Clustering text

Nisheeth

Overview

- Quickly review clustering
  - Emphasizing cluster quality assessment
- Introduce plate notation
- Introduce text clustering algorithms
- Focus on LDA
- Useful reading: MC Burton's intro to topic modeling
  - http://mcburton.net/blog/joy-of-tm/

K means clustering

- Exclusive clustering approach
- Each cluster is associated with a centroid (center point)
- Each point is assigned to the cluster with the closest centroid
- Number of clusters, K, must be specified
- The basic algorithm is very simple

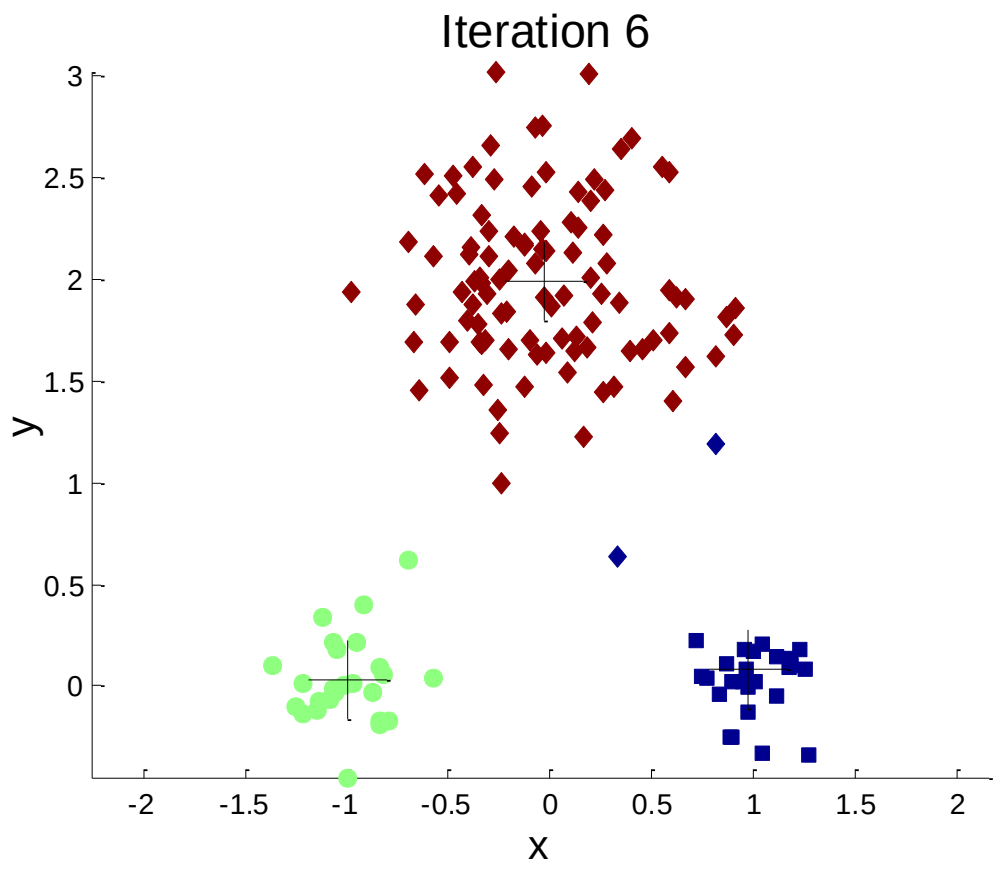
1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

