Today:
• Latent Dirichlet Allocation
  • topic models
• Social network analysis based on latent probabilistic models
• Kernel regression

Readings:
• Kernels: Bishop Ch. 6.1
  optional:
• Bishop Ch 6.2, 6.3
• “Kernel Methods for Pattern Analysis”, Shawe-Taylor & Cristianini, Chapter 2

Supervised Dimensionality Reduction
Supervised Dimensionality Reduction

- Neural nets: learn hidden layer representation, designed to optimize network prediction accuracy
- PCA: unsupervised, minimize reconstruction error
  - but sometimes people use PCA to re-represent original data before classification (to reduce dimension, to reduce overfitting)
- Fisher Linear Discriminant
  - like PCA, learns a linear projection of the data
  - but supervised: it uses labels to choose projection

Fisher Linear Discriminant

- A method for projecting data into lower dimension to hopefully improve classification
- We’ll consider 2-class case

![Diagram of Fisher Linear Discriminant with 2-class case example]
Fisher Linear Discriminant

Project data onto one dimension, to help classification
\[ y = w^T x \]

Define class means: \[ m_i = \frac{1}{N_i} \sum_{n \in C_i} x^n \]

Could choose \( w \) according to: \[ \arg \max_w w^T (m_2 - m_1) \]

Instead, Fisher Linear Discriminant chooses:
\[ \arg \max_w \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} \]

\[ m_i \equiv w^T m_i \quad s_i^2 \equiv \sum_{n \in C_i} (x^n - m_i)^2 \]

Summary: Fisher Linear Discriminant

- Choose n-1 dimension projection for n-class classification problem
- Use within-class covariances to determine the projection
- Minimizes a different error function (the projected within-class variances)
Example topics
induced from a large collection of text

What about Probabilistic Approaches?

Supervised?  Unsupervised?
Example topics
induced from a large collection of text

| Disease | Water | Fish | Sea | Swimming | Pool | Like | Shell | Shark | Tank | Shells | Sharks | Diving | Dolphins | Swam | Long | Seal | Dive | Dolphin | Underwater | Mind | World | Dream | Dreams | Thought | Imagination | Moment | Thoughts | Own | Real | Life | Imagine | Sense | Consciousness | Strange | Feeling | Whole | Being | Might | Hope | Story | Stories | Tell | Character | Characters | Author | Read | Told | Setting | Tales | Plot | Telling | Short | Fiction | True | Events | Tells | Tale | Novel | Field | Magnetic | Magnet | Wire | Needle | Current | Coil | Poles | Iron | Compass | Core | Lines | Core | Appearing | Electric | Direction | Force | Magnets | Be | Magnetism | Pole | Novel | Induced | Science | Study | Scientists | Scientific | Knowledge | Work | Research | Chemistry | Technology | Many | Mathematics | Biology | Field | Physics | Laboratory | Studies | World | Scientist | Studying | Sciences | Ball | Game | Team | Football | Baseball | Players | Play | Field | Player | Basketball | Coach | Played | Playing | Hit | Tennis | Games | Sports | Bat | Terry |

[Tennenbaum et al]

Plate Notation

![Plate Notation Diagram](image-url)
Latent Dirichlet Allocation Model

Figure 1: Graphical model representation of LDA. The boxes are “plates” representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

where \( p(z_n | \theta) \) is simply \( \theta_i \) for the unique \( i \) such that \( z_n = i \). Integrating over \( \theta \) and summing over \( z \), we obtain the marginal distribution of a document:

\[
p(w | \alpha, \beta) = \int p(\theta | \alpha) \left( \prod_{n=1}^{N} \sum_{z} p(z_n | \theta) p(w_n | z_n, \beta) \right) d\theta. \tag{3}
\]

Also extended to case where number of topics is not known in advance (hierarchical Dirichlet processes – [Blei et al, 2004])

Clustering words into topics with Latent Dirichlet Allocation

[Blei, Ng, Jordan 2003]

Probabilistic model for document set:

For each of the \( D \) documents:

1. Pick a \( \theta_d \sim P(\theta | \alpha) \) to define \( P(z | \theta_d) \)
2. For each of the \( N_d \) words \( w \)
   - Pick topic \( z_n \sim P(z | \theta_d) \)
   - Pick word \( w_n \sim P(w | z_n, \phi) \)

Training this model defines topics (i.e., \( \phi \) which defines \( P(W | Z) \))
Example topics
induced from a large collection of text

Significance:
• Learned topics reveal implicit semantic categories of words within the documents
• In many cases, we can represent documents with $10^2$ topics instead of $10^5$ words
• Especially important for short documents (e.g., emails). Topics overlap when words don’t!

[Example topics]

Significance:
• Learned topics reveal implicit semantic categories of words within the documents
• In many cases, we can represent documents with $10^2$ topics instead of $10^5$ words
• Especially important for short documents (e.g., emails). Topics overlap when words don’t!

[Example topics]

[Example topics]

[Example topics]

Significance:
• Learned topics reveal implicit semantic categories of words within the documents
• In many cases, we can represent documents with $10^2$ topics instead of $10^5$ words
• Especially important for short documents (e.g., emails). Topics overlap when words don’t!

