neq y]$$

(also called *generalization error*, *risk* or *true error)*

- **Based on a sample $\boldsymbol{S} = ((\boldsymbol{x_1}, \boldsymbol{y_1}), (\boldsymbol{x_2}, \boldsymbol{y_2}), \ldots, (\boldsymbol{x_m}, \boldsymbol{y_m}))$** of m training points $(x_t, y_t) \sim \mathrm{i.i.d.}\, \mathcal{D}$ (we can also write: $S \sim \mathcal{D}^{\mathrm{m}}$)

The Statistical Learning Model

- Unknown source distribution $\mathcal{D}$ over (x, y)
- Goal: find predictor h with small expected error:

$$L_{\mathcal{D}}(h) = \mathbb{P}_{(x,y)\sim\mathcal{D}}[h(x) \neq y]$$

- Based on sample $S = ((x_1, y_1), (x_2, y_2), \dots, (x_m, y_m))$ of m training points

$$(x_t, y_t) \sim \text{i.i.d.}\, \mathcal{D} \qquad (\text{i.e. } S \sim \mathcal{D}^m)$$

- **Statistical (batch) learning:**
  1. Receive training set $S \sim \mathcal{D}^m$
  2. Learn $h = A(S)$ using learning rule $A: (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{Y}^{\mathcal{X}}$
  3. Use h on future examples drawn form $\mathcal{D}$, suffering expected error $L_{\mathcal{D}}(h)$

- **Main assumption:**
  - i.i.d. samples
  - Samples drawn from distribution $\mathcal{D}$ we will later use the predictor on

Expected vs Empirical Error

- What we care about is the **expected error**

$$L_{\mathcal{D}}(h) = \mathbb{P}_{(x,y)\sim\mathcal{D}}[h(x) \neq y]$$

- Why not just minimize it directly?

- For a given sample S we can calculate the **empirical error**

$$L_S(h) = \frac{1}{m}\sum_{t=1}^{m}[[\, h(x_t) \neq y_t]]$$

- How do we use the empirical error?
- Is it a good estimate for the expected error?
- How good?

The Empirical Error as an Estimator for the Expected Error

- How close are the expected and empirical errors?

$$|L_S(h) - L_{\mathcal{D}}(h)|$$

Random Variable

Number: $L_{\mathcal{D}}(h) = \mathbb{P}_{(x,y)\sim\mathcal{D}}[h(x) \neq y]$

$$L_S(h) = \frac{1}{m}\sum_{t=1}^{m}\left[[h(x_t) \neq y_t]\right] \sim \frac{1}{m} Binom(m, L_{\mathcal{D}}(h))$$

$$\approx \mathcal{N}\left(L_{\mathcal{D}}(h), \sqrt{\frac{L_{\mathcal{D}}(h)\left(1-L_{\mathcal{D}}(h)\right)}{\mathrm{m}}}\right)$$

- Hoeffding Bound on trail of Binomial:

$$\mathbb{P}_{Z\sim Binom(m,p)}[|Z - \mathbb{E}[Z]| > t] \leq 2e^{-t^2/m}$$

- Conclusion: with probability $\geq 1 - \delta$,

$$|L_{\mathcal{D}}(h) - L_S(h)| \leq \sqrt{\frac{\log {}^{2}/_{\delta}}{2m}}$$

Empirical Risk Minimization

$$ERM_{\mathcal{H}}(S) = \hat{h} = \arg\min_{h \in \mathcal{H}} L_S(h)$$

- Can we conclude that w.p. $\geq 1 - \delta$,

$$\left| L_{\mathcal{D}}(\hat{\boldsymbol{h}}) - L_S(\hat{h}) \right| \leq \sqrt{\frac{\log {}^{2}/_{\delta}}{2m}} \quad ?$$

