About myself

Assistant Professor at U. of Southern California, Computer Science Department

Research interests: statistical machine learning, with applications to language processing, computer visions, etc

About Tutorial

Focus on important concepts, paradigms, models and algorithms

Conceptual level understanding

Main goal is to inspire your interests

Feel free to chat with us or contact us at lebanon@cc.gatch.edu and feisha@usc.edu
Outline

Themes

Estimate parameters

How to learn models automatically from data?

Evaluate and select models

How to get good models from data?

Advanced techniques

Identify hidden patterns and structures

Nonlinear and probabilistic approaches

Model sequential/temporal data

Hidden Markov models
Outline

Themes

Estimate parameters

How to learn models automatically from data?

Evaluate and select models

How to get good models from data?

Advanced techniques

Identify hidden patterns and structures

Nonlinear and probabilistic approaches

Model sequential/temporal data

Hidden Markov models
Learning models from data

Many statistical models

- Parametric: eg. linear and logistic regression, etc
- Non-parametric: kNN, support vector machines, etc

How to specify parameters in these models?

- Manually: cumbersome and does not scale up
- Desideratum: estimate automatically from data!

Parameter estimation techniques

- Maximum likelihood estimation
- Bayesian inference
Maximum likelihood estimation (MLE)

Intuitive example

Estimate a coin toss

I have seen 3 flips of heads, 2 flips of tails, what is the chance of head (or tail) of my next flip?

Model

Each flip is a Bernoulli random variable $X$.

$X$ can take only two values: 1 (head), 0 (tail)

$$p(X = 1) = \theta$$  $$p(X = 0) = 1 - \theta$$

Parameter to be identified from data
Principles of MLE

N (independent) trials

Observation (often called the training data)

\[ X_1 = 1 \quad X_2 = 0 \quad X_3 = 1 \quad X_4 = 1 \quad X_5 = 0 \]

Likelihood of all the \( N \) observations

\[
\theta \times (1 - \theta) \times \theta \times \theta \times (1 - \theta) = \mathcal{L} = \theta^3 (1 - \theta)^2
\]

Intuition

choose \( \theta \) such that \( \mathcal{L} \) is maximized
Maximizing the likelihood

Solution

\[ L = \theta^3 (1 - \theta)^2 \]

\[ \theta^{MLE} = \frac{3}{3 + 2} \]

maximizing \( L \) is analytically solvable

Intuition

Probability of head is the percentage of heads in the total flips.
More generally,

**Model**

\[ X \sim P(X; \theta) \]

**Training data**

\[ \mathcal{D} = \{x_1, x_2, \cdots, x_N\} \]

**Maximum likelihood estimate**

\[
\mathcal{L}(\mathcal{D}) = \prod_{i=1}^{N} P(x_i; \theta) \quad \theta^{MLE} = \arg \max_{\theta} \mathcal{L}(\mathcal{D}) \\
= \arg \max_{\theta} \sum_{i=1}^{N} \log P(x_i; \theta)
\]
Example: model of continuous random variables

Predict the sale price of a house

Model

Linear regression: Sale Price = a × SqFt + b + noise

assume Gaussian distributed $N(0, \sigma^2)$
MLE of \((a, b)\)

**Training data (sale record)**

\[
D = (S^{(1)}, F^{(1)}), (S^{(2)}, F^{(2)}), \ldots, (S^{(N)}, F^{(N)})
\]

**Likelihood**

A single observation \((S, F)\):

\[
p(S|F) = \exp\left\{-\frac{(S - (aF + b))^2}{\sigma^2}\right\}
\]

For all training data

\[
\log \mathcal{L}(D) = -\sum_{n=1}^{N} (S^{(n)} - aF^{(n)} - b)^2 + \text{const}
\]

Maximize this

minimize this, which is prediction error
By the way, this model works

\[
\text{Price (in K)} \approx 0.22 \times \text{footage} + 467
\]
Caveats for complicated models

No closed-form solution

Use numerical optimization

many easy-to-use, robust packages are available

Stuck in local optimum

Restart optimization with random initialization

Computational tractability

Difficult to compute likelihood $\mathcal{L}(D)$ exactly

Need to approximate
What if you know something about the coin before tossing?

