CSCI 5622 Machine Learning

ML Topic Modeling

DATE	TOPIC	DUE
Wed, Dec 2	Topic Modeling	Draft Paper to Peers
Mon, Dec 7	Sequence Learning & Planning	Peer Feedback Due
Wed, Dec 9	Planning & Reinforcement Learning	Peer Feedback Grades
Fri, Dec 11	----------	***Final Paper Due***
Tue, Dec 15	Peer Presentations	Peer project grades
W 16th 7:30PM	Final Presentations, ***Rm ECCR133***	Presentations
.. (Dec 16)	***Mandatory attendance & participation***	***Peer Project Questions***

www.RodneyNielsen.com/teaching/CSCI5622-F09/

Assistant Professor Adjunct, CU Dept. of Computer Science

ML Vector Space Representation

- LSA
  - Semantic info derived from word-document co-occurrence matrix
  - Dimensionality reduction essential part of semantics derivation
  - Words & Docs represented as points in Euclidean space

Topic Model Representation

ML

- Topic Modeling
  - Semantic info derived from word-document co-occurrence matrix
  - Dimensionality reduction essential part of semantics derivation
  - Words & Docs represented as probabilistic topics

Topic Modeling

- Topic Modeling
  - Generative model of documents
- Documents
  - Mixture of topics
- Topic
  - Distribution across words

Topic Modeling

- Topic Modeling
  - Generative model of documents
- Generating a document
  - Select a distribution over topics: what topics will the document be about
  - Generating the words
    - Select a word at random according to the distribution of words over the topic

ML

Topic Modeling

- Use standard statistical techniques to infer the set of topics used to generate a model

ML Ex: Topics from an Educational Corpus

Sixteen highest probability words for four of 300 topics in TASA

Topic 247

Drug Use

word	prob.
DRUGS	.069
DRUG	.060
MEDICINE	.027
EFFECTS	.026
BODY	.023
MEDICINES	.019
PAIN	.016
PERSON	.016
MARIJUANA	.014
LABEL	.012
ALCOHOL	.012
DANGEROUS	.011
ABUSE	.009
EFFECT	.009
KNOWN	.008
PILLS	.008

Topic 5

Colors

word	prob.
RED	.202
BLUE	.099
GREEN	.096
YELLOW	.073
WHITE	.048
COLOR	.048
BRIGHT	.030
COLORS	.029
ORANGE	.027
BROWN	.027
PINK	.017
LOOK	.017
BLACK	.016
PURPLE	.015
CROSS	.011
COLORED	.009

Topic 43

Mind & Memory

word	prob.
MIND	.081
THOUGHT	.066
REMEMBER	.064
MEMORY	.037
THINKING	.030
PROFESSOR	.028
FELT	.025
REMEMBERED	.022
THOUGHTS	.020
FORGOTTEN	.020
MOMENT	.020
THINK	.019
THING	.016
WONDER	.014
FORGET	.012
RECALL	.012

Topic 56

Doctor Visits

word	prob.
DOCTOR	.074
DR.	.063
PATIENT	.061
HOSPITAL	.049
CARE	.046
MEDICAL	.042
NURSE	.031
PATIENTS	.029
DOCTORS	.028
HEALTH	.025
MEDICINE	.017
NURSING	.017
DENTAL	.015
NURSES	.013
PHYSICIAN	.012
HOSPITALS	.011

Steyvers & Griffiths (2007)