DBSCAN algorithm

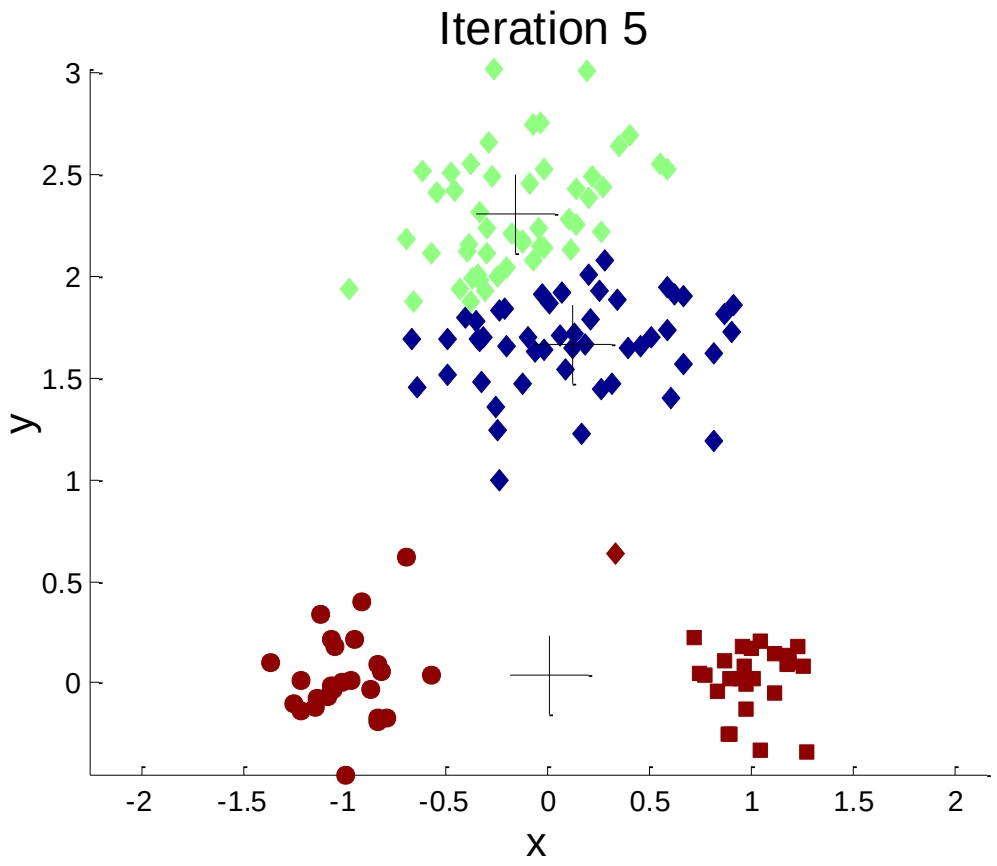
- Eliminate noise points
- Perform clustering on the remaining points

current_cluster_label $\leftarrow 1$
for all core points **do**
  if the core point has no cluster label **then**
    current_cluster_label $\leftarrow$ *current_cluster_label* $+ 1$
    Label the current core point with cluster label *current_cluster_label*
  end if
  for all points in the *Eps*-neighborhood, except i^{th} the point itself **do**
    if the point does not have a cluster label **then**
      Label the point with cluster label *current_cluster_label*
    end if
  end for
end for

Good result



Bad result



Quantifying clustering quality

- Cluster Cohesion: Measures how closely related are objects in a cluster
  - Example: SSE
- Cluster Separation: Measures how distinct or well-separated a cluster is from other clusters
- Example: Squared Error
  - Cohesion is measured by the within cluster sum of squares (SSE)