Prior knowledge
- coin is highly biased
- coin is likely to “tail”
- coin is likely to “head”

Incorporate this knowledge

Why?
- for instance, reduce the number of tosses we would need

How?
- Bayesian inference
Representing prior knowledge

Bayesian prior

We use a Beta distribution for the coin

The distribution Beta\((a, b)\) is given by

\[
p(\theta; a, b) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1 - \theta)^{b-1}
\]

\(a \) and \(b \) control “biases”

- \(a = 2\), \(b = 3\): coin biased to “tail”
- \(a = 8\), \(b = 4\): coin biased to “head”

- \(a = 0.1\), \(b = 0.1\): coin is highly biased.

\(\theta\)
Update our prior

Observed data can support or against our prior

Intuition

Our prior is that the coin is fair

We have also observed 5 tosses

How should be change our belief?

$$p(\theta; a, b)$$
Update our prior

Observed data can support or against our prior

Intuition

Our prior is that the coin is fair

We have also observed 5 tosses

How should be change our belief?

$p(\theta; a, b)$
Update our prior

Observed data can support or against our prior

Intuition

Our prior is that the coin is fair
We have also observed 5 tosses
How should we change our belief?

\[ \mathcal{L}(D; \theta) \]
Update our prior

Observed data can support or against our prior

Intuition

Our prior is that the coin is fair

We have also observed 5 tosses

How should be change our belief?

\[ p(\theta; a, b) \]

\[ L(\mathcal{D}; \theta) \]
Update our prior

Observed data can support or against our prior

Intuition

Our prior is that the coin is fair

We have also observed 5 tosses

How should be change our belief?

\[
p(\theta; a, b) \times L(D; \theta) = p(\theta|D; a, b)
\]
Make our intuition precise

Use Bayes rule to compute posterior

\[ p(\theta | \mathcal{D}; a, b) = \frac{p(\theta; a, b)p(\mathcal{D} | \theta)}{p(\mathcal{D}; a, b)} \]

prior likelihood

marginal probability

\[ p(\mathcal{D}; a, b) = \int p(\theta; a, b)p(\mathcal{D} | \theta) \, d\theta \]

often difficult to compute

Ex: coin toss

\[ p(\theta | \mathcal{D}; a, b) = \text{Beta} \left( a + \sum_i X_i, b + N - \sum_i X_i \right) \]
Make a prediction

Predictive distribution

Given prior, observations, what is the probability of head or tail for the next toss?

$$p(X|\mathcal{D}; a, b) = \int p(X; \theta)p(\theta|\mathcal{D}; a, b) \, d\theta$$

Ex: coin toss after seeing $N$ flip results

$$P(X = 1|\mathcal{D}; a, b) = \int P(X = 1|\theta)P(\theta|\mathcal{D}; a, b) = \frac{a + \sum_i X_i}{a + b + N}$$
When to use Bayesian inference?

Effect of training data

- lot of them: prior does not matter that much
- very limited amount of data: much more effective

Caveats

- Often, computationally intensive and intractable
- Use “tricks” to simplify computation
- Need more computational tools (beyond this tutorial)
Outline

Themes

Estimate parameters

Maximum likelihood estimation and Bayesian inference

Evaluate and select models

How to get good models from data?

Advanced techniques

Identify hidden patterns and structures

Nonlinear and probabilistic approaches

Model sequential/temporal data

Hidden Markov models and others
Outline

**Themes**

- Estimate parameters
  - Maximum likelihood estimation and Bayesian inference

**Evaluate and select models**

- How to get good models from data?

**Advanced techniques**

- Identify hidden patterns and structures
  - Nonlinear and probabilistic approaches

- Model sequential/temporal data
  - Hidden Markov models and others
Overfitting

Good news

For any training data set, we can always make a model as good as possible on it!

Our learning algorithm can just remember all the training data, and respond to questions like a database!

But, is that we want?

Bad news

In general, too complicated model does not do well on data that is unseen during the training.

This is called poor generalization or overfitting on training.
Example of overfitting

More parameters, more flexibility

Fit to training data better

But the curve becomes more “rugged”

\[ \text{Price} = a \times \text{SqFt}^3 + b \times \text{SqFt}^2 + c \times \text{SqFt} + d \]
Prevent overfitting

Model selection

- does well on training/seen data
- does well (expectedly) on unseen data

Some approaches

- Use a validation data set
- Cross-validation
- Regularization
Use a validation data set

Intuition

Validation data $D'$ should not overlap with training data $D$, used to estimate model parameters.

$D'$ is seen as as a proxy to the unseen data

How to use it?

Choose the model perform the best on $D'$, not $D$
Use cross-validation

What if we cannot afford to have a separate $D'$

Data (especially labeled ones) are costly to obtain.