Topic Models

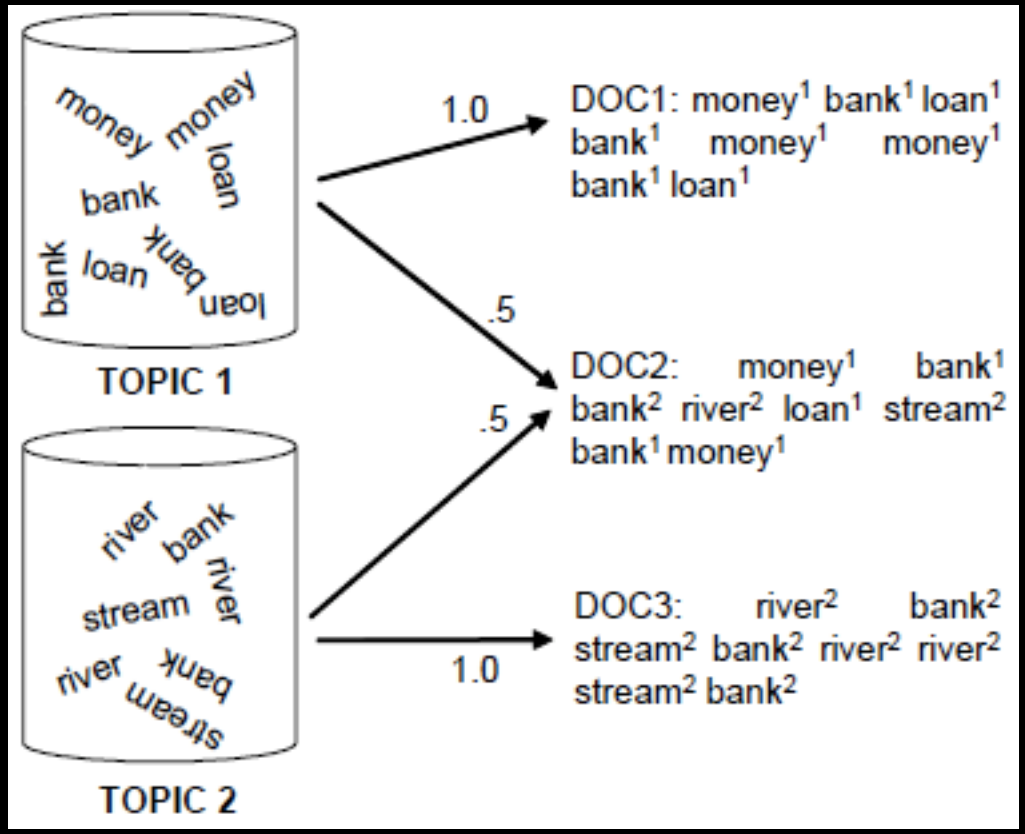
- Big advantage of Topic Modeling over Euclidean space representation
  - Topics are clearly interpretable
  - Axes of vector representation (e.g., 300 dimensions of LSA) have no interpretation

Generative Models

- Generative models for documents
  - Generate a document by probabilistic selection of words based on some underlying distribution theory
- Learning the model
  - Search for maximum likelihood hypothesis given the data
    (Find values for latent variables that maximize the probability of generating the observed data, e.g., the observed words in the documents)

