10-701

Machine Learning

http://www.cs.cmu.edu/~10701/index.html
Organizational info

• All up-to-date info is on the course web page:
  http://www.cs.cmu.edu/~10701/index.html

Instructor:
  - Ziv Bar-Joseph

• TAs:
  - Byron Boots
  - Hai-Son Le
  - Yang Xu

• See web page for contact info, office hours, etc.

• Mailing list – if not received email or to unsubscribe:
  send email to: hple+@cs.cmu.edu
Intro to ML and probability
1/23 - Intro to ML and probability
1/25 - Density estimation
1/30 - Classification theory, KNN
2/01 - Bayes and Naïve Bayes classifiers / PS1 out
2/06 - Linear regression
2/08 - ML in industry
2/13 - Logistic regression
2/15/12 - Learning theory 1 PS1 in / PS2 out
2/20/12 - Learning theory 2
2/22/12 - Decision trees
2/27/12 - SVM1
2/29/12 - SVM2 / PS2 in, PS3 out

3/21 (Wednesday): Midterm
(1:30-3:30)

3/21, 3/26 - Model and feature selection / project proposals
3/28 - Bayesian networks 1
4/02 - BN 2
4/04 - HMMs
4/09 - HMM2 structure learning / PS4 in PS5 out
4/11 MDPs
4/16 - RL
4/18 - undirected graphical models
5/02 (Wednesday): Poster session
in the afternoon (no class)
5/02 - poster session

Intro and classification
(A.K.A. ‘supervised learning’)

Clustering
(‘Unsupervised learning’)

Probabilistic representation
and modeling (‘reasoning
under uncertainty’)

Applications
of ML
Grading

- 5 Problem sets - 30%
- Project - 20%
- Midterm - 20%
- Final - 25%
- Class participation - 5%
Class assignments

• 5 Problem sets
  - Each containing both theoretical and programming assignments

• Projects
  - Groups of 1 or 2
  - Implement and apply an algorithm discussed in class to a new domain
  - Extend algorithms discussed in class in various directions
  - New theoretical results (for example, for a new setting of a problem)
  - More information on website

• Recitations
  - Monday, 6-7:20pm, NSH 1305
  - Expand on material learned in class, go over problems from previous classes etc.
What is Machine Learning?

Easy part: Machine
Hard part: Learning

- Short answer: Methods that can help generalize information from the observed data so that it can be used to make better decisions in the future
What is Machine Learning?

Longer answer: The term Machine Learning is used to characterize a number of different approaches for generalizing from observed data:

• Supervised learning
  - Given a set of features and labels learn a model that will predict a label to a new feature set

• Unsupervised learning
  - Discover patterns in data

• Reasoning under uncertainty
  - Determine a model of the world either from samples or as you go along

• Active learning
  - Select not only model but also which examples to use
Paradigms of ML

- **Supervised learning**
  - Given $D = \{X_i, Y_i\}$ learn a model (or function) $F: X_k \rightarrow Y_k$

- **Unsupervised learning**
  - Given $D = \{X_i\}$ group the data into $Y$ classes using a model (or function) $F: X_i \rightarrow Y_j$

- **Reinforcement learning (reasoning under uncertainty)**
  - Given $D = \{\text{environment, actions, rewards}\}$ learn a policy and utility functions:
    - policy: $F1: \{e,r\} \rightarrow a$
    - utility: $F2: \{a,e\} \rightarrow R$

- **Active learning**
  - Given $D = \{X_i, Y_i\}, \{X_j\}$ learn a function $F1: \{X_j\} \rightarrow x_k$ to maximize the success of the supervised learning function $F2: \{X_i, x_k\} \rightarrow Y$
Web search

Turing Center KnowItAll Project

TextRunner Search

TextRunner searches hundreds of millions of assertions extracted from over 100 million Web pages on the topics of nutrition, history of science, and general knowledge, and sorts the results by probability.

