Machine Learning
CISC 5800
Dr Daniel Leeds

What is machine learning
• Finding patterns in data
• Adapting program behavior

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:

Dog photos and the internet
Model of dog appearance
Search for “dog” from web images:
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: Naïve Bayes and Logistic Regression
• Advanced classifiers: Neural networks and support vector machines
  • Deep learning
  • Kernel methods
• Dimensionality reduction: Feature selection, information criteria
• Graphical models: Bayes Nets and Hidden Markov Model
• 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: Monday 4-5pm and by appointment
• Course web site: http://storm.cis.fordham.edu/leeds/cisc5800
• Fellow students
• Textbooks/online notes
  • Andrew Ng’s Stanford course notes
  • Matlab

Andrew Ng’s Stanford course notes
CS229 Machine Learning
Autumn 2016
Probability and basic calculus

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”

• \( 0\leq P(A) \leq 1 \)
• \( P(\text{True})=1 \)
• \( P(\text{False})=0 \)
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** LikesChocolate
- \(P(\text{LikesChocolate}) = P(\text{LikesChocolate}=\text{True}) = 0.85\)

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
- 5 like ice cream but not chocolate
- 10 don’t like chocolate nor ice cream

\[ P(I',C) = P(I=\text{False}, C=\text{True}) = 0.3 \]
\[ P(I,C') = 0.05 \]
\[ P(I,C) = P(\text{True}) = 0.6 \]
\[ P(C) = P(\text{True}) = 0.85 \]

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

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

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

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

- $P(A|B) = P(A)$
- $P(A,B) = P(A)P(B)$

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 $i \neq j$
- $\sum_i P(A = a_i) = 1$
- $P(A = a_i) = \sum_j P(A = a_i, B = b_j)$ $a$ is a value assignment for variable $A$

Probability table

<table>
<thead>
<tr>
<th>Grade</th>
<th>Honor-Student</th>
<th>$P(G,H)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>False</td>
<td>0.05</td>
</tr>
<tr>
<td>B</td>
<td>False</td>
<td>0.05</td>
</tr>
<tr>
<td>C</td>
<td>False</td>
<td>0.05</td>
</tr>
<tr>
<td>D</td>
<td>False</td>
<td>0.1</td>
</tr>
<tr>
<td>A</td>
<td>True</td>
<td>0.3</td>
</tr>
<tr>
<td>B</td>
<td>True</td>
<td>0.2</td>
</tr>
<tr>
<td>C</td>
<td>True</td>
<td>0.15</td>
</tr>
<tr>
<td>D</td>
<td>True</td>
<td>0.1</td>
</tr>
</tbody>
</table>
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” \(f(x)\)

\[
f(x) = \frac{1}{\sigma \sqrt{2 \pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
\]

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

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

\[\mu = 0, \sigma^2 = 1.0\]

Which color is \(\mu=-2, \sigma^2=0.5\)? Which color is \(\mu=0, \sigma^2=0.2\)?

\[N(\mu_1, \sigma_1^2) + N(\mu_2, \sigma_2^2) = N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)\]
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:

- \( \frac{d}{dx} x^a = a x^{a-1} \)
- \( \frac{d}{dx} kf(x) = k f'(x) \)
- \( \frac{d}{dx} [f(x) + g(x)] = f'(x) + g'(x) \)

\[ f'(x) = 2x - 5 \]

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:

- \( \frac{d}{dx} f(w) = 0 \) -- \( w \) 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: \( x^{thresh} \)

Example function:

\[
C(x) = \begin{cases} 
  True & \text{if } x > x^{thresh} \\
  False & \text{otherwise}
\end{cases}
\]

Learning our classifier parameter(s)

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

The testing set

- Does classifier correctly label new data?

Testing set should be distinct from training set!

Example “good” performance: 90% correct labels
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

![Graph showing train and test error over time spent training](image)