What is machine learning

- Finding patterns in data
- Adapting program behavior
- Advertise a customer’s favorite products
- Search the web to find pictures of dogs
- Change radio channel when user says “change channel”

Advertise a customer’s favorite products

This summer, I had two meetings, one in Portland and one in Baltimore

Today I get an e-mail from Priceline:

Search the web to find pictures of dogs

Filenames:
- Dog.jpg
- Puppy.bmp

Caption text

Pixel patterns
Change radio channel when user says “change channel”

• Distinguish user’s voice from music
• Understand what user has said

What’s covered in this class

• Theory: describing patterns in data
  • Probability
  • Linear algebra
  • Calculus/optimization

• Implementation: programming to find and react to patterns in data
  • Popular and successful algorithms
  • Matlab
  • Data sets of text, speech, pictures, user actions, neural data...

Outline of topics

• Groundwork: probability and slopes
• Classification overview: Training, testing, and overfitting
• Basic classifiers: Naive Bayes and Logistic Regression
• Advanced classifiers: Neural networks and support vector machines
  - Deep learning
  - Kernel methods
• Dimensionality reduction: Feature selection, information criteria
• Graphical models: Hidden Markov model (possibly Bayes nets)
• Expectation-Maximization

What you need to do in this class

• Class attendance
• Assignments: homeworks (4) and final project
• Exams: midterm and final

• Don’t cheat
  • You may discuss homeworks with other students, but your submitted work must be your own. Copying is not allowed.
Resources

- Office hours: Wednesday 5-6pm and by appointment
- Course web site: http://storm.cis.fordham.edu/leeds/cisc5800
- Fellow students
- Textbooks/online notes
- Matlab

Probability and basic calculus

Probability

What is the probability that a child likes chocolate?

- Ask 100 children
- Count who likes chocolate
- Divide by number of children asked

\[ P(\text{"child likes chocolate"}) = \frac{85}{100} = 0.85 \]

In short: \( P(C) = 0.85 \)  
\( C=\text{"child likes chocolate"} \)

General probability properties

\( P(A) \) means “Probability that statement A is true”

- \( 0 \leq \text{Prob}(A) \leq 1 \)
- \( \text{Prob(True)} = 1 \)
- \( \text{Prob(False)} = 0 \)

<table>
<thead>
<tr>
<th>Name</th>
<th>Chocolate?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sarah</td>
<td>Yes</td>
</tr>
<tr>
<td>Melissa</td>
<td>Yes</td>
</tr>
<tr>
<td>Darren</td>
<td>No</td>
</tr>
<tr>
<td>Stacy</td>
<td>Yes</td>
</tr>
<tr>
<td>Brian</td>
<td>No</td>
</tr>
</tbody>
</table>
Random variables

A variable can take on a value from a given set of values:

- {True, False}
- {Cat, Dog, Horse, Cow}
- {0,1,2,3,4,5,6,7}

A random variable holds each value with a given probability

Example: binary variable
- \( P(\text{LikesChocolate}) = P(\text{LikesChocolate}=\text{True}) = 0.85 \)

Complements

What is the probability that a child DOES NOT like chocolate?

Complement: \( C' = \text{“child doesn’t like chocolate”} \)

\[ P(C') = 0.15 \]

In general: \( P(A') = 1 - P(A) \)

Joint probabilities

Across 100 children:

- 55 like chocolate AND ice cream \( P(I,C) = P(I=\text{True}, C=\text{True}) = 0.55 \)
- 30 like chocolate but not ice cream \( P(I',C) = P(I=\text{False}, C=\text{True}) = 0.3 \)
- 5 like ice cream but not chocolate \( P(I,C') = 0.05 \)
- 10 don’t like chocolate nor ice cream

\[ \text{Prob}(I) = P(I=\text{True}) = 0.6 \]
\[ \text{Prob}(C) = P(C=\text{True}) = 0.85 \]

Marginal and conditional probabilities

For two binary random variables A and B

- \( P(A) = P(A,B) + P(A,B') = P(A=\text{True}, B=\text{True}) + P(A=\text{True}, B=\text{False}) \)
- \( P(B) = P(A,B) + P(A',B) \)

For marginal probability \( P(X) \), “marginalize” over all possible values of the other random variables

- \( \text{Prob}(C|I) : \text{Probability child likes chocolate given s/he likes ice cream} \)

\[ P(C|I) = \frac{P(C,I)}{P(I)} = \frac{P(C,I)}{P(C,I) + P(C',I)} \]
Independence