We would like to use them all.

Trick

Partition data into $K$ disjoint parts

Learn model with $(K-1)$ parts

Use the one left to evaluate model

Do $K$ times, average model’s performance

Choose model with the best averaged performance
Use regularization

Explicitly force parameters to be small

Ex: regularized linear regression

\[
\sum_{n=1}^{N} (S^{(n)} - aF^{(n)} - b)^2 + \lambda a^2
\]

minimize prediction error balanced with parameter size

Intuition

\(\lambda\) is big: one prefers smaller \(a\). When \(a \to 0\) (as \(\lambda \to +\infty\)), the model’s prediction is a constant \(b\), ie, quite simple

\(\lambda\) is small: one prefers to reduce error
Outline

Themes

Estimate parameters

Evaluate and select models

Advanced techniques

Identify hidden patterns and structures

Nonlinear manifold learning

Probabilistic topic models

Model sequential/temporal data

Hidden Markov models and others
Outline

Themes

Estimate parameters

Evaluate and select models

Advanced techniques

Identify hidden patterns and structures

Nonlinear manifold learning

Probabilistic topic models

Model sequential/temporal data

Hidden Markov models and others
Recap: linear methods

Principal component analysis (PCA)

Original data in high-dimensional space

Project them to low-dimensional space

\[ \mathcal{P} : x \in \mathbb{R}^D \rightarrow z \in \mathbb{R}^d \]

\[ z = U^\top x \quad \text{linear projection with} \quad d < D \]

Nonnegative matrix factorization (NMF)

Project also linearly

\[
\begin{bmatrix}
    x^{(1)} \\
    x^{(2)} \\
    \vdots \\
    x^{(N)}
\end{bmatrix}
\approx
W
\begin{bmatrix}
    h^{(1)} \\
    h^{(2)} \\
    \vdots \\
    h^{(N)}
\end{bmatrix}
\]

\# \; \text{N data vectors} \quad \# \; \text{low-dimensional codes}
Linear methods are inadequate

Issue:
projection on any 1D line will cause mixing of data from different classes.
We need nonlinear projection

Intuition:

Find a good space through nonlinear mapping;
Then do linear projection (as in PCA)

✓ Data separated!

How to find this nonlinear mapping?
Kernel PCA (kPCA)

**Kernel trick**

Avoid explicitly finding a nonlinear mapping \( z = \phi(x) \)

Define implicitly through a kernel function

\[
K(x^{(i)}, x^{(j)}) = \phi(x^{(i)})^\top \phi(x^{(j)})
\]

Use kernel function to compute PCA

**Common kernels**

Gaussian kernel

\[
K(x^{(i)}, x^{(j)}) = \exp \left\{-\frac{\|x^{(i)} - x^{(j)}\|^2}{\sigma^2}\right\}
\]

Polynomial kernel

\[
K(x^{(i)}, x^{(j)}) = \left(1 + x^{(i)} \cdot x^{(j)}\right)^d
\]
Use of kPCA

Dimensionality reduction

Visualization

Data denoising

Extract features from data

Novelty/anomaly detection

[Heiko Hoffman, 2007]

IEEE VisWeek 2010 (Salt Lake City, Utah)
Other relevances to information visualization

Wide applicability

Kernel functions can be defined on many data types.

- strings/texts, graphs (network topology etc), etc

Does not have to be mathematical functions

- can be defined over similarity or dissimilarity measures
- can incorporate human inputs

Ex: relationships between INTERNET requests (text data) and CPU utilization (numerical data)
Another form of nonlinearity

Manifold assumption

Data lies on a smooth low-dimensional manifold

“Swiss roll”: a 2D sheet rolled in 3D space

PCA projection: linear approach failed to “unroll”
How to unroll?

The key is being myopic

- A is closer to C in Euclidean space
- on manifold (and unrolled space), A is closer to B
How to unroll?

The key is being myopic

- A is closer to C in Euclidean space

- A and D are close on both spaces

- on manifold (and unrolled space), A is closer to B
How to unroll?