Our IJCAI '07 paper on TextRunner is here: [Open Information Extraction from the Web](http://www.cs.washington.edu/research/textrunner/)

Example quotes:

"What did Thomas Edison invent?"
"What kills bacteria?"
"Johannes Kepler"

Search individual fields:

Argument 1

Predicate

Argument 2

Search

Search query:

[Search]

*questions/comments/bugs*

Show advanced search options

Primarily supervised learning
Web search, cont’d

Primarily supervised learning
Recommender systems are primarily supervised learning systems.
Semi supervised learning

At present, NELL has accumulated a knowledge base of 967,123 beliefs that it has read from various web pages. It is not perfect, but NELL is learning. You can track NELL’s progress below or @cmunell on Twitter, browse and download its knowledge base, read more about our technical approach, or join the discussion group.
Grand and Urban Challenges road race

Supervised and reinforcement learning
Helicopter control

Reinforcement learning
Biology

Which part is the gene?

Supervised and unsupervised learning (can also use active learning)
Most investors trying the approach are using "machine learning," a branch of artificial intelligence in which a computer program analyzes huge chunks of data and makes predictions about the future.
Common Themes

• Mathematical framework
  - Well defined concepts based on explicit assumptions
• Representation
  - How do we encode text? Images?
• Model selection
  - Which model should we use? How complex should it be?
• Use of prior knowledge
  - How do we encode our beliefs? How much can we assume?
(brief) intro to probability
• Random variable
  - referring to an element / event whose status is unknown:
    \[ A = \text{“it will rain tomorrow”} \]
• Domain (usually denoted by \( \Omega \))
  - The set of values a random variable can take:
    - \( A = \text{“The stock market will go up this year”} \): Binary
    - \( A = \text{“Number of Steelers wins in 2012”} \): Discrete
    - \( A = \text{“% change in Google stock in 2012”} \): Continuous
Axioms of probability (Kolmogorov’s axioms)

A variety of useful facts can be derived from just three axioms:

1. $0 \leq P(A) \leq 1$
2. $P(\text{true}) = 1$, $P(\text{false}) = 0$
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

There have been several other attempts to provide a foundation for probability theory. Kolmogorov’s axioms are the most widely used.
Using the axioms

- How can we use the axioms to prove that:
  \[ P(\neg A) = 1 - P(A) \]
Priors

Degree of belief in an event in the absence of any other information

P(rain tomorrow) = 0.2
P(no rain tomorrow) = 0.8
Conditional probability

- $P(A = 1 \mid B = 1)$: The fraction of cases where $A$ is true if $B$ is true
Conditional probability

- In some cases, given knowledge of one or more random variables we can improve upon our prior belief of another random variable
- For example:

  \[ p(\text{slept in movie}) = 0.5 \]
  \[ p(\text{slept in movie | liked movie}) = 1/4 \]
  \[ p(\text{didn’t sleep in movie | liked movie}) = 3/4 \]

<table>
<thead>
<tr>
<th>Slept</th>
<th>Liked</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Joint distributions

- The probability that a set of random variables will take a specific value is their joint distribution.
- Notation: $P(A \land B)$ or $P(A,B)$
- Example: $P(\text{liked movie, slept})$

If we assume independence then

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

However, in many cases such an assumption maybe too strong (more later in the class)
Joint distribution (cont)

P(class size > 20) = 0.6
P(summer) = 0.4

P(class size > 20, summer) = ?

Evaluation of classes

<table>
<thead>
<tr>
<th>Size</th>
<th>Time</th>
<th>Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>R</td>
<td>2</td>
</tr>
<tr>
<td>70</td>
<td>R</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>S</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>S</td>
<td>3</td>
</tr>
<tr>
<td>56</td>
<td>R</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td>S</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>S</td>
<td>3</td>
</tr>
<tr>
<td>23</td>
<td>R</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>R</td>
<td>2</td>
</tr>
<tr>
<td>45</td>
<td>R</td>
<td>1</td>
</tr>
</tbody>
</table>
Joint distribution (cont)

\[ P(\text{class size} > 20) = 0.6 \]
\[ P(\text{summer}) = 0.4 \]
\[ P(\text{class size} > 20, \text{summer}) = 0.1 \]

Evaluation of classes

<table>
<thead>
<tr>
<th>Size</th>
<th>Time</th>
<th>Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>R</td>
<td>2</td>
</tr>
<tr>
<td>70</td>
<td>R</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>S</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>S</td>
<td>3</td>
</tr>
<tr>
<td>56</td>
<td>R</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td>S</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>S</td>
<td>3</td>
</tr>
<tr>
<td>23</td>
<td>R</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>R</td>
<td>2</td>
</tr>
<tr>
<td>45</td>
<td>R</td>
<td>1</td>
</tr>
</tbody>
</table>
Joint distribution (cont)