If the truth value of B does not affect the truth value of A, we say A and B are independent.

\[ P(A | B) = P(A) \]
\[ P(A, B) = P(A) \cdot P(B) \]

Multi-valued random variables

A random variable can hold more than two values, each with a given probability

- \( P(\text{Animal}=\text{Cat})=0.5 \)
- \( P(\text{Animal}=\text{Dog})=0.3 \)
- \( P(\text{Animal}=\text{Horse})=0.1 \)
- \( P(\text{Animal}=\text{Cow})=0.1 \)

Probability rules: multi-valued variables

For given random variable A:

- \( P(A = a_i \text{ and } A = a_j) = 0 \) if \( i \neq j \)
- \( \sum_i P(A = a_i) = 1 \)
- \( P(A = a_i) = \sum_j P(A = a_i, B = b_j) \), where \( \alpha \) is a value assignment for variable A

Probability table

- \( P(\text{G}=\text{C}, \text{H}=\text{True})=0.15 \)
- \( P(\text{H}=\text{True})=0.75 \)
- \( P(\text{G}=\text{C} | \text{H}=\text{True}) = \frac{15}{75} = 0.2 \)
- \( P(\text{H}=\text{True} | \text{G}=\text{C}) = \frac{15}{2} = 0.75 \)
Continuous random variables

A random variable can take on a continuous range of values
• From 0 to 1
• From 0 to $\infty$
• From $-\infty$ to $\infty$

Probability expressed through a “probability density function” $f(x)$

Common probability distributions

• Uniform: $f_{\text{uniform}}(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$

• Gaussian: $f_{\text{gauss}}(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

The Gaussian function $f_{\text{gauss}}(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

• Mean $\mu$ – center of distribution
• Standard deviation $\sigma$ – width of distribution

Calculus: finding the slope of a function

What is the minimum value of: $f(x)=x^2-5x+6$

Find value of $x$ where slope is 0

General rules:
• $\frac{d}{dx}x^a = ax^{a-1}$
• $\frac{d}{dx}kf(x) = kf'(x)$
• $\frac{d}{dx}[f(x) + g(x)] = f'(x) + g'(x)$
Calculus: finding the slope of a function

What is the minimum value of: \( f(x) = x^2 - 5x + 6 \)

- \( f'(x) = 2x - 5 \)
- What is the slope at \( x = 5 \)? \( f'(5) = 5 \)
- What is the slope at \( x = -3 \)? \( f'(-3) = -11 \)
- What value of \( x \) gives slope of 0? \( x = 2.5 \)

More on derivatives: \( \frac{d}{dx} f(x) = f'(x) \)

- \( \frac{d}{dx} f(w) = 0 \) -- \( w \) is not related to \( x \), so derivative is 0
- \( \frac{d}{dx} f(g(x)) = g'(x) \cdot f'(g(x)) \)
- \( \frac{d}{dx} \log x = \frac{1}{x} \)
- \( \frac{d}{dx} e^x = e^x \)

Introduction to classifiers

The goal of a classifier

- Learn function \( C \) to maximize correct labels (\( Y \)) based on features (\( X \))

\[ C(x) = y \]
Giraffe detector

- Label $x$ : height
- Class $y$ : True or False ("is giraffe" or "is not giraffe")

Learn optimal classification parameter(s)
- Parameter: $x_{\text{thresh}}$

Example function:

$$ C(x) = \begin{cases} \text{True} & \text{if } x > x_{\text{thresh}} \\ \text{False} & \text{otherwise} \end{cases} $$

Learning our classifier parameter(s)

- Adjust parameter(s) based on observed data
- Training set: contains features and corresponding labels

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5</td>
<td>True</td>
</tr>
<tr>
<td>2.2</td>
<td>True</td>
</tr>
<tr>
<td>1.8</td>
<td>True</td>
</tr>
<tr>
<td>1.2</td>
<td>False</td>
</tr>
<tr>
<td>0.9</td>
<td>False</td>
</tr>
</tbody>
</table>

The testing set

- Does classifier correctly label new data?

Training set

Example "good" performance: 90% correct labels

Testing set should be distinct from training set!

Be careful with your training set

- What if we train with only baby giraffes and ants?
- What if we train with only T rexes and adult giraffes?
Training vs. testing

- **Training**: learn parameters from set of data in each class
- **Testing**: measure how often classifier correctly identifies new data

- More training reduces classifier error $\varepsilon$
- Too much training data causes worse testing error – overfitting