Generative Models

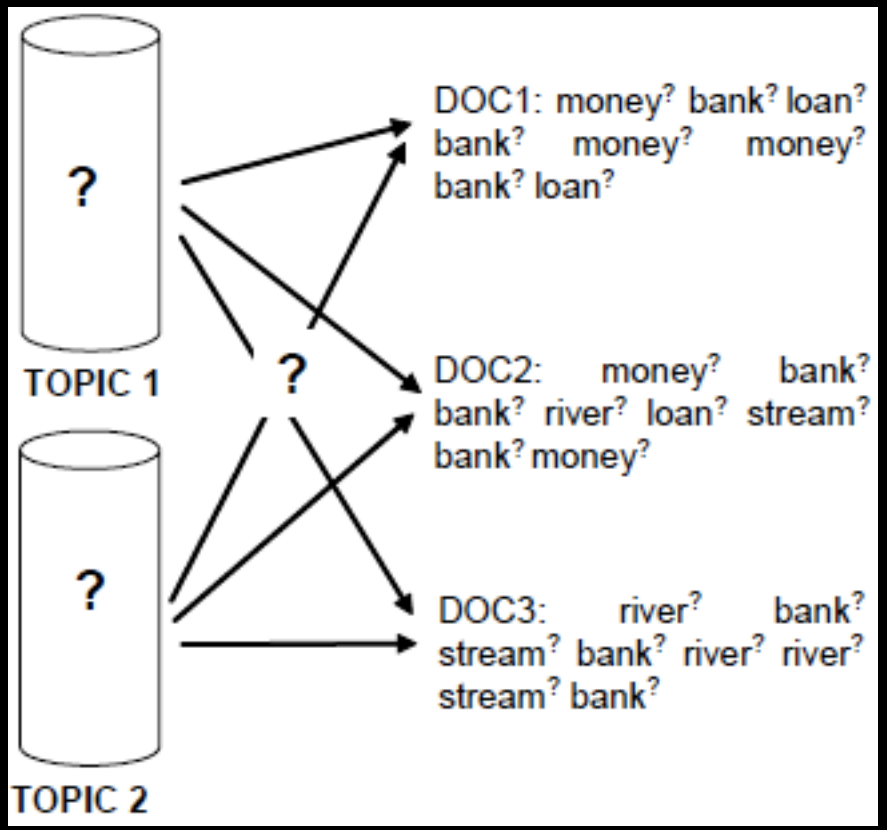
- Probabilistic generative process



Steyvers & Griffiths (2007)

Statistical Inference

- Learning the values of model's latent variables



Steyvers & Griffiths (2007)

Bag of Words Assumption

ML

- The order of the words is irrelevant
- Only the frequency or probability of a word is important

ML Probabilistic Topic Models

- Wide variety of probabilistic topic models
  - All assume documents are mixtures of topics
  - Vary in distributional assumptions

Notation

ML

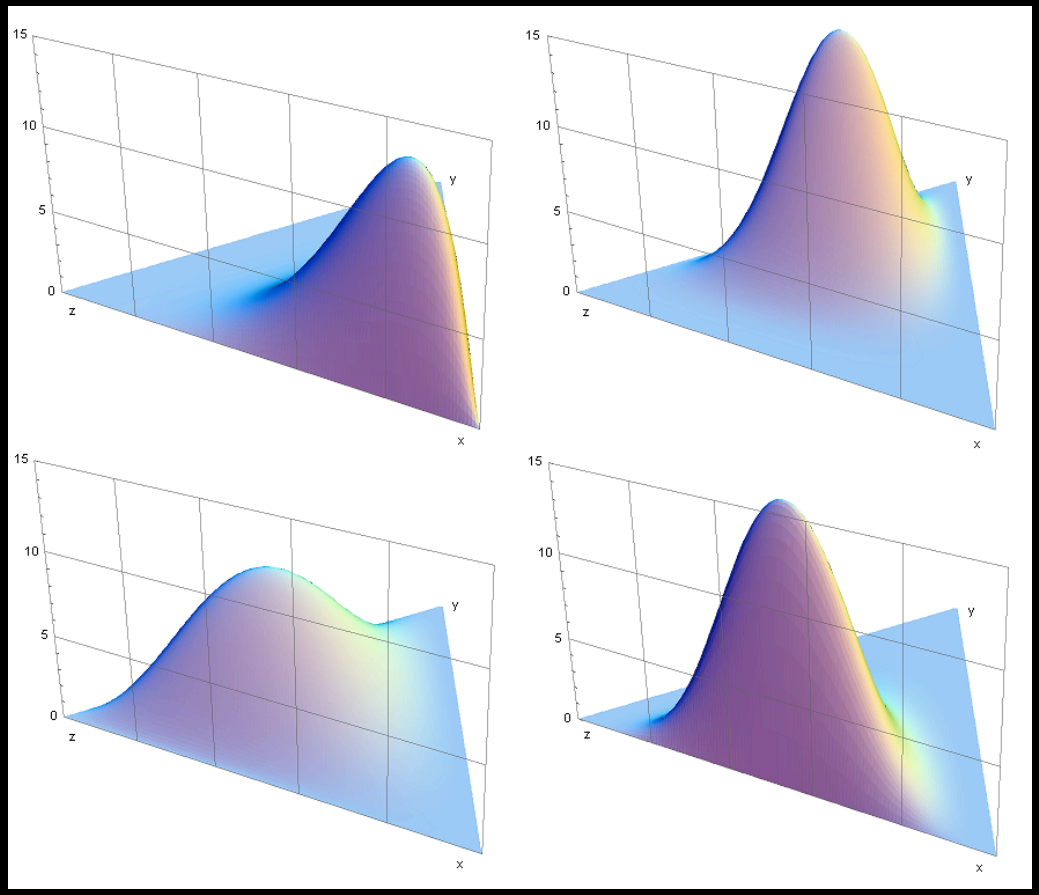
- $P(z_i = j)$: the probability that the i^{th} word was generated by the j^{th} topic
- $P(w_i | z_i = j)$: the probability of generating word w_i given the j^{th} topic
- $P(w_i) = \Sigma_j \; P(z_i = j)P(w_i | z_i = j)$