$P(\text{class size} > 20) = 0.6$
$P(\text{eval} = 1) = 0.3$
$P(\text{class size} > 20, \text{eval} = 1) = 0.3$

<table>
<thead>
<tr>
<th>Size</th>
<th>Time</th>
<th>Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>R</td>
<td>2</td>
</tr>
<tr>
<td>70</td>
<td>R</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>S</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>S</td>
<td>3</td>
</tr>
<tr>
<td>56</td>
<td>R</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td>S</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>S</td>
<td>3</td>
</tr>
<tr>
<td>23</td>
<td>R</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>R</td>
<td>2</td>
</tr>
<tr>
<td>45</td>
<td>R</td>
<td>1</td>
</tr>
</tbody>
</table>
Joint distribution (cont)

P(class size > 20) = 0.6
P(eval = 1) = 0.3
P(class size > 20, eval = 1) = 0.3

Evaluation of classes

<table>
<thead>
<tr>
<th>Size</th>
<th>Time</th>
<th>Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>R</td>
<td>2</td>
</tr>
<tr>
<td>70</td>
<td>R</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>S</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>S</td>
<td>3</td>
</tr>
<tr>
<td>56</td>
<td>R</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td>S</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>S</td>
<td>3</td>
</tr>
<tr>
<td>23</td>
<td>R</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>R</td>
<td>2</td>
</tr>
<tr>
<td>45</td>
<td>R</td>
<td>1</td>
</tr>
</tbody>
</table>
Chain rule

• The joint distribution can be specified in terms of conditional probability:

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

• Together with Bayes rule (which is actually derived from it) this is one of the most powerful rules in probabilistic reasoning.
Bayes rule

• One of the most important rules for this class.
• Derived from the chain rule:
  \[ P(A,B) = P(A \mid B)P(B) = P(B \mid A)P(A) \]
• Thus,

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

Thomas Bayes was an English clergyman who set out his theory of probability in 1764.
Bayes rule (cont)

Often it would be useful to derive the rule a bit further:

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

This results from:
\[P(B) = \sum_A P(B, A)\]
Using Bayes rule

- Cards game:

Place your bet on the location of the King!
Using Bayes rule

- Cards game:

Do you want to change your bet?
Using Bayes rule

Computing the (posterior) probability: \( P(C = k \mid \text{selB}) \)

\[
P(C = k \mid \text{selB}) = \frac{P(\text{selB} \mid C = k)P(C = k)}{P(\text{selB})}
\]

\[
= \frac{P(\text{selB} \mid C = k)P(C = k)}{P(\text{selB} \mid C = k)P(C = k) + P(\text{selB} \mid C = 10)P(C = 10)}
\]
Using Bayes rule

\[ P(C=k \mid \text{selB}) = \frac{P(\text{selB} \mid C = k)P(C = k)}{P(\text{selB} \mid C = k)P(C = k) + P(\text{selB} \mid C = 10)P(C = 10)} \]

\[ = \frac{1/2 \times 1/3}{1/2 \times 1/3 + 1/2 \times 2/3} \]

\[ = \frac{1}{3} \]
Probability Density Function

- Discrete distributions

\[ \sum_{i} P(X = x_i) = 1 \]

- Continuous: Cumulative Density Function (CDF): \( F(a) \)

\[ P(x \leq a) = \int_{-\infty}^{a} f(\tau) d\tau \]
Cumulative Density Functions

- Total probability

\[ P(\Omega) = \int_{-\infty}^{\infty} f(x) \, dx = 1 \]

- Probability Density Function (PDF)

\[ \frac{d}{dx} F(x) = f(x) \]

- Properties:

\[ P(a \leq x \leq b) = \int_{b}^{a} f(x) \, dx = F(b) - F(a) \]

\[ \lim_{x \to -\infty} F(x) = 0 \]

\[ \lim_{x \to \infty} F(x) = 1 \]

\[ F(a) \geq F(b) \quad \forall a \geq b \]
Expectations

• Mean/Expected Value:
  \[ E[x] = \bar{x} = \int x f(x) \, dx \]

• Variance:
  \[ Var(x) = E[(x - \bar{x})^2] = E[x^2] - (\bar{x})^2 \]