The key is being myopic

A is closer to C in Euclidean space

A and D are close on both spaces

on manifold (and unrolled space), A is closer to B

Intuition: local distances are more trustworthy than global distances
from local distance to distance on manifold

To estimate global distances more accurately

Build a nearest neighbor graph

each vertex is a data point

annotate edges with trustable local distances

Distances on manifold are modeled as shortest paths on the graph

all pairwise shortest paths can be computed efficiently
from local distance to distance on manifold

To estimate global distances more accurately

Build a nearest neighbor graph

- each vertex is a data point
- annotate edges with trustable local distances

Distances on manifold are modeled as shortest paths on the graph

- all pairwise shortest paths can be computed efficiently
Unroll!

**IsoMap (Isometric Mapping) Algorithm**

- Build nearest neighbor graph
- Estimate distances on manifold (geodesic)
- Apply multidimensional scaling (MDS) to preserve the geodesic distance
- Equivalent to PCA, i.e., diagonalize a kernel matrix computed from distances
- Yield nonlinear projection into low-dimensional space
Examples

Toy examples

Swissroll

Handwritten digit

[Balasubramanian et al, 2002]

[Tenenbaum et al, 2000]
All digits

Sample images (LeCun)

Two-dimensional embedding of 60K images

[van der Matten et al, 2009]
**Example**

**Embedding music similarity**

- **Graph vertices are artists, albums, tracks**
- **Similarity determines edge length.**
- **267K vertices, 3.2M edges**
- **Use fast variants of MDS to scale up**

**Algorithm**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>% of CPU time</th>
<th>Average % of CPU time</th>
<th>Average % of CPU time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Songs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Closer</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sequential</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FSE</td>
<td>60</td>
<td>5.0%</td>
<td>52.8</td>
</tr>
<tr>
<td>LMDS</td>
<td>60</td>
<td>4.5%</td>
<td>52.7</td>
</tr>
<tr>
<td>LMDS</td>
<td>100</td>
<td>4.1%</td>
<td>87.4</td>
</tr>
<tr>
<td>LMDS</td>
<td>200</td>
<td>3.3%</td>
<td>175.0</td>
</tr>
<tr>
<td>LMDS</td>
<td>400</td>
<td>3.2%</td>
<td>355.1</td>
</tr>
</tbody>
</table>

**Table 2: Speed and accuracy of music embedding for various algorithms.**

All embeddings are 20-dimensional ($d = 20$). The CPU time was measured on a 2.4 GHz Pentium 4. FSE uses a fixed rectangle size $n = 3d$, so for the table. For the same $n$, FSE and LMDS are competitive. However, LMDS scales better for accuracy by increasing $n$.

ALaplacian Eigenmaps N/A 13.0% 8003.4

A Gaussian kernel with $\sigma = 2$ was used to convert distances to similarity for the Laplacian eigenmap. The slowness of the Laplacian eigenmap prevented extensive tuning of the parameters.

**Figure 2:** LMDS Projection of the entire music dissimilarity graph into 2D. The coordinates of 23 artists are shown.

Given that LMDS outperforms FSE for large $n$, this paper presents qualitative results from the LMDS $n = 400$ projection. First, the top two dimensions are plotted to form a visualization of music space. This visualization is shown in Figure 4.2, which shows the

[Platt, 2004]
Other manifold based learning algorithms

There are many others

Spectral methods: Locally linear embedding (LLE), Hessian LLE, etc

SDP methods: Maximum variance unfolding (MVU), etc

General framework

Build nearest neighbor graphs

Compute distances, curvatures etc

Construct embeddings

Diagonize matrix or

Solve more complicated optimization
Probabilistic approaches for identifying hidden structures

Motivation: how to model text corpus?

What are the common semantic aspects?

How to represent each piece of text in these aspects?

Do they show interesting structures?
Thinking reversely: how a piece of text is composed?

[Blei, Ng and Jordan, 2003]

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.
Generative process of a text corpus

for each document $d$

    draw a topic vector $\theta_d \sim \text{Dirichlet}(\alpha) \in R^K$

for each word $n$

    select a topic $z_{dn} \sim \text{Multinomial}(\theta_d)$

    select a word $w_{dn} \sim \text{Multinomial}(\Phi(z_{dn}, \cdot))$
Latent Dirichlet Allocation (LDA): formal definition

Generative process of a text corpus

for each document $d$

draw a topic vector $\theta_d \sim \text{Dirichlet}(\alpha) \in \mathbb{R}^K$

for each word $n$

select a topic $z_{dn} \sim \text{Multinomial}(\theta_d)$

select a word $w_{dn} \sim \text{Multinomial}(\Phi(z_{dn}, \cdot))$
Latent Dirichlet Allocation (LDA): formal definition