Dirichlet Distribution

$$Dir(\alpha_1,\ldots,\alpha_T) = \frac{\Gamma\left(\sum_j \alpha_j\right)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^{T} p_j^{\alpha_j - 1}$$

Dirichlet Distribution

$\alpha = (6, 2, 2)$



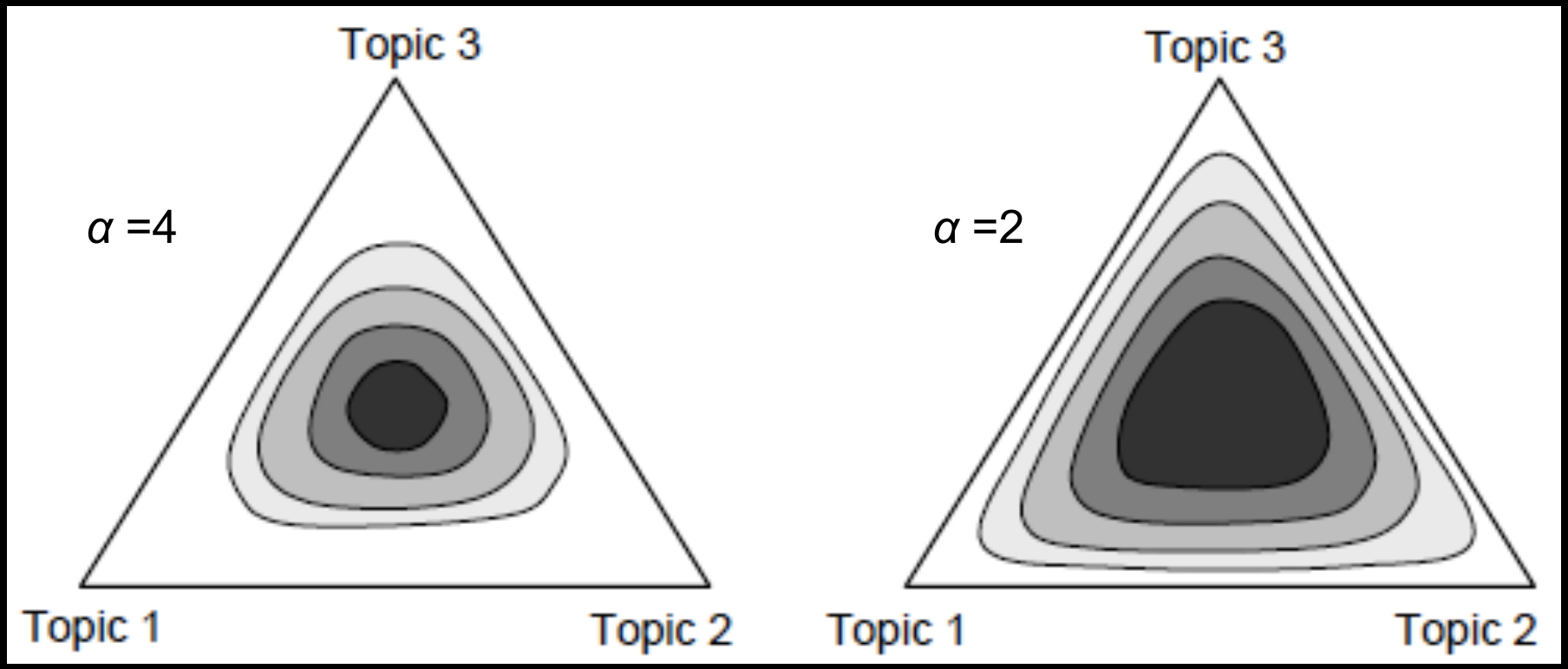