$$WSS = \sum_i \sum_{x \in C_i} (x - m_i)^2$$

  - Separation is measured by the between cluster sum of squares

$$BSS = \sum_i |C_i| (m - m_i)^2$$

    - Where $|C_i|$ is the size of cluster i

Quantifying clustering quality

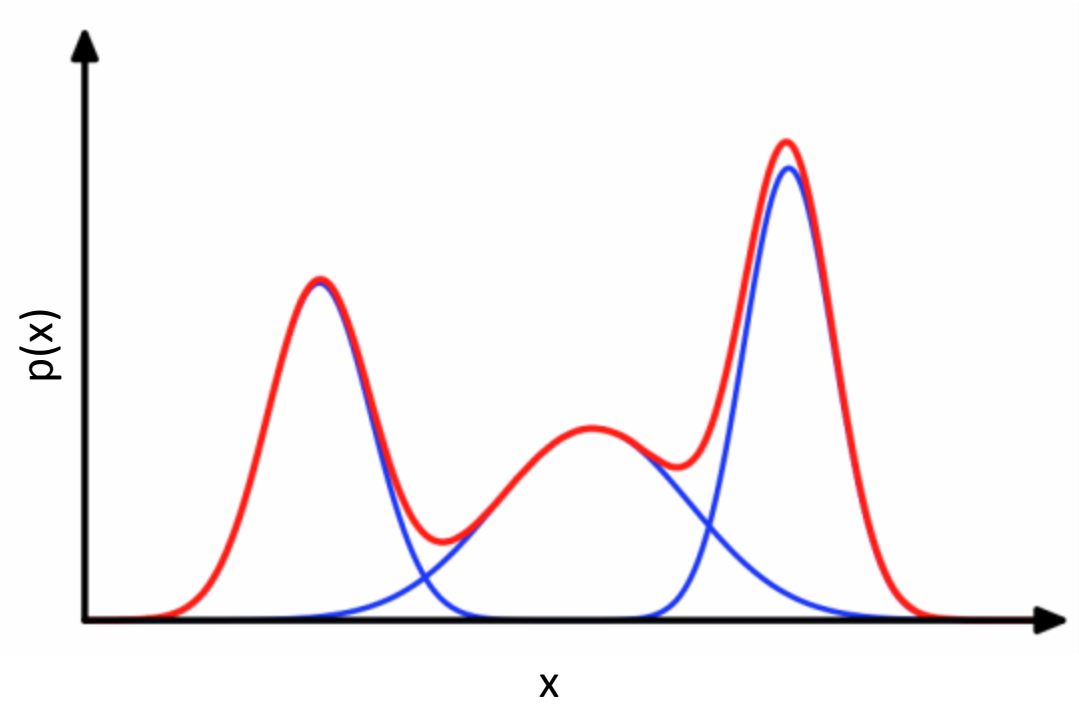
Table 5.9. K-means Clustering Results for LA Document Data Set

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

entropy For each cluster, the class distribution of the data is calculated first, i.e., for cluster j we compute p_{ij}, the 'probability' that a member of cluster j belongs to class i as follows: $p_{ij} = m_{ij}/m_j$, where m_j is the number of values in cluster j and m_{ij} is the number of values of class i in cluster j. Then using this class distribution, the entropy of each cluster j is calculated using the standard formula $e_j = \sum_{i=1}^{L} p_{ij} \log_2 p_{ij}$, where the L is the number of classes. The total entropy for a set of clusters is calculated as the sum of the entropies of each cluster weighted by the size of each cluster, i.e., $e = \sum_{i=1}^{K} \frac{m_i}{m} e_j$, where m_j is the size of cluster j, K is the number of clusters, and m is the total number of data points.

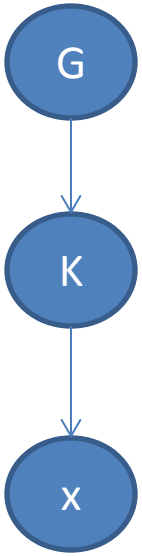
purity Using the terminology derived for entropy, the purity of cluster j, is given by $purity_j = \max p_{ij}$ and the overall purity of a clustering by $purity = \sum_{i=1}^{K} \frac{m_i}{m} purity_j$.

What does the model tell you?