Uniform Convergence and the Union Bound

- For each h we have:

$$\mathbb{P}_S(|L_S(h) - L_{\mathcal{D}}(h)| \geq t) \leq 2e^{-t^2/m}$$

- And so:

$$\mathbb{P}_S(\exists_{h\in\mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| \geq t) \leq \sum_{h\in\mathcal{H}} \mathbb{P}_S(|L_S(h) - L_{\mathcal{D}}(h)| \geq t)$$
$$\leq |\mathcal{H}| \cdot 2e^{-t^2/m}$$

- **Theorem**: For any hypothesis class $\mathcal{H}$ and any $\mathcal{D}$,

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[\forall_{h\in\mathcal{H}}, |L_{\mathcal{D}}(h) - L_S(h)| \leq \sqrt{\frac{\log|\mathcal{H}| + \log {}^{2}/_{\delta}}{2m}}\right] \geq 1 - \delta$$

- Another way to view the derivation:

$$\mathbb{P}_S\left[|L_S(h) - L_{\mathcal{D}}(h)| \geq \sqrt{\frac{\log^2/_{\delta}}{2m}}\right] \leq \delta \stackrel{\text{def}}{=} \frac{\delta'}{|\mathcal{H}|}$$

And then $\log 2/\delta = \log 2|\mathcal{H}|/\delta' = \log|\mathcal{H}| + \log 2/\delta$

Empirical Risk Minimization

- Theorem: For any $\mathcal{H}$ and any $\mathcal{D}$, $\forall^{\delta}_{S\sim\mathcal{D}^m}$,

$$L_{\mathcal{D}}(\hat{h}) \leq L_S(\hat{h}) + \sqrt{\frac{\log|\mathcal{H}| + \log {}^{2}/_{\delta}}{2m}}$$

Post-Hoc Guarantee

- Theorem: For any $\mathcal{H}$ and any $\mathcal{D}$, $\forall^{\delta}_{S\sim\mathcal{D}^m}$,

$$L_{\mathcal{D}}(\hat{h}) \leq \inf_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + 2\sqrt{\frac{\log|\mathcal{H}| + \log {}^{2}/_{\delta}}{2m}}$$

A-priori Guarantee

Proof: if indeed $\forall_{h\in\mathcal{H}}, |L_{\mathcal{D}}(h) - L_S(h)| \leq \sqrt{\cdots}$, then:

$$L_{\mathcal{D}}(\hat{h}) \leq L_S(\hat{h}) + \sqrt{\cdots} \leq L_S(h^*) + \sqrt{\cdots} \leq L_{\mathcal{D}}(h^*) + \sqrt{\cdots} + \sqrt{\cdots}$$

$\hat{h}$ minimizes of L_S and so
$L_S(\hat{h}) \leq L_S(h)$ for any h, including h^*

Post-Hoc Guarantee

- Theorem: For any $\mathcal{H}$ and any $\mathcal{D}$, $\forall^{\delta}_{S\sim\mathcal{D}^m}$,

$$L_{\mathcal{D}}(\hat{h}) \leq L_S(\hat{h}) + \sqrt{\frac{\log|\mathcal{H}| + \log {}^{2}/_{\delta}}{2m}}$$

- Performance Guarantee: Without ANY assumptions about $\mathcal{D}$ (i.e. about reality), if we somehow find a predictor h with low $L_S(h)$, we can be ensured, (with high probability) that it will perform well on future examples.

- Instead, use independent test set S' (e.g. split available examples into a training set S and test set S'). From Hoeffding:

$$L_{\mathcal{D}}(A(S)) \leq L_{S'}(A(S)) + \sqrt{\frac{\log 1/\delta}{2|S'|}}$$

Random, but depends only on S,independent of S'

- Even better using tighter Binomial tail bounds, or even better numerically with Gaussian approximation of Binomial or entropy-based bound **[see homework]**

A-Priori Learning Guarantee

- Theorem: For any $\mathcal{H}$ and any $\mathcal{D}$, $\forall^{\delta}_{S\sim\mathcal{D}^m}$,

$$L_{\mathcal{D}}(\hat{h}) \leq \underbrace{\inf_{h\in\mathcal{H}} L_{\mathcal{D}}(h)}_{\text{approximation error}} + \underbrace{2\sqrt{\frac{\log|\mathcal{H}| + \log {}^{2}/_{\delta}}{2m}}}_{\text{estimation error}}$$