$\alpha = (3, 7, 5)$

$\alpha = (2, 3, 4)$

$\alpha = (6, 2, 6)$

ML Symmetric Dirichlet Distribution

- Symmetric Dirichlet distribution over 3 topics

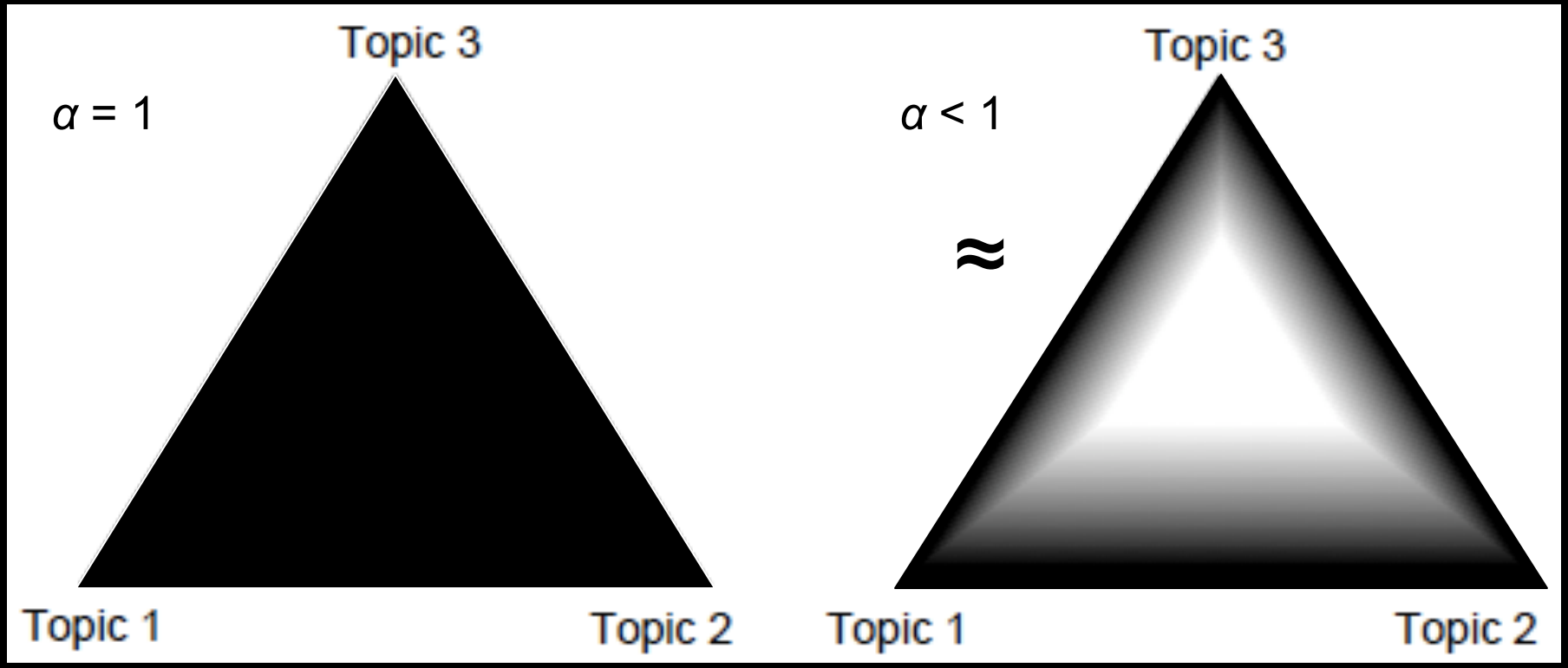


- Probability mass greater over black areas

Steyvers & Griffiths (2007)