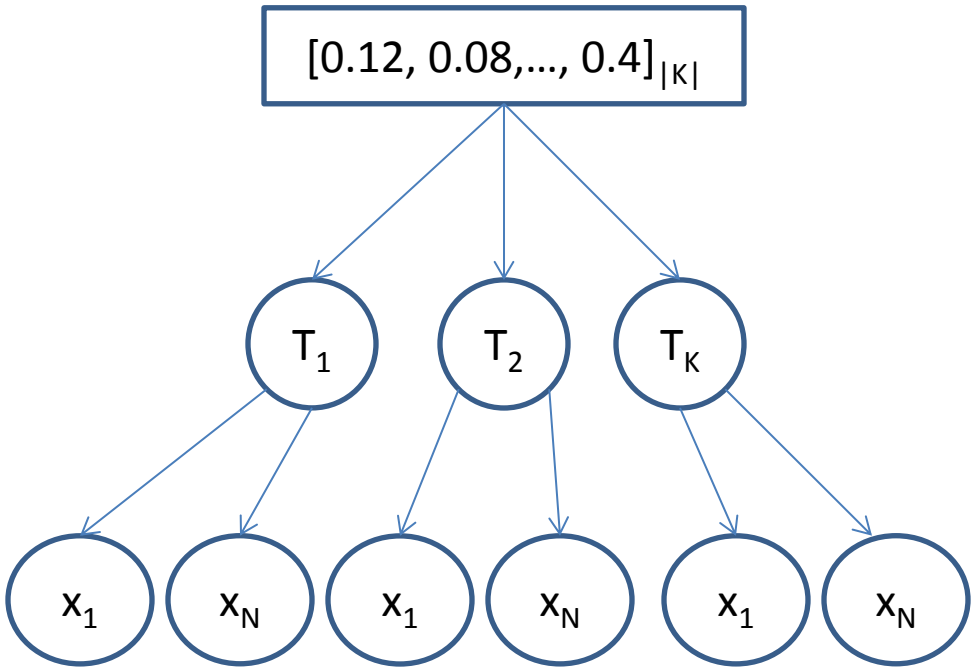


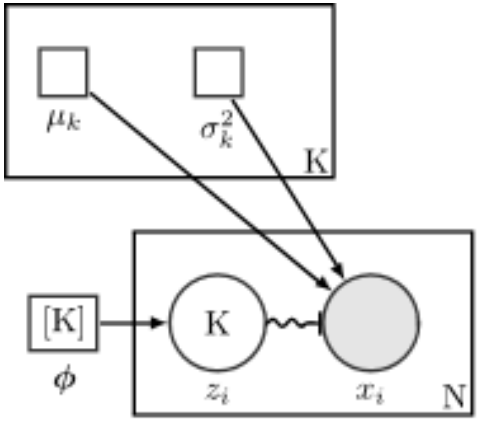
With some probability, pick a Gaussian

With some probability, pick a point from the Gaussian



Shift to plate notation





Coin toss example

- Say you toss a coin N times
- You want to figure out its bias
- Bayesian approach
  - Find the generative model
  - Each toss ~ Bern(θ)
  - θ ~ Beta(α,β)
- Draw the generative model in plate notation

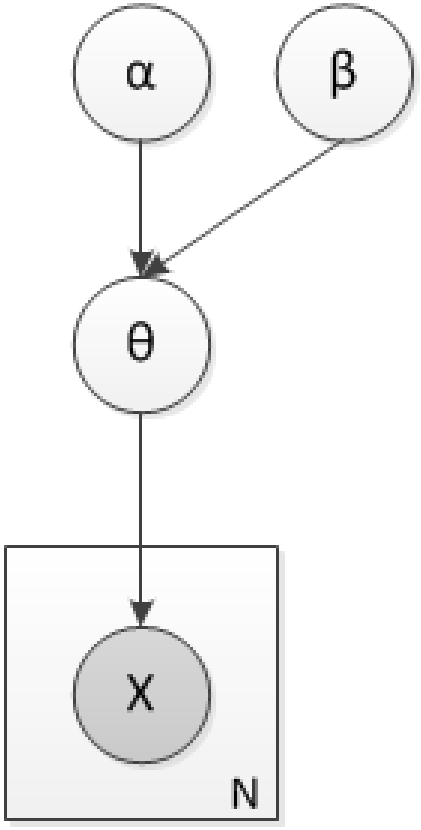


Plate notation

- Random variables as circles
- Parameters, fixed values as squares
- Repetitions of conditional probability structures as rectangular ‘plates’
- *Switch* conditioning as squiggles
- Random variables observed in practice are shaded

Conjugacy

- Algebraic convenience in Bayesian updating
- Posterior ← Prior x Likelihood
- We want the distributions to be parametric, the parameter is what is learned
  - we want the posterior to have the same parametric form as the prior
  - Conjugate prior = f(.) such that $f(\theta)g(x|\theta) \sim f(\theta^{new})$

Useful conjugate priors

likelihood	conjugate prior	posterior
$p(x \mid \theta)$	$p_0(\theta)$	$p(\theta \mid x)$
Normal (θ, σ)	Normal (μ_0, σ_0)	Normal (μ_1, σ_1)
Binomial (N, θ)	Beta (r, s)	Beta $(r + n, s + N - n)$
Poisson (θ)	Gamma (r, s)	Gamma $(r + n, s + 1)$
Multinomial $(\theta_1, \ldots, \theta_k)$	Dirichlet $(\alpha_1, \ldots, \alpha_k)$	Dirichlet $(\alpha_1 + n_1, \ldots, \alpha_k + n_k)$

This one is important for us

Remember the query-Likelihood model?

