PAC Learning
+
Big Picture
## Q&A

**Q:** Why do we shuffle the examples in SGD?

**A:** This is how we do sampling without replacement

1. **Theoretically** we can show sampling without replacement is not significantly worse than sampling with replacement (Shamir, 2016)
2. **Practically** sampling without replacement tends to work better

**Q:** What is “bias”?

**A:** That depends. The word “bias” shows up all over machine learning! Watch out...

1. The additive term in a linear model (i.e. b in $w^Tx + b$)
2. Inductive bias is the principle by which a learning algorithm generalizes to unseen examples
3. Bias of a model in a societal sense may refer to racial, socio-economic, gender biases that exist in the predictions of your model
4. The difference between the expected predictions of your model and the ground truth (as in “bias-variance tradeoff”)

Reminders

• Homework 5: Neural Networks
  – Out: Fri, Mar 1
  – Due: Fri, Mar 22 at 11:59pm

• Today’s In-Class Poll

• Matt’s office hours for Mon, 3/18 are rescheduled to Tue (3/19) -- see Piazza/GCal
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 |
|----------------------|------------|----------|
| **Thm. 1** $N \geq \frac{1}{\epsilon} \left[ \log(|\mathcal{H}|) + \log\left(\frac{1}{\delta}\right) \right]$ labeled examples are sufficient so that with probability $(1 - \delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$. |            |          |
| Infinite $|\mathcal{H}|$ |            |          |
Question:
Suppose \( H = \) class of conjunctions over \( x \) in \( \{0,1\}^M \)

Example hypotheses:
\[
h(x) = x_1 (1-x_3) x_5 \\
h(x) = x_1 (1-x_2) x_4 (1-x_5)
\]

If \( M = 10, \varepsilon = 0.1, \delta = 0.01 \), how many examples suffice according to Theorem 1?

Answer:
A. \( 10*(2*\ln(10)+\ln(100)) \approx 92 \)
B. \( 10*(3*\ln(10)+\ln(100)) \approx 116 \)
C. \( 10*(10*\ln(2)+\ln(100)) \approx 116 \)
D. \( 10*(10*\ln(3)+\ln(100)) \approx 156 \)
E. \( 100*(2*\ln(10)+\ln(10)) \approx 691 \)
F. \( 100*(3*\ln(10)+\ln(10)) \approx 922 \)
G. \( 100*(10*\ln(2)+\ln(10)) \approx 924 \)
H. \( 100*(10*\ln(3)+\ln(10)) \approx 1329 \)

**Thm. 1** \[ N \geq \frac{1}{\varepsilon} \left[ \log(|\mathcal{H}|) + \log\left(\frac{1}{\delta}\right) \right] \] labeled examples are sufficient so that with probability \( (1-\delta) \) all \( h \in \mathcal{H} \) with \( \hat{R}(h) = 0 \) have \( R(h) \leq \varepsilon \).
**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>$N \geq \frac{1}{\epsilon} \left[ \log(</td>
</tr>
<tr>
<td>Infinite $</td>
<td>\mathcal{H}$</td>
<td>-</td>
</tr>
</tbody>
</table>
Sample Complexity Results

Realizable

1. Bound is **inversely linear in epsilon** (e.g. halving the error requires double the examples)
2. Bound is **only logarithmic in |\mathcal{H}|** (e.g. quadrupling the hypothesis space only requires double the examples)

Agnostic

1. Bound is **inversely quadratic in epsilon** (e.g. halving the error requires 4x the examples)
2. Bound is **only logarithmic in |\mathcal{H}|** (i.e. same as Realizable case)