ML Symmetric Dirichlet Distribution

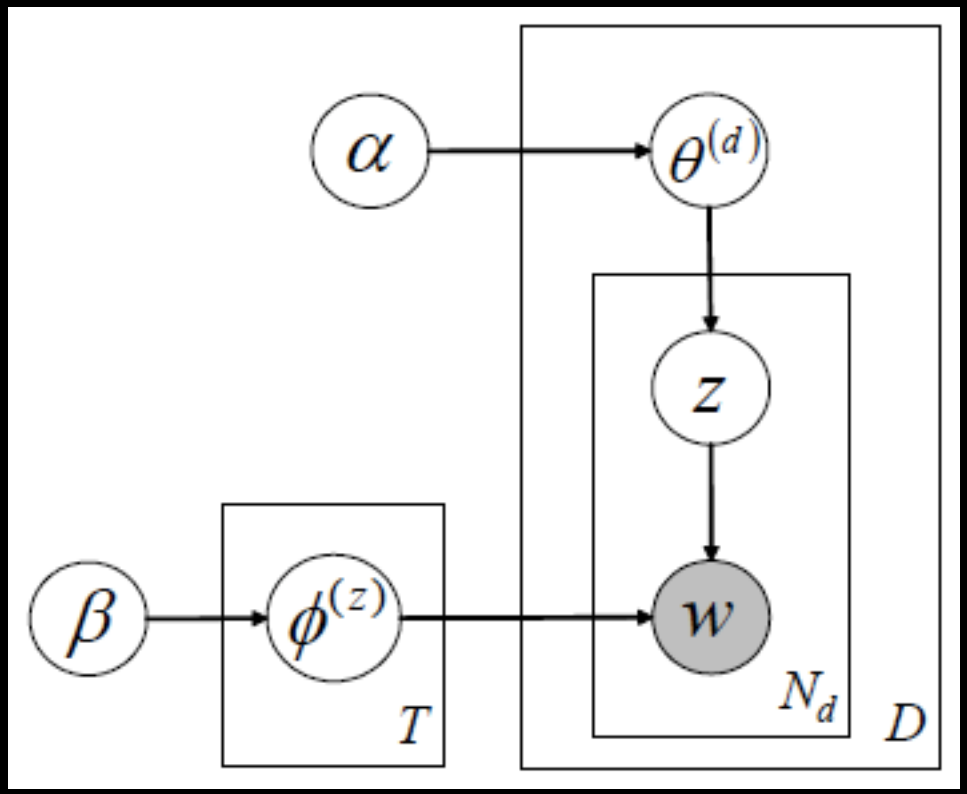
- Symmetric Dirichlet distribution over 3 topics



- Probability mass greater over black areas

ML Topic Model Plate Notation

- Graphic model of topic models



$\Theta^{(d)}$ = Topic distribution for document *d*

$\phi^{(z)}$ = Word distribution for topic *z*

Extracting Topics

ML

- EM algorithm
- Markov Chain Monte Carlo
- Gibbs Sampling

Calculating Similarity

ML

- Words are similar if they occur in the same topics
- Documents are similar if they are comprised on similar topics