- Rank documents by the probability that the query could be generated by the document model (i.e. same topic)
- Given query, start with P(D|Q)
- Using Bayes' Rule

$$p(D|Q) \overset{rank}{=} P(Q|D)P(D)$$

- Assuming prior is uniform, unigram model

$$P(Q|D) = \prod_{i=1}^{n} P(q_i|D)$$

- Alternative formulation: **multinomial** unigram model

$$P(Q|D) = \prod_{i=1}^{n} P(q_i|D)^{tf(q_i,q)}$$

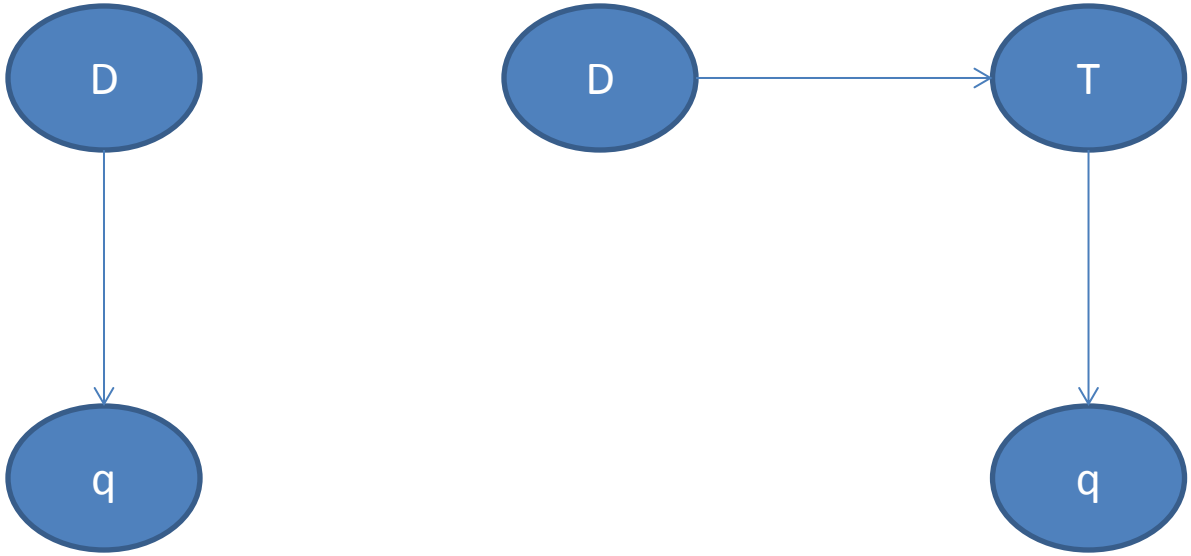
Multinomial unigram model

- Each word assumed generated from a single multinomial distribution
- In plate notation

