Introduction

Le Song

Machine Learning II: Advanced Topics
CSE 8803ML, Spring 2012
What is Machine Learning (ML)

- Study of algorithms that improve their performance at some task with experience
Common to Industrial scale problems

13 million wikipedia pages

800 million users

6 billion photos

24 hours video uploaded per minutes
Simple Machine Learning Problems
Classification and Regression

Numeric values:
- 40 F
- Wind: NE at 14 km/h
- Humidity: 83%
Clustering

Image Databases
Example training images for each orientation
Similarity Search

Given an image, find similar ones
Dimensionality reduction

Each image has thousands or millions of pixels.

Can we give each image a coordinate, such that similar images are near each other?
More Complicated Machine Learning Problems
According to a study at Cambridge University, it doesn’t matter in what order the letters in a word are, the only important thing is that the first and last letter be at the right place. The rest can be a total messes and you can still read it without a problem. This is because the human mind does not read every letter by itself, but the word as a whole.
"Machine Learning is the preferred method for speech recognition ..."
Future stock price may depend on previous prices, and the price histories of many similar stocks.
Spam Filtering

Warning: This message may not be from whom it claims to be. Beware of follow-up links in it or of providing the sender with any personal information. Learn more

From:
Office of United Nations General Assembly.

I write to inform you that your e-mail address is among the lucky addresses approved. The sum of ($850,000.00 British pounds) will be transferred to your account. This is because the United Nation is collaborating with the refugees beneficiary council are supporting individuals due to the worldwide economy meltdown.

Contact Person: Mr. Gary Epps
United Nations General Assembly
E-mail: gary.epps@un.org

You are requested to him the following information below:
Your full Name
Residential Address
Office Address
Occupation
Age
Sex
Country
Phone number
Scanned valid ID which may be either your
Webpage Classification

Company homepage vs. University homepage

Professor homepage vs. Postdoc homepage
Webpage classification
Image Classification
What are the fundamental questions in graphical models

- **Representation:**
  - Graphical models represent exponentially large probability distribution compactly
  - What are the types of models?
  - What does the model mean/imply/assume? (semantics)
  - Key concept: conditional independence

- **Inference:**
  - How do we answer questions/query with the model?
  - What is the probability of X given some observations?
  - What is the most likely explanation for what is happening?
  - What decision should I make?

- **Learning**
  - What are the right/good parameters for the model?
  - How do I obtain the structure of the model?
Where do we start

- **Representation:**
  - From directed graphical models/Bayesian networks
  - Undirected graphical models (Markov random fields, conditional random fields)

- **Inference:**
  - Exact inference (message passing, junction tree algorithms)
  - Approximate Inference (loopy belief propagation, sampling, variational inference)

- **Learning:**
  - Parameter learning
  - Structure learning (Chow-liu tree, graphical lasso)

- **Example graphical model**
  - Gaussian mixture, Hidden Markov models
  - Latent Dirichlet Allocation
  - Collaborative filtering
Learning Problems with nonlinear relations and patterns
Nonlinear classifier

Nonlinear Decision Boundaries

Linear SVM Decision Boundaries
Nonlinear regression and want to estimate the variance
Need advanced methods such as Gaussian processes and kernel regression
Nonconventional clusters

Need more advanced methods, such as kernel methods or spectral clustering to work
Nonlinear principal component analysis

PCA

Nonlinear PCA
Predict Depth from Image Features

Skewed and multimodal continuous depth data

Histogram
nonparametric
Gaussian

Depth (meters in log10 scale)

farther away

Spatial dependency

depth variables

image pixel variables

[Saxena, Chung and Ng 2005]
Protein Structure Prediction

- Met
- Leu
- Val
- His
- Ser
- His

Amino-acid sequence variables

Angular variables

adjacent angle not differ too much

multimodal non-Euclidean data

Ramachandran Plot

[Boomsma et al. 08]
What are the fundamental questions in kernel

- Design kernels:
  - Kernels make designing algorithm and incorporating domain specific knowledge two separate modules
  - What is a kernel?
  - How to design kernels for different data types?
  - How to construct more sophisticated kernels based on simple kernels?

- Algorithms with kernels:
  - How to design algorithms with kernels to capture nonlinear pattern?