- If we assume, based on our expert knowledge, that there exists a good predictor $h^* \in \mathcal{H}$, then with enough examples we can learn a predictor that's almost as good, and we know how many examples we'll need

- For any $\delta, \epsilon > 0$, using

$$m = 2\frac{\log|\mathcal{H}| + \log {}^{2}/_{\delta}}{\epsilon^2}$$

samples is enough to ensure $L_{\mathcal{D}}(\hat{h}) \leq L_{\mathcal{D}}(h^*) + \epsilon$ w.p. $\geq 1 - \delta$

Cardinality and Learning

- We saw:
  - All finite hypothesis classes are "learnable"—if we assume there is a good predictor in the class, with enough samples we'll be able to learn it (fairly powerful: includes all 100-line programs)
  - Sample complexity $\propto \log|\mathcal{H}|$

- Is cardinality the only thing controlling learnability and sample complexity?
- Is this sample complexity bound always tight?
- Are all classes of the same cardinality equally complex?
- Are there infinite classes that learnable?

Probably Approximately Correct (PAC)

- Definition: A hypothesis class $\mathcal{H}$ is **PAC-Learnable** (in the realizable case) if there exists a learning rule A such that $\forall \epsilon, \delta > 0$, $\exists m(\epsilon, \delta)$, $\forall \mathcal{D}$ s.t. $L_{\mathcal{D}}(h) = 0$ for some $h \in \mathcal{H}$ (i.e. $\mathcal{D}$ is realizable by $\mathcal{H}$), $\forall^{\delta}_{S \sim \mathcal{D}^{m(\epsilon,\delta)}}$,
$$L_{\mathcal{D}}\big(A(S)\big) \leq \epsilon$$

- Definition: A hypothesis class $\mathcal{H}$ is **agnostically PAC-Learnable** if there exists a learning rule A such that $\forall \epsilon, \delta > 0$, $\exists m(\epsilon, \delta)$, $\forall \mathcal{D}$, $\forall^{\delta}_{S \sim \mathcal{D}^{m(\epsilon,\delta)}}$,
$$L_{\mathcal{D}}\big(A(S)\big) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$$

Probably Approximately Correct (PAC)

- Definition: A hypothesis class $\mathcal{H}$ is **PAC-Learnable** (in the realizable case) if there exists a learning rule A such that $\forall \epsilon, \delta > 0$, $\exists m(\epsilon, \delta)$, $\forall \mathcal{D}$ s.t. $L_{\mathcal{D}}(h) = 0$ for some $h \in \mathcal{H}$ (i.e. $\mathcal{D}$ is realizable by $\mathcal{H}$), $\forall^{\delta}_{S \sim \mathcal{D}^{m(\epsilon,\delta)}}$,
$$L_{\mathcal{D}}\big(A(S)\big) \leq \epsilon$$

- Definition: A hypothesis class $\mathcal{H}$ is **agnostically PAC-Learnable** if there exists a learning rule A such that $\forall \epsilon, \delta > 0$, $\exists m(\epsilon, \delta)$, $\forall \mathcal{D}$, $\forall^{\delta}_{S \sim \mathcal{D}^{m(\epsilon,\delta)}}$,
$$L_{\mathcal{D}}\big(A(S)\big) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$$

- Sample complexity of a learning rule:
$$m_{A,\mathcal{H}}(\epsilon, \delta) = \min m \; s.t. \; \forall \mathcal{D}, \forall^{\delta}_{S \sim \mathcal{D}^{m(\epsilon,\delta)}}, L_{\mathcal{D}}\big(A(S)\big) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$$
- Sample complexity for learning a hypothesis class:
$$m_{\mathcal{H}}(\epsilon, \delta) = \min_{A} m_{A,\mathcal{H}}(\epsilon, \delta)$$

Cardinality and Sample Complexity

- We saw:
  - All finite hypothesis classes are PAC learnable
  - $m_{\mathcal{H}}(\epsilon, \delta) \leq m_{ERM,\mathcal{H}}(\epsilon, \delta) \leq O\left(\frac{\log|\mathcal{H}|+\log 1/\delta}{\epsilon^2}\right)$

- Are there infinite classes that are PAC learnable?
- Is the bound on $m_{\mathcal{H}}$ always tight? Can $m_{\mathcal{H}}$ be smaller than $\log|\mathcal{H}|$?
- Are all classes of the same cardinality equally complex?