Generative process of a text corpus

for each document $d$

\[ \text{draw a topic vector } \theta_d \sim \text{Dirichlet}(\alpha) \in \mathbb{R}^K \]

for each word $n$

\[ \text{select a topic } z_{dn} \sim \text{Multinomial}(\theta_d) \]

\[ \text{select a word } w_{dn} \sim \text{Multinomial}(\Phi(z_{dn}, \cdot)) \]
**Latent Dirichlet Allocation (LDA): formal definition**

Generative process of a text corpus

For each document $d$

- Draw a topic vector $\theta_d \sim \text{Dirichlet}(\alpha) \in \mathbb{R}^K$

For each word $n$

- Select a topic $z_{dn} \sim \text{Multinomial}(\theta_d)$

- Select a word $w_{dn} \sim \text{Multinomial}(\Phi(z_{dn}, \cdot))$
Latent Dirichlet Allocation (LDA): formal definition

Generative process of a text corpus

for each document $d$

draw a topic vector $\theta_d \sim \text{Dirichlet}(\alpha) \in \mathbb{R}^K$

for each word $n$

select a topic $z_{dn} \sim \text{Multinomial}(\theta_d)$

select a word $w_{dn} \sim \text{Multinomial}(\Phi(z_{dn}, \cdot))$
Latent Dirichlet Allocation (LDA): formal definition

Generative process of a text corpus

for each document $d$

draw a topic vector $\theta_d \sim \text{Dirichlet}(\alpha) \in \mathbb{R}^K$

for each word $n$

select a topic $z_{dn} \sim \text{Multinomial}(\theta_d)$

select a word $w_{dn} \sim \text{Multinomial}(\Phi(z_{dn}, \cdot))$
Latent Dirichlet Allocation (LDA): formal definition

Generative process of a text corpus

for each document $d$

draw a topic vector $\theta_d \sim \text{Dirichlet}(\alpha) \in R^K$

for each word $n$

select a topic $z_{dn} \sim \text{Multinomial}(\theta_d)$

select a word $w_{dn} \sim \text{Multinomial}(\Phi(z_{dn}, \cdot))$
Inference and learning in LDA

Maximum likelihood estimation

\[ \alpha^*, \Phi^* = \arg \max \ p(w_d | \alpha, \Phi) \]

\[ = \arg \max \ \int p(\theta_d | \alpha) \left( \prod_{n=1}^{N} \sum_{z_{dn}} p(z_{dn} | \theta_d) p(w_{dn} | z_{dn}, \Phi) \right) \, d\theta_d \]

Difficulty

Intractable integral and posterior \( p(\theta, z | \alpha, \Phi, w) \)

Solution

Variational inference (Blei et al 2003)

MCMC sampling (Griffiths and Steyvers, 2004)
# LDA for analyzing Enron emails

<table>
<thead>
<tr>
<th>topic 182</th>
<th>topic 113</th>
<th>topic 23</th>
<th>topic 54</th>
<th>topic 18</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>WORD</strong></td>
<td><strong>PROB.</strong></td>
<td><strong>WORD</strong></td>
<td><strong>PROB.</strong></td>
<td><strong>WORD</strong></td>
</tr>
<tr>
<td>TEXANS</td>
<td>.0145</td>
<td>GOD</td>
<td>.0357</td>
<td>ENVIRONMENTAL</td>
</tr>
<tr>
<td>WIN</td>
<td>.0143</td>
<td>LIFE</td>
<td>.0272</td>
<td>AIR</td>
</tr>
<tr>
<td>FOOTBALL</td>
<td>.0137</td>
<td>MAN</td>
<td>.0116</td>
<td>MTBE</td>
</tr>
<tr>
<td>FANTASY</td>
<td>.0129</td>
<td>PEOPLE</td>
<td>.0103</td>
<td>EMISSIONS</td>
</tr>
<tr>
<td>SPORTSLINE</td>
<td>.0129</td>
<td>CHRIST</td>
<td>.0092</td>
<td>CLEAN</td>
</tr>
<tr>
<td>PLAY</td>
<td>.0123</td>
<td>FAITH</td>
<td>.0083</td>
<td>EPA</td>
</tr>
<tr>
<td>TEAM</td>
<td>.0114</td>
<td>LORD</td>
<td>.0079</td>
<td>PENDING</td>
</tr>
<tr>
<td>GAME</td>
<td>.0112</td>
<td>JESUS</td>
<td>.0075</td>
<td>SAFETY</td>
</tr>
<tr>
<td>SPORTS</td>
<td>.0110</td>
<td>SPIRITUAL</td>
<td>.0066</td>
<td>WATER</td>
</tr>
<tr>
<td>GAMES</td>
<td>.0109</td>
<td>VISIT</td>
<td>.0065</td>
<td>GASOLINE</td>
</tr>
</tbody>
</table>