- Efficiency
  - How to make kernel methods work for large datasets?
Where do we start

- **Kernels:**
  - Linear kernel, polynomial kernel, Gaussian RBF kernel
  - Rules to combine kernels
  - Kernels for sequences, graphs, images

- **Kernel algorithms**
  - Kernel Support Vector Machine, Kernel ridge regression
  - Kernel PCA and clustering
  - Kernel canonical correlation analysis
  - Kernel and graphical models: Gaussian processes

- **Fast kernel methods**
  - Incomplete Cholesky decomposition
  - Random feature for Gaussian RBF kernel
Syllabus

- Cover several advanced machine learning topics, eg., graphical models, kernel methods, boosting, bagging and tensor data analysis ...

- The focus of the class will be graphical models and kernel methods.
  - Graphical models provide a unified view for a wide range of problems with a very large number of attributes.
  - Kernel methods provide a general framework for designing algorithms for finding nonlinear relations and patterns.
  - Mixture models, HMM, Kalman filters, Markov random fields, conditional random fields, LDA, collaborative filtering, kernel SVM, kernel PCA, spectral clustering, kernel canonical correlations analysis, sequence and graph comparison
Prerequisites

- Probabilities
  - Distributions, densities, marginalization, conditioning ....
- Basic statistics
  - Moments, classification, regression, maximum likelihood estimation
- Algorithms
  - Dynamic programming, basic data structures, complexity
- Programming
  - Mostly your choice of language, but Matlab will be very useful

We provide some background, but the class will be fast paced

Ability to deal with “abstract mathematical concepts

Ideally you’ve taken Machine Learning I (FALL 2011: CS 7641, CSE 6740, ISyE 6740)
Textbooks

- Required Textbook:
  - Pattern Recognition and Machine Learning, Chris Bishop

- Secondary Textbook:
  - Kernel Methods for Pattern Analysis, John Shawe-Taylor & Nello Cristianini

- Other useful books
  - Gaussian Processes for Machine Learning, Carl Edward Rasmussen & Christopher Williams
  - Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, Bernhard Scholkopf and Alex Smola
  - Probabilistic Graphical Models: Principles and Techniques, Daphne Koller & Nir Friedman
Grading

- 3 homeworks (30%)
  - First one goes out end of Jan
  - Second one out end of Feb
  - Third one out end of Mar
  - Start early

- Final exam (30%)

- Project (40%)
  - Details out around Feb 9th
  - Projects done individually, or groups of two students
  - Need to submit a project report, give a 15-minute project presentation, and a poster
Homeworks

- Due in the beginning of class
  - Half credit within 48 hours
  - Zero credit after 48 hours

- All homeworks must be handed in, even for zero credit

Collaboration

- You may discuss the questions
- Each student writes their own answers
- Write on your homework anyone with whom you collaborate
- Each student must write their own codes for the programming part
Staff

- Instructor: Le Song, Klaus 1340
- TA: Krishnakumar Balasubramanian, Klaus 1305
- Guest Lecturer: Mariya Ishteva, Klaus 1336
- Administrative Assistant: Michael Terrell, Klaus 1321

Mailing list: cse8803ml2@googlegroups.com

More information:
http://www.cc.gatech.edu/~lsong/teaching/8803ML.html
Today

- Probabilities
- Independence
- Conditional Independence
Random Variables (RV)

- Data may contain many different attributes
  - Age, grade, color, location, coordinate, time ...
- Random variable formalize attributes
- Upper-case for rv (eg. $X, Y$), lower-case for values (eg. $x, y$)
- $P(X)$ for distribution, $p(X)$ for density

Properties of random variable $X$:
- $Val(X)$ = possible values of random variable $X$
- For discrete (categorical): $\sum_{i=1\ldots|Val(x)|} P(X = x_i) = 1$
- For continuous: $\int_{Val(x)} p(X = x) dx = 1$
  - $P(x) \geq 0$
- Shorthand: $P(x)$ for $P(X = x)$
Interpretations of probability

- **Frequentists**
  - \( P(x) \) is the frequency of \( x \) in the limit
  - Many arguments against this interpretation
    - What is the frequency of the event “it will rain tomorrow?”

- **Subjective interpretation**
  - \( P(x) \) is my degree of belief that \( x \) will happen
  - What does “degree of belief mean”?
  - If \( P(x) = 0.8 \), then I am willing to bet

For this class, we don’t care the type of interpretation.
Conditional probability

After we have seen $x$, how do we feel $y$ will happen?

