PAC Learning
+
The Big Picture
Reminders

• Midterm Exam
  – Thursday Evening 6:30 – 9:00 (2.5 hours)
  – Room and seat assignments will be announced on Piazza
  – You may bring one 8.5 x 11 cheatsheet
Midterm Exam

• **Time / Location**
  – **Time:** Evening Exam  
    Thu, March 22 at 6:30pm – 9:00pm  
  – **Room:** We will contact each student individually with your room assignment. The rooms are not based on section.  
  – **Seats:** There will be assigned seats. Please arrive early.  
  – Please watch Piazza carefully for announcements regarding room / seat assignments.

• **Logistics**
  – Format of questions:  
    • Multiple choice  
    • True / False (with justification)  
    • Derivations  
    • Short answers  
    • Interpreting figures  
    • Implementing algorithms on paper  
  – No electronic devices  
  – You are allowed to **bring** one 8½ x 11 sheet of notes (front and back)
LEARNING THEORY
Questions For Today

1. Given a classifier with zero training error, what can we say about generalization error? 
   (Sample Complexity, Realizable Case)

2. Given a classifier with low training error, what can we say about generalization error? 
   (Sample Complexity, Agnostic Case)

3. Is there a theoretical justification for regularization to avoid overfitting? 
   (Structural Risk Minimization)
Sample Complexity Results

Definition 0.1. The **sample complexity** of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

**Four Cases we care about…**

| Finite $|\mathcal{H}|$ | Realizable | Agnostic |
|------------------------|------------|----------|
| $N \geq \frac{1}{\varepsilon} \left[ \log(|\mathcal{H}|) + \log\left(\frac{1}{\delta}\right) \right]$ labeled examples are sufficient so that with probability $(1 - \delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) > 0$. |
| $N \geq \frac{1}{2\varepsilon^2} \left[ \log(|\mathcal{H}|) + \log\left(\frac{2}{\delta}\right) \right]$ labeled examples are sufficient so that $|R(h) - \hat{R}(h)| < \varepsilon$ with probability $(1 - \delta)$ for all $h \in \mathcal{H}$. |

We need a new definition of “complexity” for a Hypothesis space for these results (see VC Dimension).
**Sample Complexity Results**

**Definition 0.1.** The sample complexity of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

**Four Cases we care about...**

| Finite $|\mathcal{H}|$ | Realizable | Agnostic |
|----------------------|------------|----------|
| $N \geq \frac{1}{\epsilon} \left[ \log(|\mathcal{H}|) + \log(\frac{1}{\delta}) \right]$ labeled examples are sufficient so that with probability $(1 - \delta)$ all $h \in \mathcal{H}$ with $R(h) \geq \epsilon$ have $\hat{R}(h) > 0$. | $N \geq \frac{1}{2\epsilon^2} \left[ \log(|\mathcal{H}|) + \log(\frac{3}{\delta}) \right]$ labeled examples are sufficient so that with probability $(1 - \delta)$ for all $h \in \mathcal{H}$ we have that $|R(h) - \hat{R}(h)| < \epsilon$. |

| Infinite $|\mathcal{H}|$ | Realizable | Agnostic |
|----------------------|------------|----------|
| $N = O\left( \frac{1}{\epsilon} \left[ VC(\mathcal{H}) \log(\frac{1}{\epsilon}) + \log(\frac{1}{\delta}) \right] \right)$ labeled examples are sufficient so that with probability $(1 - \delta)$ all $h \in \mathcal{H}$ with $R(h) \geq \epsilon$ have $\hat{R}(h) > 0$. | $N = O\left( \frac{1}{\epsilon^2} \left[ VC(\mathcal{H}) + \log(\frac{1}{\delta}) \right] \right)$ labeled examples are sufficient so that with probability $(1 - \delta)$ for all $h \in \mathcal{H}$ we have that $|R(h) - \hat{R}(h)| \leq \epsilon$. |
VC DIMENSION
What if $H$ is infinite?

E.g., linear separators in $\mathbb{R}^d$

E.g., thresholds on the real line

E.g., intervals on the real line
Shattering, VC-dimension

Definition:

A set of points $S$ is shattered by $H$ is there are hypotheses in $H$ that split $S$ in all of the $2^{|S|}$ possible ways; i.e., all possible ways of classifying points in $S$ are achievable using concepts in $H$.

Definition: VC-dimension (Vapnik-Chervonenkis dimension)

The **VC-dimension** of a hypothesis space $H$ is the cardinality of the largest set $S$ that can be shattered by $H$.

If arbitrarily large finite sets can be shattered by $H$, then $VCdim(H) = \infty$
**Definition:** VC-dimension (Vapnik-Chervonenkis dimension)

The **VC-dimension** of a hypothesis space $H$ is the cardinality of the largest set $S$ that can be shattered by $H$.

If arbitrarily large finite sets can be shattered by $H$, then $\text{VCdim}(H) = \infty$

To show that VC-dimension is $d$:
- there exists a set of $d$ points that can be shattered
- there is no set of $d+1$ points that can be shattered.

**Fact:** If $H$ is finite, then $\text{VCdim}(H) \leq \log(|H|)$.  

---

*Slide from Nina Balcan*
Shattering, VC-dimension

If the VC-dimension is $d$, that means there exists a set of $d$ points that can be shattered, but there is no set of $d+1$ points that can be shattered.

E.g., $H=$ Thresholds on the real line

$\text{VCdim}(H) = 1$

E.g., $H=$ Intervals on the real line

$\text{VCdim}(H) = 2$
If the VC-dimension is $d$, that means there exists a set of $d$ points that can be shattered, but there is no set of $d+1$ points that can be shattered.