[Rosen-Zvi et al, 2005]
### Joint author-topic analysis

**Machine Learning for Information Visualization**

**Guy Lebanon and Fei Sha**

**AUTHOR = Enron General Announcements (509 emails)**

<table>
<thead>
<tr>
<th>PROB.</th>
<th>TOPIC</th>
<th>WORDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>.9420</td>
<td>39</td>
<td>ENRON, EMPLOYEES, DAY, CARD, BUILDING, CALL, PLANTS, MEMBERSHIP, TRANSFER, CENTER</td>
</tr>
<tr>
<td>.0314</td>
<td>200</td>
<td>DECEMBER, JANUARY, MARCH, NOVEMBER, FEBRUARY, WEEK, FRIDAY, SEPTEMBER, WEDNESDAY, TUESDAY</td>
</tr>
<tr>
<td>.0028</td>
<td>147</td>
<td>MAIL, CUSTOMER, SERVICE, LIST, SEND, ADDRESS, CONTACT, RECEIVE, BUSINESS, REPLY</td>
</tr>
<tr>
<td>.0026</td>
<td>125</td>
<td>MEETING, CALL, MONDAY, CONFERENCE, FRIDAY, TIME, THURSDAY, OFFICE, MORNING, TUESDAY</td>
</tr>
</tbody>
</table>

**AUTHOR = Outlook Migration Team (132 emails)**

<table>
<thead>
<tr>
<th>PROB.</th>
<th>TOPIC</th>
<th>WORDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>.9910</td>
<td>82</td>
<td>OUTLOOK, MIGRATION, NOTES, OWA, INFORMATION, EMAIL, BUTTON, SEND, MAILBOX, ACCESS</td>
</tr>
<tr>
<td>.0016</td>
<td>91</td>
<td>ENRON, CORP, SERVICES, BROADBAND, EBS, ADDITION, BUILDING, INCLUDES, ATTACHMENT, COMPETITION</td>
</tr>
<tr>
<td>.0005</td>
<td>77</td>
<td>EMAIL, ADDRESS, INTERNET, SEND, ECT, MESSAGING, BUSINESS, ADMINISTRATION, QUESTIONS, SUPPORT</td>
</tr>
<tr>
<td>.0004</td>
<td>83</td>
<td>ISSUE, GENERAL, ISSUES, CASE, DUE, INVOLVED, DISCUSSION, MENTIONED, PLACE, POINT</td>
</tr>
</tbody>
</table>

**AUTHOR = The Motley Fool (145 emails)**

<table>
<thead>
<tr>
<th>PROB.</th>
<th>TOPIC</th>
<th>WORDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>.3593</td>
<td>17</td>
<td>ANALYST, SERVICES, INDUSTRY, TELECOM, ENERGY, MARKETS, FOOL, BANDWIDTH, ESOURCE, TRAINING</td>
</tr>
<tr>
<td>.0773</td>
<td>177</td>
<td>ACCOUNT, ONLINE, OFFER, TRADE, TIME, INVESTMENT, ACCOUNTS, FREE, INFORMATION, ACCESS</td>
</tr>
<tr>
<td>.0713</td>
<td>169</td>
<td>HTTP, WWW, GIF, IMAGES, ASP, SPACER, EMAIL, CGI, HTML, CLICK</td>
</tr>
<tr>
<td>.0660</td>
<td>200</td>
<td>DECEMBER, JANUARY, MARCH, NOVEMBER, FEBRUARY, WEEK, FRIDAY, SEPTEMBER, WEDNESDAY, TUESDAY</td>
</tr>
</tbody>
</table>