$P(y|x)$ means $P(Y = y | X = x)$

A conditional distribution are a family of distributions

- For each $X = x$, it is a distribution $P(Y|x)$
Two of the most important rules: I. The chain rule

- \( P(y, x) = P(y|x)P(x) \)

More generally:

- \( P(x_1, x_2, \ldots, x_k) = P(x_1)P(x_2|x_1) \ldots P(x_k|x_{k-1}, \ldots, x_2, x_1) \)
Two of the most important rules: II. Bayes rule

\[ P(y|x) = \frac{P(x|y)P(y)}{P(x)} = \frac{P(x,y)}{\sum_{x \in \text{Val}(x)} P(x,y)} \]

- likelihood
- Prior
- posterior
- Normalization constant

More generally, additional variable z:

\[ P(y|x,z) = \frac{P(x|y,z)P(y|z)}{P(x|z)} \]
Most important concept: independence

- $X$ and $Y$ independent, if $P(Y|X) = P(Y)$
  - $P(Y|X) = P(Y) \rightarrow (X \perp Y)$

Proposition: $X$ and $Y$ independent if and only if $P(X, Y) = P(X)P(Y)$
Most important concept: conditional independence

- Independence is rarely true; conditional independence is more prevalent

- $X$ and $Y$ conditionally independent given $Z$ if
  \[ P(Y|X,Z) = P(Y|Z) \]
  \[ P(Y|X,Z) = P(Y|Z) \rightarrow (X \perp Y | Z) \]

- $(X \perp Y | Z)$ if and only if $P(X,Y|Z) = P(X|Z)P(Y|Z)$
Joint distribution, Marginalization

- Two random variables – Grades (G) & Intelligence (I)
  
  \[ P(G, I) = \]

<table>
<thead>
<tr>
<th>G</th>
<th>I</th>
<th>VH</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.7</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>B</td>
<td>0.15</td>
<td>0.05</td>
<td>0.05</td>
</tr>
</tbody>
</table>

- For \( n \) binary variables, the table (multiway array) gets really big
  
  \[ P(X_1, X_2, \ldots, X_n) \] has \( 2^n \) entries!

- Marginalization – Compute marginal over a single variable
  
  \[ P(G = B) = P(G = B, I = VH) + P(G = B, I = H) = 0.2 \]
Marginalization – the general case

- Compute marginal distribution $P(X_i)$ from $P(X_1, X_2, ..., X_i, X_{i+1}, ..., X_n)$

$$P(X_1, X_2, ..., X_i) = \sum_{x_{i+1},...,x_n} P(X_1, X_2, ..., X_i, x_{i+1}, ..., x_n)$$

$$P(X_i) = \sum_{x_1,...,x_{i-1}} P(x_1, ..., x_{i-1}, X_i)$$

If binary variables, need to sum over $2^{n-1}$ terms!
Summary: basic concepts for R.V.

- Outcome: assign $x_1, ..., x_n$ to $X_1, ... X_n$

- Conditional probability: $P(X, Y) = P(X) P(Y|X)$

- Bayes rule: $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$

- Chain rule:
  
  $P(X_1, ..., X_n) = P(X_1)P(X_2|X_1) ... P(X_n|X_1, ..., X_{n-1})$
Summary: conditional independence

- $X$ is independent of $Y$ given $Z$ if
  \[ P(X = x | Y = y, Z = z) = P(X = x | Z = z) \]
  \[ \forall x \in \text{Val}(X), y \in \text{Val}(Y), z \in \text{Val}(Z) \]

- Shorthand:
  - $(X \perp Y | Z)$
  - For $(X \perp Y | \emptyset)$, write $(X \perp Y)$

- Proposition: $(X \perp Y | Z)$ if and only if
  \[ P(X, Y | Z) = P(X | Z)P(Y | Z) \]
Further properties of conditional independence

- Symmetry:
  \( (X \perp Y \mid Z) \Rightarrow (Y \perp X \mid Z) \)

- Decomposition:
  \( (X \perp Y, W \mid Z) \Rightarrow (X \perp Y \mid Z) \)

- Weak union:
  \( (X \perp Y, W \mid Z) \Rightarrow (X \perp Y \mid Z, W) \)

- Contraction:
  \( (X \perp W \mid Y, Z) \& (X \perp Y \mid Z) \Rightarrow (X \perp Y, W \mid Z) \)

- Intersection:
  \( (X \perp Y \mid W, Z) \& (X \perp W \mid Y, Z) \Rightarrow (X \perp Y, W \mid Z) \)

- Only for strictly positive distributions!

\( P(x) > 0, \forall x \)