• In general:
  \[ E[x^2] = \int x^2 f(x) \, dx \]
  \[ E[g(x)] = \int g(x) f(x) \, dx \]
Multivariate

- Joint for \((x, y)\)

\[
P((x, y) \in A) = \int \int_A f(x, y) \, dx \, dy
\]

- Marginal:

\[
f(x) = \int f(x, y) \, dy
\]

- Conditionals:

\[
f(x|y) = \frac{f(x, y)}{f(y)}
\]

- Chain rule:

\[
f(x, y) = f(x|y)f(y) = f(y|x)f(x)
\]
Bayes Rule

- Standard form:

\[ f(x|y) = \frac{f(y|x)f(x)}{f(y)} \]

- Replacing the bottom:

\[ f(x|y) = \frac{f(y|x)f(x)}{\int f(y|x)f(x)\,dx} \]
Binomial

- Distribution:

\[ x \sim \text{Binomial}(p, n) \]

\[ P(x = k) = \binom{n}{k} p^k (1 - p)^{n-k} \]

- Mean/Var:

\[ E[x] = np \]

\[ Var(x) = np(1 - p) \]
Uniform

• Anything is equally likely in the region $[a,b]$

• Distribution:

$$x \sim U(a, b)$$

• Mean/Var

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

$$E[x] = \frac{a + b}{2}$$

$$Var(x) = \frac{a^2 + ab + b^2}{3}$$
Gaussian (Normal)

- If I look at the height of women in country xx, it will look approximately Gaussian
- Small random noise errors, look Gaussian/Normal

- Distribution:
  \[ x \sim N(\mu, \sigma^2) \]
  \[ f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]
- Mean/var
  \[ E[x] = \mu \]
  \[ Var(x) = \sigma^2 \]
Why Do People Use Gaussians

• Central Limit Theorem: (loosely)
  - Sum of a large number of IID random variables is approximately Gaussian
Multivariate Gaussians

- Distribution for vector $x$

$$x = (x_1, \ldots, x_N)^T, \quad x \sim N(\mu, \Sigma)$$

- PDF:

$$f(x) = \frac{1}{(2\pi)^{N/2} |\Sigma|^{1/2}} e^{-\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu)}$$

$$E[x] = \mu = (E[x_1], \ldots, E[x_N])^T$$

$$\text{Var}(x) \rightarrow \Sigma = \begin{pmatrix}
\text{Var}(x_1) & \text{Cov}(x_1, x_2) & \cdots & \text{Cov}(x_1, x_N) \\
\text{Cov}(x_2, x_1) & \text{Var}(x_2) & \cdots & \text{Cov}(x_2, x_N) \\
\vdots & \ddots & \ddots & \vdots \\
\text{Cov}(x_N, x_1) & \text{Cov}(x_N, x_2) & \cdots & \text{Var}(x_N)
\end{pmatrix}$$
Multivariate Gaussians

\[ f(x) = \frac{1}{(2\pi)^{N/2} |\Sigma|^{1/2}} e^{-\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu)} \]

\[ E[x] = \mu = (E[x_1], \ldots, E[x_N])^T \]

\[ Var(x) \rightarrow \Sigma = \begin{pmatrix}
Var(x_1) & Cov(x_1, x_2) & \cdots & Cov(x_1, x_N) \\
Cov(x_2, x_1) & Var(x_2) & \cdots & Cov(x_2, x_N) \\
\vdots & \vdots & \ddots & \vdots \\
Cov(x_N, x_1) & Cov(x_N, x_2) & \cdots & Var(x_N)
\end{pmatrix} \]

\[ \text{cov}(x_1, x_2) = \frac{1}{n} \sum_{i=1}^{n} (x_{1,i} - \mu_1)(x_{2,i} - \mu_2) \]
Covariance examples

Anti-correlated

Covariance: -9.2

Correlated

Covariance: 18.33

Independent (almost)

Covariance: 0.6
Sum of Gaussians

• The sum of two Gaussians is a Gaussian:

\[ x \sim N(\mu, \sigma^2) \quad y \sim N(\mu_y, \sigma_y^2) \]

\[ ax + b \sim N(a\mu + b, (a\sigma)^2) \]

\[ x + y \sim N(\mu + \mu_y, \sigma^2 + \sigma_y^2) \]
Important points

- Random variables
- Chain rule
- Bayes rule
- Joint distribution, independence, conditional independence