**AUTHOR = Individual A (411 emails)**

<table>
<thead>
<tr>
<th>PROB.</th>
<th>TOPIC</th>
<th>WORDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>.1855</td>
<td>105</td>
<td>CUSTOMERS, RATE, PG, CPUC, SCE, UTILITY, ACCESS, CUSTOMER, DECISION, DIRECT</td>
</tr>
<tr>
<td>.1289</td>
<td>54</td>
<td>FERC, MARKET, ISO, COMMISSION, ORDER, FILING, COMMENTS, PRICE, CALIFORNIA, FILED</td>
</tr>
<tr>
<td>.0920</td>
<td>44</td>
<td>MILLION, BILLION, YEAR, NEWS, CORP, CONTRACTS, GAS, COMPANY, COMPANIES, WATER</td>
</tr>
<tr>
<td>.0719</td>
<td>124</td>
<td>STATE, PUBLIC, DAVIS, SAN, GOVERNOR, COMMISSION, GOV, SUMMER, COSTS, HOUR</td>
</tr>
</tbody>
</table>

**AUTHOR = Individual B (193 emails)**

<table>
<thead>
<tr>
<th>PROB.</th>
<th>TOPIC</th>
<th>WORDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>.2590</td>
<td>178</td>
<td>CAPACITY, GAS, EL, PASO, PIPELINE, MMBTU, CALIFORNIA, SHIPPERS, MMCF, RATE</td>
</tr>
<tr>
<td>.0902</td>
<td>74</td>
<td>GAS, CONTRACT, DAY, VOLUMES, CHANGE, DAILY, DAN, MONTH, KIM, CONTRACTS</td>
</tr>
<tr>
<td>.0645</td>
<td>70</td>
<td>GOOD, TIME, WORK, TALK, DON, BACK, WEEK, DIDN, THOUGHT, SEND</td>
</tr>
<tr>
<td>.0599</td>
<td>116</td>
<td>SYSTEM, FACILITIES, TIME, EXISTING, SERVICES, BASED, ADDITIONAL, CURRENT, END, AREA</td>
</tr>
</tbody>
</table>

**AUTHOR = Individual C (159 emails)**

<table>
<thead>
<tr>
<th>PROB.</th>
<th>TOPIC</th>
<th>WORDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>.1268</td>
<td>42</td>
<td>MEXICO, ARGENTINA, ANDREA, BRAZIL, TAX, OFFICE, LOCAL, RICHARD, COPY, STAFF</td>
</tr>
<tr>
<td>.1045</td>
<td>189</td>
<td>AGREEMENT, ENA, LANGUAGE, CONTRACT, TRANSACTION, DEAL, FORWARD, REVIEW, TERMS, QUESTIONS</td>
</tr>
<tr>
<td>.0815</td>
<td>176</td>
<td>MARK, TRADING, LEGAL, LONDON, DERIVATIVES, ENRONONLINE, TRADE, ENTITY, COUNTERPARTY, HOUSTON</td>
</tr>
<tr>
<td>.0784</td>
<td>135</td>
<td>SUBJECT, REQUIRED, INCLUDING, BASIS, POLICY, BASED, APPROVAL, APPROVED, RIGHTS, DAYS</td>
</tr>
</tbody>
</table>

---

**IEEE VisWeek 2010 (Salt Lake City, Utah)**
Example of dynamic topic model (DTM)

Topic model results from analyzing Science papers

[Figure showing dynamic topic model results]

[Text: "Science in the USSR" (1957)]

[Text: "The Atom and Humanity" (1945)]

[Text: "The Costs of the Soviet Empire" (1985)]

[Text: "Alchemy" (1891)]

[Text: "Mass and Energy" (1907)]

[Text: "The Z Boson" (1990)]

[Text: "Speed of Railway Trains in Europe" (1889)]

[Text: "Structure of the Proton" (1974)]

[Text: "Quantum Criticality: Competing Ground States in Low Dimensions" (1980)]

[Text: "Farming and Food Supplies in Time of War" (1915)]

[Text: "Science in the USSR" (1957)]

[Text: "Post-Cold War Nuclear Dangers" (1995)]

[Blei and Lafferty, 2009]
Outline

Themes

- Estimate parameters
- Evaluate and select models

Advanced techniques

- Identify hidden patterns and structures
  - Nonlinear manifold learning
  - Probabilistic topic models

Model sequential/temporal data

- Hidden Markov models and others
Outline

Themes

Estimate parameters
Evaluate and select models

Advanced techniques

Identify hidden patterns and structures

Nonlinear manifold learning
Probabilistic topic models

Model sequential/temporal data

Hidden Markov models and others
Temporal/sequential data

Time series

Perceptual: speech, music, etc
Financial data: stock prices, etc
Sensor data: network traffic, etc
Surveillance data: videos and other imageries

Assumption

Data at different times are correlated.

Treat data at different time slices independently are thus inadequate in capturing all information.
What kind of correlation?