$$p(\boldsymbol{w}) = \prod_{i=1}^{n} p(w_i)$$

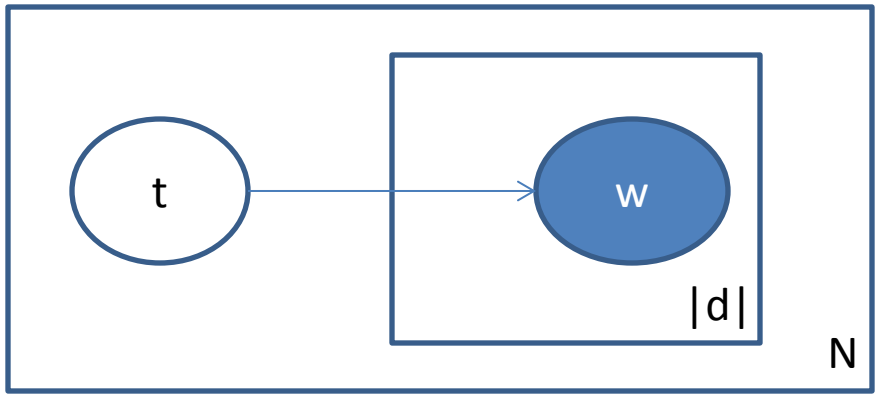
- Probabilistic alternative to tf.idf

Going beyond tf.idf in text processing



Mixture of unigrams

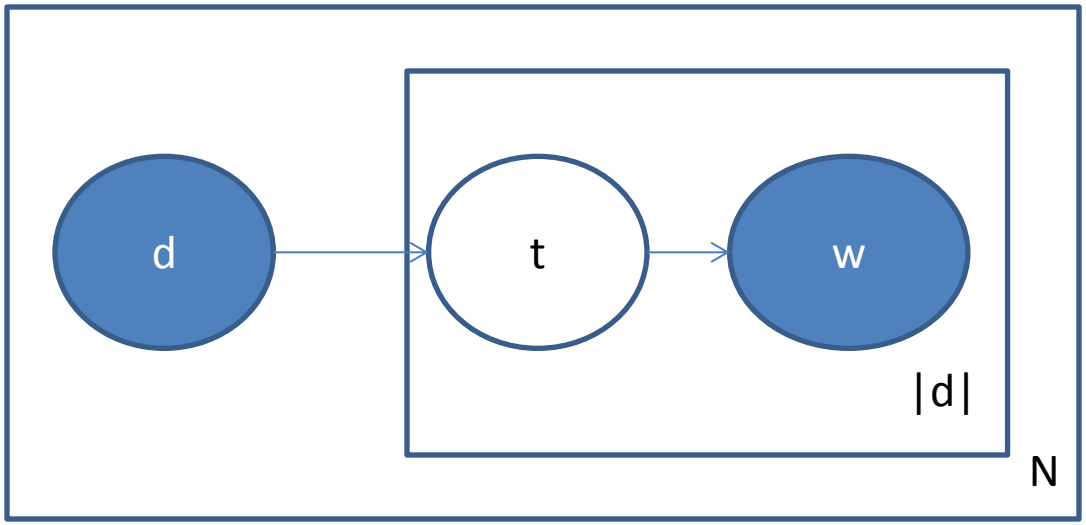
- Document label generated from a topic
- Words generated from topic-specific word distributions
- Strong assumption: one document generated from one topic only



$$p(\boldsymbol{w}) = \sum_{t} p(t) \prod_{i=1}^{|d|} p(w_i|t)$$

Probabilistic latent semantic analysis

- Assume topics are drawn from documents
- Assume words are drawn from topics



$$p(d, \boldsymbol{w}) = p(d) \sum_{t} \prod_{i=1}^{|d|} p(w_i|t) p(t|d)$$

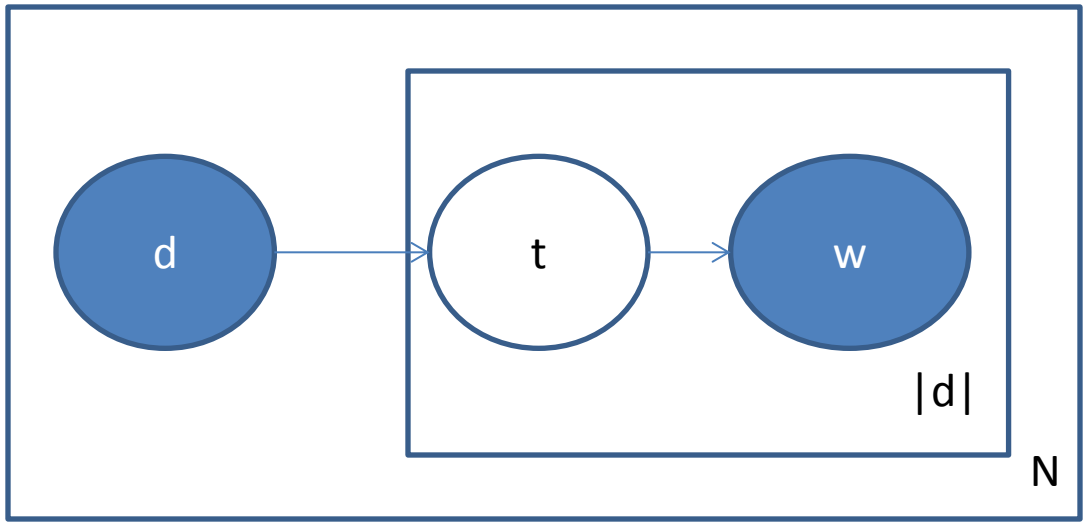
Problem of PLSI

- Mixture weights are considered as document specific, thus no natural way to assign probability to a previously unseen document.
- Number of parameters to be estimated grows linearly with size of training set
  - overfits data
  - multiple local maxima.
- Not a fully generative model of documents.