<table>
<thead>
<tr>
<th>Finite</th>
<th>|\mathcal{H}|</th>
<th>Infinite (|\mathcal{H}|)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thm. 1</td>
<td>(N \geq \frac{1}{\epsilon} \left[ \log(</td>
<td>\mathcal{H}</td>
</tr>
<tr>
<td>Thm. 2</td>
<td>(N \geq \frac{1}{2\epsilon^2} \left[ \log(</td>
<td>\mathcal{H}</td>
</tr>
</tbody>
</table>
**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 |
|----------------------|------------|----------|
| Thm. 1 | $N \geq \frac{1}{\epsilon} \left[ \log(|\mathcal{H}|) + \log\left(\frac{1}{\delta}\right) \right]$ labeled examples are sufficient so that with probability $(1 - \delta)$ all $h \in \mathcal{H}$ will have $R(h) \leq \epsilon$. |
| Thm. 2 | $N \geq \frac{1}{2\epsilon^2} \left[ \log(|\mathcal{H}|) + \log\left(\frac{2}{\delta}\right) \right]$ labeled examples are sufficient so that for all $h \in \mathcal{H}$ we have $R(h) \leq \epsilon$. |

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 |
|----------------------|------------|----------|
| **Thm. 1** $N \geq \frac{1}{\varepsilon} \left[ \log(|\mathcal{H}|) + \log\left(\frac{1}{\delta}\right) \right]$ labeled examples are sufficient so that with probability $(1 - \delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \varepsilon$. | **Thm. 2** $N \geq \frac{1}{2\varepsilon^2} \left[ \log(|\mathcal{H}|) + \log\left(\frac{2}{\delta}\right) \right]$ labeled examples are sufficient so that with probability $(1 - \delta)$ for all $h \in \mathcal{H}$ we have that $|R(h) - \hat{R}(h)| \leq \varepsilon$. |

| Infinite $|\mathcal{H}|$ | **Thm. 3** $N = O\left(\frac{1}{\varepsilon^2} \left[ VC(\mathcal{H}) \log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right) \right] \right)$ labeled examples are sufficient so that with probability $(1 - \delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \varepsilon$. | **Thm. 4** $N = O\left(\frac{1}{\varepsilon^2} \left[ VC(\mathcal{H}) + \log\left(\frac{1}{\delta}\right) \right] \right)$ labeled examples are sufficient so that with probability $(1 - \delta)$ for all $h \in \mathcal{H}$ we have that $|R(h) - \hat{R}(h)| \leq \varepsilon$. |
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 $H$ if 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 $\text{VCdim}(H) = \infty$.
**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 $\text{VCdim}(H) \leq \log(|H|)$.
Shattering, VC-dimension

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

$\text{VCdim}(H) \geq 3$
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: VCdim of linear separators in \( \mathbb{R}^d \) is \( d+1 \)
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

$\text{VCdim}(H) = 2$

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., $H =$ Union of $k$ intervals on the real line \[ \text{VCdim}(H) = 2k \]

A sample of size $2k$ shatters (treat each pair of points as a separate case of intervals)

\[ \text{VCdim}(H) \geq 2k \]

\[ \text{VCdim}(H) < 2k + 1 \]
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...

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

| Infinite $|\mathcal{H}|$ | | |
| **Thm. 3** $N = O\left(\frac{1}{\epsilon} \left[ \text{VC}(\mathcal{H}) \log\left(\frac{1}{\epsilon}\right) + \log\left(\frac{1}{\delta}\right) \right]\right)$ labeled examples are sufficient so that with probability $(1 - \delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$. |
| **Thm. 4** $N = O\left(\frac{1}{\epsilon^2} \left[ \text{VC}(\mathcal{H}) + \log\left(\frac{1}{\delta}\right) \right]\right)$ labeled examples are sufficient so that with probability $(1 - \delta)$ for all $h \in \mathcal{H}$ we have that $|R(h) - \hat{R}(h)| \leq \epsilon$. |
SLT-style Corollaries

Thm. 1  \[ N \geq \frac{1}{\epsilon} \left[ \log(|\mathcal{H}|) + \log\left(\frac{1}{\delta}\right) \right] \] labeled examples are sufficient so that with probability \((1 - \delta)\) all \(h \in \mathcal{H}\) with \(\hat{R}(h) = 0\) have \(R(h) \leq \epsilon\).