[Example topics]

Significance:
• Learned topics reveal implicit semantic categories of words within the documents
• In many cases, we can represent documents with $10^2$ topics instead of $10^5$ words
• Especially important for short documents (e.g., emails). Topics overlap when words don’t!

[Example topics]

Significance:
• Learned topics reveal implicit semantic categories of words within the documents
• In many cases, we can represent documents with $10^2$ topics instead of $10^5$ words
• Especially important for short documents (e.g., emails). Topics overlap when words don’t!

[Example topics]

Significance:
• Learned topics reveal implicit semantic categories of words within the documents
• In many cases, we can represent documents with $10^2$ topics instead of $10^5$ words
• Especially important for short documents (e.g., emails). Topics overlap when words don’t!

[Example topics]

Significance:
• Learned topics reveal implicit semantic categories of words within the documents
• In many cases, we can represent documents with $10^2$ topics instead of $10^5$ words
• Especially important for short documents (e.g., emails). Topics overlap when words don’t!

[Example topics]

Significance:
• Learned topics reveal implicit semantic categories of words within the documents
• In many cases, we can represent documents with $10^2$ topics instead of $10^5$ words
• Especially important for short documents (e.g., emails). Topics overlap when words don’t!

[Example topics]

Significance:
• Learned topics reveal implicit semantic categories of words within the documents
• In many cases, we can represent documents with $10^2$ topics instead of $10^5$ words
• Especially important for short documents (e.g., emails). Topics overlap when words don’t!

[Example topics]

Significance:
• Learned topics reveal implicit semantic categories of words within the documents
• In many cases, we can represent documents with $10^2$ topics instead of $10^5$ words
• Especially important for short documents (e.g., emails). Topics overlap when words don’t!

[Example topics]
Author-Recipient-Topic model for Email

Latent Dirichlet Allocation (LDA)  
[Boe, Ng, Jordan, 2003]

Author-Recipient Topic (ART)  
[McCallum, Corrada, Wang, 2005]

Enron Email Corpus

- 250k email messages
- 23k people

Date: Wed, 11 Apr 2001 06:56:00 -0700 (PDT)
From: debra.perlingiere@enron.com
To: steve.hooser@enron.com
Subject: Enron/TransAltaContract dated Jan 1, 2001

Please see below. Katalin Kiss of TransAlta has requested an electronic copy of our final draft? Are you OK with this? If so, the only version I have is the original draft without revisions.

DP
Debra Perlingiere
Enron North America Corp.
Legal Department
1400 Smith Street, EB 3885
Houston, Texas 77002
dperlin@enron.com
### Topics, and prominent sender/receivers discovered by ART

**[McCallum et al, 2005]**

**Top words within topic:**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>attached 0.0742</td>
<td>day 0.0419</td>
<td>game 0.0170</td>
</tr>
<tr>
<td>agreement 0.0493</td>
<td>friday 0.0418</td>
<td>draft 0.0156</td>
</tr>
<tr>
<td>review 0.0340</td>
<td>morning 0.0369</td>
<td>week 0.0113</td>
</tr>
<tr>
<td>questions 0.0257</td>
<td>monday 0.0282</td>
<td>team 0.0113</td>
</tr>
<tr>
<td>draft 0.0245</td>
<td>office 0.0282</td>
<td>eric 0.0130</td>
</tr>
<tr>
<td>letter 0.0239</td>
<td>wednesday 0.0267</td>
<td>make 0.0125</td>
</tr>
<tr>
<td>comments 0.0207</td>
<td>tuesday 0.0261</td>
<td>free 0.0107</td>
</tr>
<tr>
<td>copy 0.0165</td>
<td>time 0.0218</td>
<td>year 0.0106</td>
</tr>
<tr>
<td>revised 0.0161</td>
<td>good 0.0214</td>
<td>pick 0.0097</td>
</tr>
<tr>
<td>document 0.0156</td>
<td>thursday 0.0191</td>
<td>philip 0.0095</td>
</tr>
</tbody>
</table>