E.g.

- $\mathcal{X} = \{1, \dots, 100\}, \mathcal{H} = \{\pm 1\}^{\mathcal{X}}$
- $\mathcal{X} = \{1, \dots, 2^{100} \approx 10^{30}\}$, $\mathcal{H} = \{\, [[x \leq \theta]] \mid \theta \in 1 \dots 2^{100} \,\}$

The Growth Function

- For $C = (x_1, x_2, \dots, x_m) \in \mathcal{X}^m$:
  $\Gamma_{\mathcal{H}}(C) = \left| \left\{ \left(h(x_1), h(x_2), \dots, h(x_m) \right) \in \{\pm 1\}^m \mid h \in \mathcal{H} \right\} \right|$
- $\Gamma_{\mathcal{H}}(m) = \max_{C \in \mathcal{X}^m} \Gamma_{\mathcal{H}}(C)$

E.g.

- $\mathcal{X} = \{1, \dots, 100\}, \mathcal{H} = \{\pm 1\}^{\mathcal{X}}$

$\Gamma(m) = \min(2^m, 2^{100})$

- $\mathcal{X} = \{1, \dots, 2^{100} \approx 10^{30}\}$, $\mathcal{H} = \left\{ \left[[x \leq \theta] \right] \middle| \theta \in 1 \dots 2^{100} \right.$

$\Gamma(m) = \min(\mathrm{m} + 1, 2^{100})$

Uniform Convergence using the Growth Function

- Theorem: For any hypothesis class $\mathcal{H}$ and any $\mathcal{D}$,

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[\forall_{h\in\mathcal{H}}, |L_{\mathcal{D}}(h) - L_S(h)| \leq 4\sqrt{\frac{\log|\Gamma_{\mathcal{H}}(2m)| + \log {}^{2}/_{\delta}}{m}}\right] \geq 1-\delta$$

Proof: homework

- Conclusion: For any $\mathcal{H}$ and any $\mathcal{D}$, $\forall^{\delta}_{S\sim\mathcal{D}^m}$,

$$L_{\mathcal{D}}(\hat{h}) \leq L_S(\hat{h}) + 4\sqrt{\frac{\log|\Gamma_{\mathcal{H}}(2m)| + \log {}^{2}/_{\delta}}{m}}$$

and

$$L_{\mathcal{D}}(\hat{h}) \leq \inf_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + 8\sqrt{\frac{\log|\Gamma_{\mathcal{H}}(2m)| + \log {}^{2}/_{\delta}}{m}}$$

Shattering and VC Dimension

- $C = \{x_1, \dots, x_m\}$ is **shattered** by $\mathcal{H}$ if $\Gamma_{\mathcal{H}}(C) = 2^m$, i.e. we can get all 2^m behaviors:

$$\forall_{y_1,\dots,y_m \in \pm 1}, \exists_{h \in H} \text{ s.t. } \forall_i h(x_i) = y_i$$

- The VC-dimension of $\mathcal{H}$ is the largest number of points that can be shattered by $\mathcal{H}$:

$$\mathrm{VCdim}(\mathcal{H}) = \max m \;\; s.t. \, \Gamma_{\mathcal{H}}(m) = 2^m$$

- If $\mathcal{H}$ is infinite and $\forall_m \Gamma_{\mathcal{H}}(m) = 2^m$, then $\mathrm{VCdim}(\mathcal{H}) = \infty$