Solve the inequality in Thm.1 for \(\epsilon\) to obtain Corollary 1

Corollary 1 (Realizable, Finite \(|\mathcal{H}|\)). For some \(\delta > 0\), with probability at least \((1 - \delta)\), for any \(h\) in \(\mathcal{H}\) consistent with the training data (i.e. \(\hat{R}(h) = 0\)),

\[ R(h) \leq \frac{1}{N} \left[ \ln(|\mathcal{H}|) + \ln \left(\frac{1}{\delta}\right) \right] \]

We can obtain similar corollaries for each of the theorems...
Corollary 1 (Realizable, Finite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1 - \delta)$, for any $h$ in $\mathcal{H}$ consistent with the training data (i.e. $\hat{R}(h) = 0$),

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

Corollary 2 (Agnostic, Finite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1 - \delta)$, for all hypotheses $h$ in $\mathcal{H}$,

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{1}{2N} \left[ \ln(|\mathcal{H}|) + \ln \left( \frac{2}{\delta} \right) \right]}$$
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 $\hat{R}(h) = 0$),

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

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

$$R(h) \leq \hat{R}(h) + O \left( \sqrt{\frac{1}{N} \left[ \text{VC}(\mathcal{H}) + \ln \left( \frac{1}{\delta} \right) \right]} \right)$$  \hspace{1cm} (2)
SLT-style Corollaries

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 $\hat{R}(h) = 0$),

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

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

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

Should these corollaries inform how we do model selection?
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
The Big Picture

CLASSIFICATION AND REGRESSION
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?
- boolean
- categorical
- ordinal
- real
- ordering
- multiple discrete
- multiple continuous
- both discrete & cont.

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

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

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

Application Areas
Key challenges?
NLP, Speech, Computer Vision, Robotics, Medicine, Search
Classification and Regression: 
The Big Picture

Whiteboard

– Decision Rules / Models
– Objective Functions
– Regularization
– Update Rules
– Nonlinear Features
PROBABILISTIC LEARNING
Function Approximation
Previously, we assumed that our output was generated using a deterministic target function:

\[
\begin{align*}
\mathbf{x}(i) & \sim p^*(\cdot) \\
y(i) & = c^*(\mathbf{x}(i))
\end{align*}
\]

Our goal was to learn a hypothesis \( h(\mathbf{x}) \) that best approximates \( c^*(\mathbf{x}) \)

Probabilistic Learning
Today, we assume that our output is sampled from a conditional probability distribution:

\[
\begin{align*}
\mathbf{x}(i) & \sim p^*(\cdot) \\
y(i) & \sim p^*(\cdot|\mathbf{x}(i))
\end{align*}
\]

Our goal is to learn a probability distribution \( p(y|x) \) that best approximates \( p^*(y|x) \)
PROBABILITY
# Random Variables: Definitions

<table>
<thead>
<tr>
<th>Discrete Random Variable</th>
<th>( X )</th>
<th>Random variable whose values come from a countable set (e.g. the natural numbers or {True, False})</th>
</tr>
</thead>
</table>
| Probability mass function (pmf) | \( p(x) \) | Function giving the probability that discrete r.v. \( X \) takes value \( x \).  \[
p(x) := P(X = x)
\]
## Random Variables: Definitions

<table>
<thead>
<tr>
<th>Continuous Random Variable</th>
<th>( X )</th>
<th>Random variable whose values come from an interval or collection of intervals (e.g. the real numbers or the range (3, 5))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability density function (pdf)</td>
<td>( f(x) )</td>
<td>Function the returns a nonnegative real indicating the relative likelihood that a continuous r.v. ( X ) takes value ( x )</td>
</tr>
</tbody>
</table>