**Top author-recipients exhibiting this topic:**

- G. Nemeck 0.0737
- B. Tyholiz 0.0551
- G. Nemeck 0.0325

### Topics, and prominent sender/receivers discovered by ART

**Beck = “Chief Operations Officer”**

**Dasovich = “Government Relations Executive”**

**Shapiro = “Vice Presidency of Regulatory Affairs”**

**Steffes = “Vice President of Government Affairs”**

**Top words within topic:**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>operations 0.0321</td>
<td>market 0.0567</td>
<td>state 0.0404</td>
<td>blackberry 0.0726</td>
</tr>
<tr>
<td>team 0.0234</td>
<td>power 0.0563</td>
<td>california 0.0367</td>
<td>net 0.0557</td>
</tr>
<tr>
<td>office 0.0173</td>
<td>price 0.0280</td>
<td>power 0.0337</td>
<td>www 0.0409</td>
</tr>
<tr>
<td>list 0.0144</td>
<td>system 0.0206</td>
<td>energy 0.0239</td>
<td>website 0.0375</td>
</tr>
<tr>
<td>bob 0.0129</td>
<td>prices 0.0182</td>
<td>electricity 0.0203</td>
<td>report 0.0373</td>
</tr>
<tr>
<td>open 0.0106</td>
<td>high 0.0124</td>
<td>davis 0.0183</td>
<td>wireless 0.0364</td>
</tr>
<tr>
<td>meeting 0.0107</td>
<td>based 0.0120</td>
<td>utilities 0.0158</td>
<td>handheld 0.0362</td>
</tr>
<tr>
<td>gas 0.0107</td>
<td>buy 0.0117</td>
<td>commission 0.0136</td>
<td>stan 0.0282</td>
</tr>
<tr>
<td>business 0.0106</td>
<td>customers 0.0110</td>
<td>governor 0.0132</td>
<td>fy 0.0271</td>
</tr>
<tr>
<td>houston 0.0099</td>
<td>costs 0.0106</td>
<td>prices 0.0089</td>
<td>numod 0.0260</td>
</tr>
</tbody>
</table>

- S. Beck 0.2156
- L. Kitchen 0.1231
- J. Dasovich 0.3338
- R. Haylett 0.1432
- T. Geaccone 0.0737

- S. Beck 0.0826
- J. Dasovich 0.1133
- R. Shapiro 0.2440
- R. Haylett 0.0420
- T. Geaccone 0.0737

- S. Beck 0.0530
- M. Taylor 0.0218
- J. Dasovich 0.1394
- D. Fossum 0.0420

**Beck = “Chief Operations Officer”**

**Dasovich = “Government Relations Executive”**

**Shapiro = “Vice Presidency of Regulatory Affairs”**

**Steffes = “Vice President of Government Affairs”**
Discovering Role Similarity

Traditional SNA

ART

connection strength \((A, B) = \)

Similarity in recipients they sent email to

Similarity in authored topics, conditioned on recipient

Tracy Geaconne ⇔ Dan McCarty

Geaconne = “Secretary”
McCarty = “Vice President”
Discovering Role Similarity
Lynn Blair ⇔ Kimberly Watson

Traditional SNA
ART

Different
(send to different
individuals)

Similar
(discuss same
topics)

Blair = “Gas pipeline logistics”
Watson = “Pipeline facilities planning”

What you should know

- Unsupervised dimension reduction using all features
  - Principle Components Analysis
    - Minimize reconstruction error
  - Singular Value Decomposition
    - Efficient PCA
  - Independent components analysis
  - Canonical correlation analysis
  - Probabilistic models with latent variables

- Supervised dimension reduction
  - Fisher Linear Discriminant
    - Project to n-1 dimensions to discriminate n classes
  - Hidden layers of Neural Networks
    - Most flexible, local minima issues

- LOTS of ways of combining discovery of latent features with classification tasks
Kernel Functions

- Kernel functions provide a way to manipulate data as though it were projected into a higher dimensional space, by operating on it in its original space.
- This leads to efficient algorithms.
- And is a key component of algorithms such as:
  - Support Vector Machines
  - kernel PCA
  - kernel CCA
  - kernel regression
  - ...

Linear Regression

Wish to learn \( f: X \rightarrow Y \), where \( X=<X_1, \ldots, X_n> \), \( Y \) real-valued

Learn \( \hat{f}(x) = \sum_{i=1}^{N} x_i w_i = \langle x, w \rangle = x^T w \)

where \( w = \arg \min_{w} \sum_{i=1}^{M} (y_i - x_i^T w)^2 + \lambda \sum_{k} w_k^2 \)
Linear Regression

Wish to learn \( f: X \rightarrow Y \), where \( X=\langle X_1, \ldots, X_n \rangle \), \( Y \) real-valued

Learn
\[
\hat{f}(x) = \sum_{i=1}^{N} x_i w_i = \langle x, w \rangle = x^T w
\]

where \( w = \arg \min_{w} \sum_{l=1}^{N} (y^l - x^T l w)^2 + \lambda \sum_{k} w_k^2 \)

\[
\hat{w} = \arg \min_{w} \|y - Xw\|^2 + \lambda \|w\|^2
\]

note \( l^{th} \) row of \( X \) is \( l^{th} \) training example \( x^T l \)

\[
\|w\|^2 = \sum_{k} w_k^2 = \|w\|_2^2
\]

---

Linear Regression

Wish to learn \( f: X \rightarrow Y \), where \( X=\langle X_1, \ldots, X_n \rangle \), \( Y \) real-valued

Learn
\[
\hat{f}(x) = \sum_{i=1}^{N} x_i w_i = \langle x, w \rangle = x^T w
\]

where \( w = \arg \min_{w} \|y - Xw\|^2 + \lambda \|w\|^2 \)

solve by taking derivative wrt \( w \), setting to zero...

\[
w = (X^T X + \lambda I)^{-1} X^T y
\]

so: \( \hat{f}(x_{\text{new}}) = x_{\text{new}}^T w = x_{\text{new}}^T (X^T X + \lambda I)^{-1} X^T y \)