VC Dimension: Examples

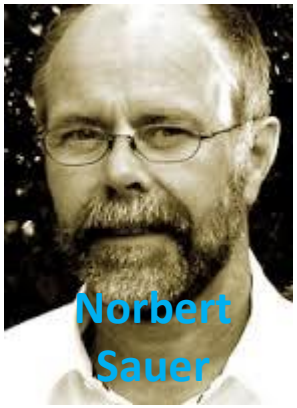
- $\mathcal{X} = \{1, \dots, 100\}, \mathcal{H} = \{\pm 1\}^{\mathcal{X}}$
  - VCdim=100
- Discrete Threshold: $\mathcal{X} = \{1, \dots, 2^{100} \approx 10^{30}\}$, $\mathcal{H} = \{ [[x \leq \theta]] \mid \theta \in 1 \dots 2^{100} \}$
  - VCdim=1
- Continuous Thresholds: $\mathcal{X} = \mathbb{R}, \mathcal{H} = \{ h_\theta(x) = [[x < \theta]] \mid \theta \in \mathbb{R} \}$
  - Only one point can be shattered; VCdim=1
- Intervals: $\mathcal{X} = \mathbb{R}, \mathcal{H} = \{ h_{a,b}(x) = [[a \leq x \leq b]] \mid a, b \in \mathbb{R} \}$
  - Can shatter any two points
  - With three points, can't realize + - +
- Axis aligned rectangles
  - Can shatter 1, 2, 3 points
  - Some sets of 4 points can't be shattered—is this a problem?
  - Some sets of 4 points can be shattered
  - Can't shatter 5 points

Sauer-Shelah-VC Lemma

- If $VCdim(\mathcal{H}) = D$, then:

for $m > D$

$$\Gamma_{\mathcal{H}}(m) \leq \sum_{i=0}^{D} \binom{m}{i} \leq \left(\frac{em}{D}\right)^{D}$$

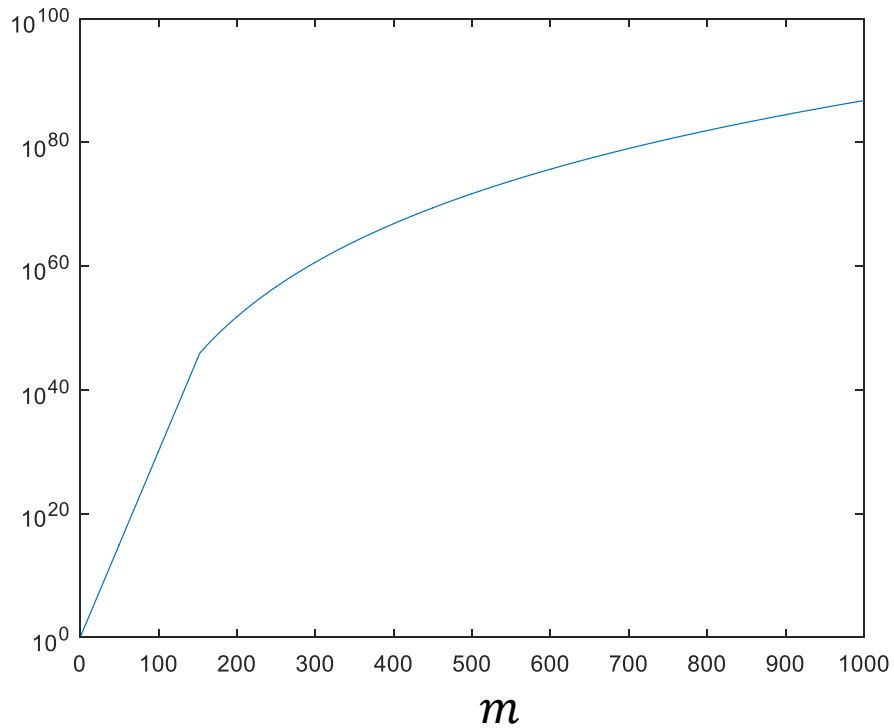


Sauer-Shelah-VC Lemma

- If $VCdim(\mathcal{H}) = D$, then:

for $m > D$

$$\Gamma_{\mathcal{H}}(m) \leq \sum_{i=0}^{D} \binom{m}{i} \leq \left(\frac{em}{D}\right)^{D}$$