*E.g.* , $H =$ Union of $k$ intervals on the real line $\quad \text{VCdim}(H) = 2k$

$\quad \begin{array}{ccccccccccc}
- & + & - & + & - & + \\
\hline
\end{array}$

$\text{VCdim}(H) \geq 2k \quad \text{A sample of size } 2k \text{ shatters}
\text{(treat each pair of points as a separate case of intervals)}$

$\text{VCdim}(H) < 2k + 1$

$\quad \begin{array}{cccccccc}
+ & - & + & - & + \\
\hline
\ldots
\end{array}$
Shattering, VC-dimension

E.g., $H =$ linear separators in $\mathbb{R}^2$

$\text{VCdim}(H) \geq 3$
Shattering, VC-dimension

E.g., $H =$ linear separators in $\mathbb{R}^2$

$\text{VCdim}(H) < 4$

Case 1: one point inside the triangle formed by the others. Cannot label inside point as positive and outside points as negative.

Case 2: all points on the boundary (convex hull). Cannot label two diagonally as positive and other two as negative.

Fact: $\text{VCdim}$ of linear separators in $\mathbb{R}^d$ is $d+1$
## Sample Complexity Results

**Definition 0.1.** The *sample complexity* of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

### Four Cases we care about...

<table>
<thead>
<tr>
<th></th>
<th>Realizable</th>
<th>Agnostic</th>
</tr>
</thead>
<tbody>
<tr>
<td>**Finite $</td>
<td>\mathcal{H}</td>
<td>$**</td>
</tr>
<tr>
<td>**Infinite $</td>
<td>\mathcal{H}</td>
<td>$**</td>
</tr>
</tbody>
</table>
Corollary 3 (Realizable, Infinite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1 - \delta)$, for any hypothesis $h$ in $\mathcal{H}$ consistent with the data (i.e. with $\hat{R}(h) = 0$),

$$R(h) \leq O \left( \frac{1}{N} \left[ \text{VC}(\mathcal{H}) \ln \left( \frac{N}{\text{VC}(\mathcal{H})} \right) + \ln \left( \frac{1}{\delta} \right) \right] \right) \quad (1)$$

Corollary 4 (Agnostic, Infinite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1 - \delta)$, for all hypotheses $h$ in $\mathcal{H}$,

$$R(h) \leq \hat{R}(h) + O \left( \sqrt{\frac{1}{N} \left[ \text{VC}(\mathcal{H}) + \ln \left( \frac{1}{\delta} \right) \right]} \right) \quad (2)$$
Generalization and Overfitting

Whiteboard:

– Empirical Risk Minimization
– Structural Risk Minimization
– Motivation for Regularization
Questions For Today

1. Given a classifier with zero training error, what can we say about generalization error? *(Sample Complexity, Realizable Case)*

2. Given a classifier with low training error, what can we say about generalization error? *(Sample Complexity, Agnostic Case)*

3. Is there a theoretical justification for regularization to avoid overfitting? *(Structural Risk Minimization)*
Learning Theory Objectives

You should be able to...

• Identify the properties of a learning setting and assumptions required to ensure low generalization error
• Distinguish true error, train error, test error
• Define PAC and explain what it means to be approximately correct and what occurs with high probability
• Apply sample complexity bounds to real-world learning examples
• Distinguish between a large sample and a finite sample analysis
• Theoretically motivate regularization
CLASSIFICATION AND REGRESSION
Classification and Regression: The Big Picture

Whiteboard

- **Decision Rules / Models** (probabilistic generative, probabilistic discriminative, perceptron, SVM, regression)
- **Objective Functions** (likelihood, conditional likelihood, hinge loss, mean squared error)
- **Regularization** (L1, L2, priors for MAP)
- **Update Rules** (SGD, perceptron)
- **Nonlinear Features** (preprocessing, kernel trick)
### Learning Paradigms:
What data is available and when? What form of prediction?
- supervised learning
- unsupervised learning
- semi-supervised learning
- reinforcement learning
- active learning
- imitation learning
- domain adaptation
- online learning
- density estimation
- recommender systems
- feature learning
- manifold learning
- dimensionality reduction
- ensemble learning
- distant supervision
- hyperparameter optimization

### Problem Formulation:
What is the structure of our output prediction?
<table>
<thead>
<tr>
<th>Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>boolean</td>
<td>Binary Classification</td>
</tr>
<tr>
<td>categorical</td>
<td>Multiclass Classification</td>
</tr>
<tr>
<td>ordinal</td>
<td>Ordinal Classification</td>
</tr>
<tr>
<td>real</td>
<td>Regression</td>
</tr>
<tr>
<td>ordering</td>
<td>Ranking</td>
</tr>
<tr>
<td>multiple discrete</td>
<td>Structured Prediction</td>
</tr>
<tr>
<td>multiple continuous</td>
<td>(e.g. dynamical systems)</td>
</tr>
<tr>
<td>both discrete &amp; continuous</td>
<td>(e.g. mixed graphical models)</td>
</tr>
</tbody>
</table>

### Facets of Building ML Systems:
How to build systems that are robust, efficient, adaptive, effective?
1. Data prep
2. Model selection
3. Training (optimization / search)
4. Hyperparameter tuning on validation data
5. (Blind) Assessment on test data

### Application Areas:
Key challenges? NLP, Speech, Computer Vision, Robotics, Medicine, Search

### Theoretical Foundations:
What principles guide learning?
- probabilistic
- information theoretic
- evolutionary search
- ML as optimization

### Big Ideas in ML:
Which are the ideas driving development of the field?
- inductive bias
- generalization / overfitting
- bias-variance decomposition
- generative vs. discriminative
- deep nets, graphical models
- PAC learning
- distant rewards