ML Calculating Document Similarity

- Kullback-Leibler divergence between two distributions $p = \theta_1$ & $q = \theta_2$, (over topics)

$$D(p,q) = \sum_{j=1}^{T} p_j \log_2 \frac{p_j}{q_j}$$

- Symmetric Kullback-Leibler divergence

$$KL(p,q) = \frac{1}{2}\left[D(p,q) + D(p,q)\right]$$

ML Calculating IR Doc-Query Similarity

- Symmetric Kullback-Leibler divergence

$$KL(p,q) = \frac{1}{2}\left[D(p,q) + D(p,q)\right]$$

- Probability of the query given the document

$$P(q,d_i) = \prod_{w_k \in q} P(w_k|d_i)$$

$$= \prod_{w_k \in q} \sum_{j=1}^{T} P(w_k|z=j)P(z=j|d_i)$$

Calculating Word Similarity

- Symmetric Kullback-Leibler divergence between conditional topic distributions for the words, w_1 & w_2; $p = \theta_1 = P(z \mid w = w_1)$

$$KL(p,q) = \frac{1}{2}\left[D(p,q) + D(p,q)\right]$$

$$D(p,q) = \sum_{j=1}^{T} p_j \log_2 \frac{p_j}{q_j}$$

Calculating Word Similarity

ML

- Or …

$$P(w_2|w_1) = \sum_{j=1}^{T} P(w_2|z=j)P(z=j|w_1)$$

ML Questions

- Questions???

Dimensionality Reduction

- Why reduce the number of features?
  - Many applications involve very high dimensional data resulting in problematic processing times
    - Ex: Principal component analysis
  - Eliminating redundant features and noise usually improves the accuracy of the final classifier
    - Ex: Feature selection

Feature Selection

- Exponential complexity to find optimal features
- Forward selection
- Backward elimination
- Combination
- Decision tree-based selection
- Genetic algorithm selection

- Wrapper
- Filter / Relevancy
- Hybrid

Dimensionality Reduction

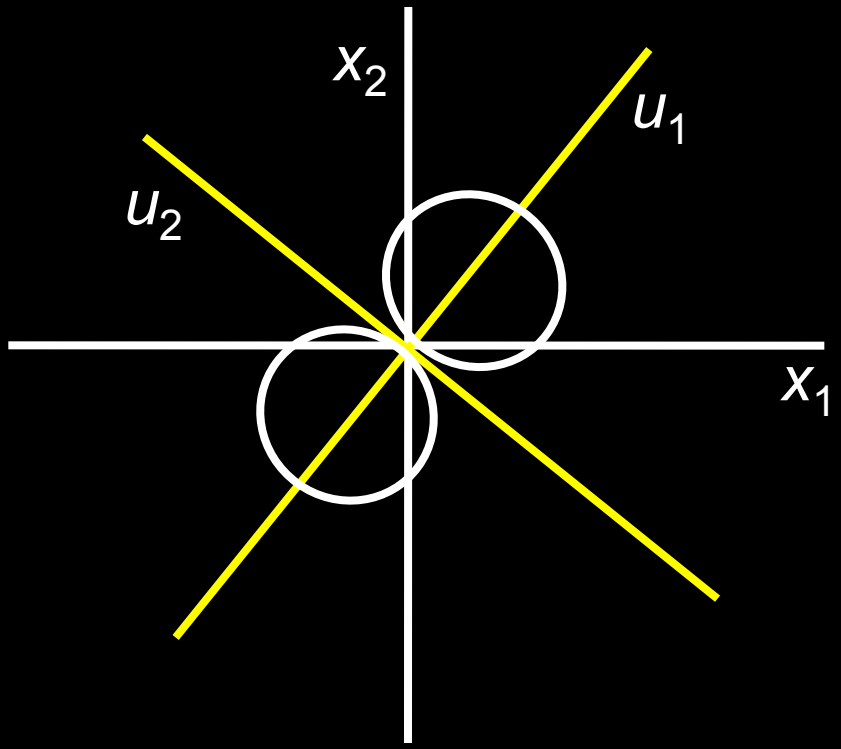
- Goal:
  - Reduce the volume of data used in the learning process
  - With minimal or no impact on your evaluation metric (e.g., accuracy)
    - Or with an improvement in the analytical performance?