One simplifying yet extremely powerful notion

Markov property

Ex: can you predict the next letter?

I
I a
I am
I am t
...

I am talking about Markov

At any given location in the text stream, the next letter to appear depends on only a few letters back, not the whole history
More formally

State variable

Ex: identity of the letter $s = \{s^0, s^1, s^2, ..., s^{26}\}$

At time $t$, the state variable is denoted by $s_t$

Markov property

$$p(s_t | s_1, s_2, \ldots, s_{t-1}) = p(s_t | s_{t-1})$$

$p(s_t | s_{t-1})$

referred as transition probabilities

there are $27 \times 27$ of them for transitions between letters.
Graphically

Evolving states form a Markov chain

State transition diagram

Ex: for 3 possible states

Transition probability matrix

\[
\begin{bmatrix}
A_{11} & A_{12} & A_{13} \\
A_{21} & A_{22} & A_{23} \\
A_{31} & A_{32} & A_{33}
\end{bmatrix}
\]
Hidden Markov model (HMM)

The need for “hidden”

States cannot be directly observed.

Ex: speech recognition

States are speakers’ utterances

Observed data are listener’s acoustic perception of sound wave.

How to identify the hidden Markov chain underlying the observation?

classical problem in HMM
**Representation of HMM**

**Graphically**

For $T$ time slices Markov chain

![Graphical representation of HMM](image)

**Probabilistically**

represent the joint distribution between observed data and hidden states

$$P(s_1, s_2, \ldots, s_T, x_1, x_2, \ldots, x_T)$$
How Markov is going to help us?

Properties of conditional independence

\[ P(x_t \mid \{s_i\}_{i=1}^T, \{x_i\}_{i=1}^T - x_t) = P(x_t \mid s_t) \]

Factorization of the joint probability

\[ P(s_1, s_2, \ldots, s_T, x_1, x_2, \ldots, x_T) = P(s_1) \prod_{t=2}^T P(s_t \mid s_{t-1}) \prod_{t=1}^T P(x_t \mid s_t) \]
How to infer hidden states from observations?

The most likely sequence of states

Viterbi path

$$P(s_1^*, s_2^*, \ldots, s_T^*) = \arg \max P(s_1, s_2, \ldots, s_T| x_1, x_2, \ldots, x_T)$$

Computable with dynamic programming
How to infer hidden states from observations?

The most likely sequence of states

Viterbi path

$$P(s_1^*, s_2^*, \ldots, s_T^*) = \arg \max P(s_1, s_2, \ldots, s_T | x_1, x_2, \ldots, x_T)$$

Computable with dynamic programming
How to infer hidden states from observations?

The most likely sequence of states

Viterbi path

\[ P(s_1^*, s_2^*, \ldots, s_T^*) = \arg \max P(s_1, s_2, \ldots, s_T | x_1, x_2, \ldots, x_T) \]

Computable with dynamic programming
Parameter estimation of HMMs from data

Maximum likelihood estimation

Commonly used learning algorithm

No closed-form solution, iterative procedure used

Very efficient with dynamic programming.

This procedure is often called Forward/Backward algorithm.

Other techniques

Provide better estimate of parameters

Commonly used in speech recognition
Example: activity recognition

Walk

“left foot” — “right foot”

Jump

“knee bent” — “in the air” — “standup”
A simple example of using HMM for information visualization

Build HMMs for different sequences

Clustering HMMs

Revealing relationships between sequences

400 sequences generated by 4 HMMs [Tino et al, 2005]

Intuitively,
Summary

Themes

Estimate parameters

How to learn models automatically from data?

Evaluate and select models

How to get good models from data?

Advanced techniques

Identify hidden patterns and structures

Nonlinear and probabilistic approaches

Model sequential/temporal data

Hidden Markov models
Resources

Recent textbook on the subjects

C. Bishop. Pattern Recognition and Machine Learning

D. Koller & N. Friedman. Probabilistic Graphical Models

Hastie et al. Elements of Statistical Learning


Software and toolboxes

Manifold Learning Matlab Demo (@UCLA)

Hidden Markov Model Toolbox for Matlab (@UBC)

Weka: data mining in Java

MLOSS: Machine Learning Open Source Software
Acknowledgement

Graphics

Many from Bishop’s textbook
Numerous web resources

Organizer of the tutorial chairs

Niklas Elmqvist
David Gotz
Markus Hadwiger

do-speeker

Guy Lebanon