Conclusion: VC Learning Guarantees

- Recall:

$$\forall^{\delta}_{S\sim\mathcal{D}^m}, \qquad L_{\mathcal{D}}(\hat{h}) \leq L_S(\hat{h}) + 4\sqrt{\frac{\log|\Gamma_{\mathcal{H}}(2m)| + \log {}^{2}/_{\delta}}{m}}$$

From Sauer, $\log|\Gamma_{\mathcal{H}}(2m)| \leq \log\left(\frac{em}{\mathrm{VCdim}}\right)^{\mathrm{VCdim}} \leq O(\mathrm{VCdim}\cdot\log m)$.

We therefor have:

$$\forall^{\delta}_{S\sim\mathcal{D}^m}, \qquad L_{\mathcal{D}}(\hat{h}) \leq L_S(\hat{h}) + O\left(\sqrt{\frac{\mathrm{VCdim}(\mathcal{H})\log m + \log {}^{1}/_{\delta}}{m}}\right)$$

With a very complex proof, this can be improved to:

$$\forall^{\delta}_{S\sim\mathcal{D}^m}, \qquad L_{\mathcal{D}}(\hat{h}) \leq L_S(\hat{h}) + O\left(\sqrt{\frac{\mathrm{VCdim}(\mathcal{H}) + \log {}^{1}/_{\delta}}{m}}\right)$$

VC Learning Guarantees

- Conclusion: If $\mathrm{VCdim}(\mathcal{H}) < \infty$ then $\mathcal{H}$ is **agnostically PAC learnable** using $ERM_{\mathcal{H}}$

$$\forall^{\delta}_{S\sim\mathcal{D}^m}, \qquad L_{\mathcal{D}}(\hat{h}) \leq L_S(\hat{h}) + O\left(\sqrt{\frac{\mathrm{VCdim}(\mathcal{H}) + \log {}^{1}/_{\delta}}{m}}\right)$$

The sample complexity is bounded by:

$$m(\epsilon, \delta) \leq O\left(\frac{\mathrm{VCdim}(\mathcal{H}) + \log(1/\delta)}{\epsilon^2}\right)$$

- Finite classes are PAC-learnable, with $m_{ERM,\mathcal{H}}(\epsilon, \delta) = O\left(\frac{\log|\mathcal{H}| + \log {}^{1}/_{\delta}}{\epsilon^2}\right)$
- VC classes are PAC-learnable, with $m_{ERM,\mathcal{H}}(\epsilon, \delta) = O\left(\frac{\mathrm{VCdim}(\mathcal{H}) + \log {}^{1}/_{\delta}}{\epsilon^2}\right)$
- **Can a class with infinite VC-dimension be learnable?**
  Can the sample complexity be lower then the VC dimension?

VC-Dimension: More Examples

- Circles in $\mathbb{R}^2$: $\mathcal{H} = \left\{ h_{c,r}(x) = \left[\left[\|x - c\| \leq r\right]\right] \mid c \in \mathbb{R}^2, r \in \mathbb{R} \right\}$
  - Can shatter 3 points
- Circles and their complement
  - Can shatter 4 points
- Circles around origin: $\mathcal{H} = \left\{ h_{c,r}(x) = \left[\left[\|x\| \leq r\right]\right] \mid r \in \mathbb{R} \right\}$
  - Can shatter only 1 point
- Axis aligned ellipses:

$$\mathcal{H} = \left\{ h_{c,r[1],r[2]}(x) = \left[\left[\frac{(x[1] - c[1])^2}{r[1]^2} + \frac{(x[2] - c[2])^2}{r[2]^2} \leq 1\right]\right] \mid c \in \mathbb{R}^2, r[1], r[2] \in \mathbb{R} \right\}$$

  - Can shatter 4 points
- General ellipses
  - Can shatter 5 points

- Upper bounds?