- For any continuous random variable: \( P(X = x) = 0 \)
- Non-zero probabilities are only available to intervals:

\[
P(a \leq X \leq b) = \int_{a}^{b} f(x) \, dx
\]
Random Variables: Definitions

**Cumulative distribution function**

<table>
<thead>
<tr>
<th>$F(x)$</th>
<th>Function that returns the probability that a random variable $X$ is less than or equal to $x$: $F(x) = P(X \leq x)$</th>
</tr>
</thead>
</table>

- **For discrete** random variables:

  \[
  F(x) = P(X \leq x) = \sum_{x' < x} P(X = x') = \sum_{x' < x} p(x')
  \]

- **For continuous** random variables:

  \[
  F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x') \, dx'
  \]
Notational Shortcuts

A convenient shorthand:

\[ P(A|B) = \frac{P(A, B)}{P(B)} \]

\[ \Rightarrow \text{For all values of } a \text{ and } b: \]

\[ P(A = a|B = b) = \frac{P(A = a, B = b)}{P(B = b)} \]
But then how do we tell $P(E)$ apart from $P(X)$?

Instead of writing: $P(A|B) = \frac{P(A, B)}{P(B)}$

We should write: $P_{A|B}(A|B) = \frac{P_{A,B}(A, B)}{P_{B}(B)}$

...but only probability theory textbooks go to such lengths.
COMMON PROBABILITY DISTRIBUTIONS
Common Probability Distributions

- For Discrete Random Variables:
  - Bernoulli
  - Binomial
  - Multinomial
  - Categorical
  - Poisson

- For Continuous Random Variables:
  - Exponential
  - Gamma
  - Beta
  - Dirichlet
  - Laplace
  - Gaussian (1D)
  - Multivariate Gaussian
Common Probability Distributions

Beta Distribution

probability density function:

\[ f(\phi|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1 - x)^{\beta-1} \]
Common Probability Distributions

Dirichlet Distribution

probability density function:

\[
f(\phi|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1 - x)^{\beta-1}
\]
Common Probability Distributions

Dirichlet Distribution

probability density function:

\[ p(\phi | \alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \phi_k^{\alpha_k - 1} \]

where \( B(\alpha) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)} \)
EXPECTATION AND VARIANCE
Expectation and Variance

The **expected value** of $X$ is $E[X]$. Also called the mean.

- **Discrete random variables:**
  
  \[ E[X] = \sum_{x \in \mathcal{X}} xp(x) \]

  Suppose $X$ can take any value in the set $\mathcal{X}$.

- **Continuous random variables:**
  
  \[ E[X] = \int_{-\infty}^{+\infty} x f(x) \, dx \]
The variance of $X$ is $Var(X)$.

\[ Var(X) = E[(X - E[X])^2] \]

- **Discrete random variables:**
  \[ Var(X) = \sum_{x \in \mathcal{X}} (x - \mu)^2 p(x) \]

- **Continuous random variables:**
  \[ Var(X) = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) \, dx \]

$\mu = E[X]$
Joint probability
Marginal probability
Conditional probability

MULTIPLE RANDOM VARIABLES
Joint Probability

- Key concept: two or more random variables may interact. Thus, the probability of one taking on a certain value depends on which value(s) the others are taking.
- We call this a joint ensemble and write
  \[ p(x, y) = \text{prob}(X = x \text{ and } Y = y) \]
Marginal Probabilities

• We can ”sum out” part of a joint distribution to get the *marginal distribution* of a subset of variables:

\[ p(x) = \sum_y p(x, y) \]

• This is like adding slices of the table together.

• Another equivalent definition: \[ p(x) = \sum_y p(x|y)p(y). \]
Conditional Probability

- If we know that some event has occurred, it changes our belief about the probability of other events.
- This is like taking a "slice" through the joint table.

\[
p(x|y) = \frac{p(x, y)}{p(y)}
\]
Independence and Conditional Independence

- Two variables are independent iff their joint factors:
  \[ p(x, y) = p(x)p(y) \]

- Two variables are conditionally independent given a third one if for all values of the conditioning variable, the resulting slice factors:
  \[ p(x, y|z) = p(x|z)p(y|z) \quad \forall z \]