ML Principal Component Analysis

- Principle Components definition: for a given set of data vectors, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, the *d*' principal axes [components] are those orthonormal axes onto which the variance retained under projection is maximal (Hotelling, 1933)

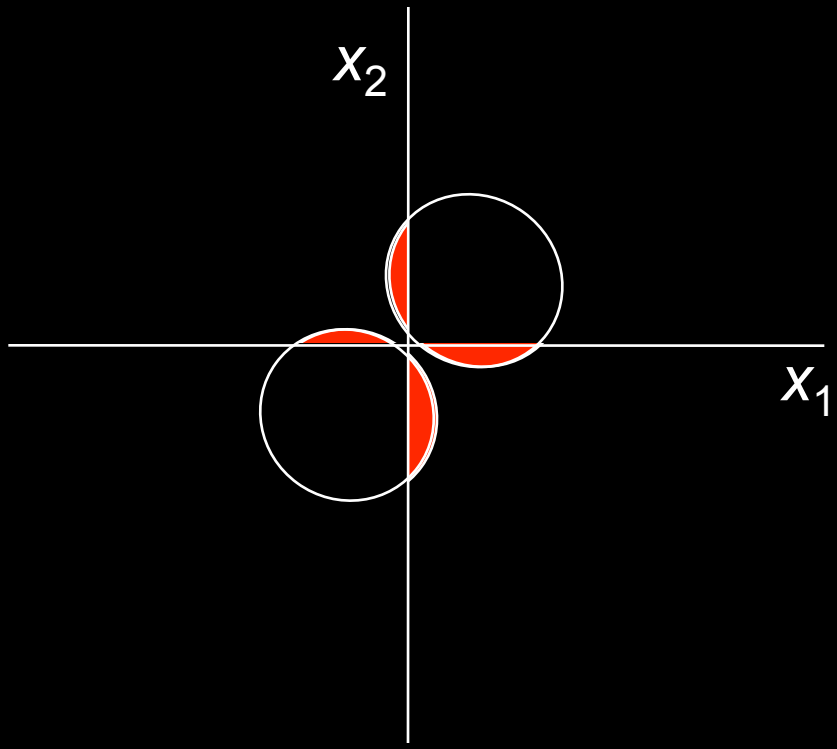
Principal Component Analysis

- Compute principal components that maximize the variance retained



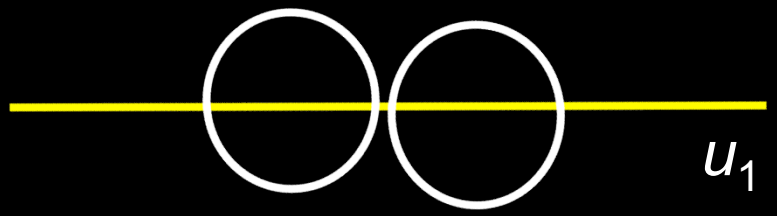
Principal Component Analysis

- In this example, retaining only x_1 or only x_2 results in the inability to learn a perfect classifier



ML Principal Component Analysis

- Projecting onto the first principal component in this example *does* allow perfect classification



ML Principal Component Analysis

- Another nice property of PCA is that it results in the closest approximation to the original matrix **X** – it is the least squares error fit
- I.e., it minimizes the error $\Sigma_{i=1..n} \| \mathbf{x}^{(i)} - \mathbf{z}^{(i)} \|$, where **z** is an approximation of **x** after the projection onto **u**

ML Questions

- Questions???

ML Conference Deadlines

- AAAI
  - http://www.aaai.org/Conferences/AAAI/aaai10.php
  - January 18, 2010: Electronic abstracts due
  - January 21, 2010: Electronic papers due

ML Presentations

- Volunteers
  - Wed, Dec 2: Davide
  - Mon, Dec 7: Nirav & Kun
  - Wed, Dec 9: Tony
  - Xinyu

Projects

ML

- Will stay until all questions are addressed, or
- Please make an appointment if you have any questions at all regarding your project