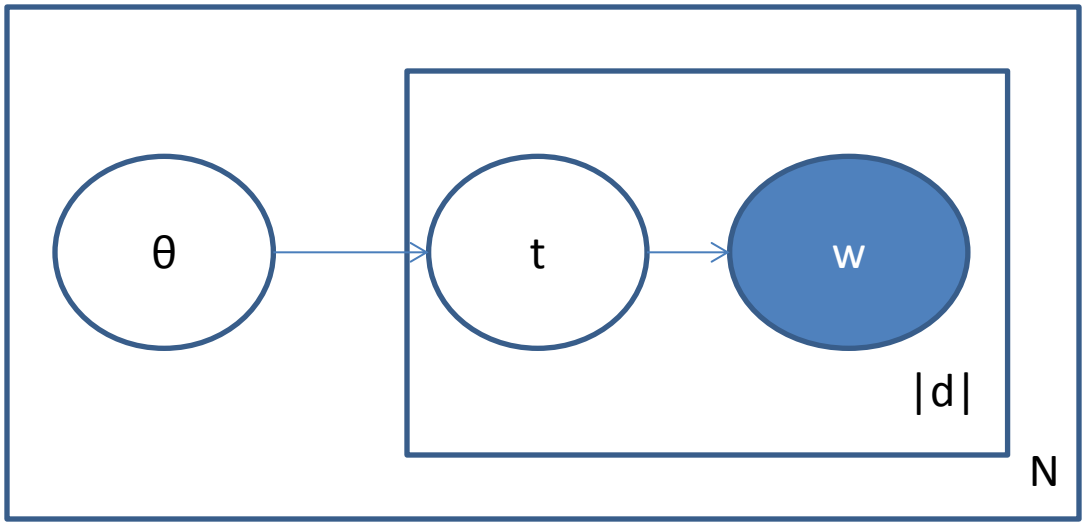
Latent Dirichlet allocation

- LDA is a generative probabilistic model of a corpus.
  - Documents are considered random mixtures over latent topics
  - Topic are characterized by a distribution over words.

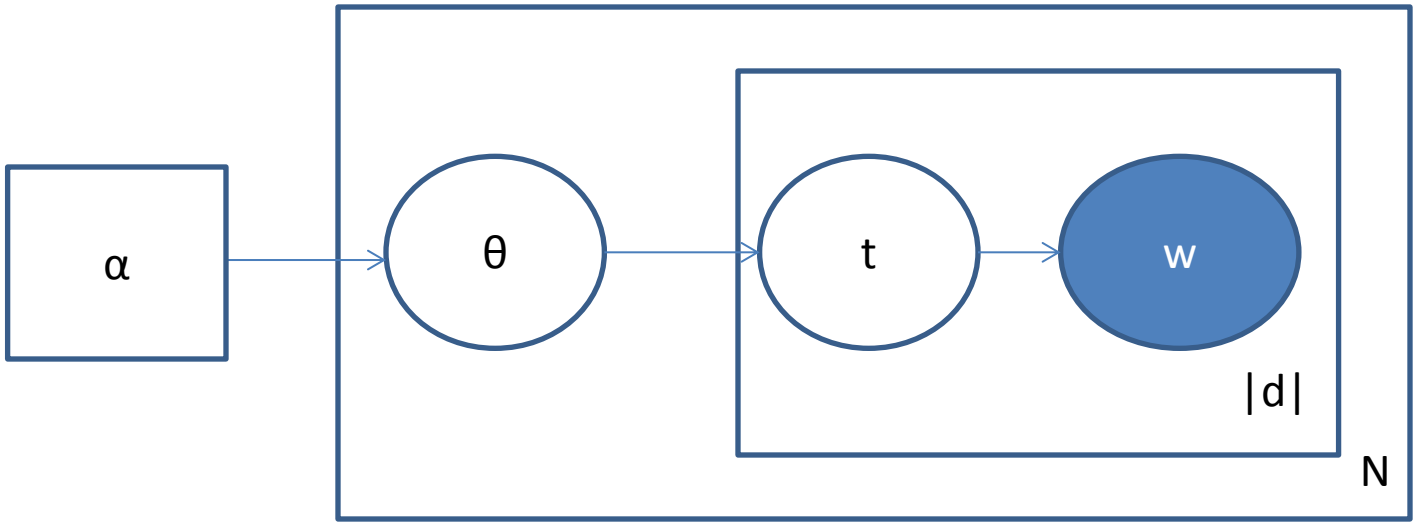
LDA in plate notation



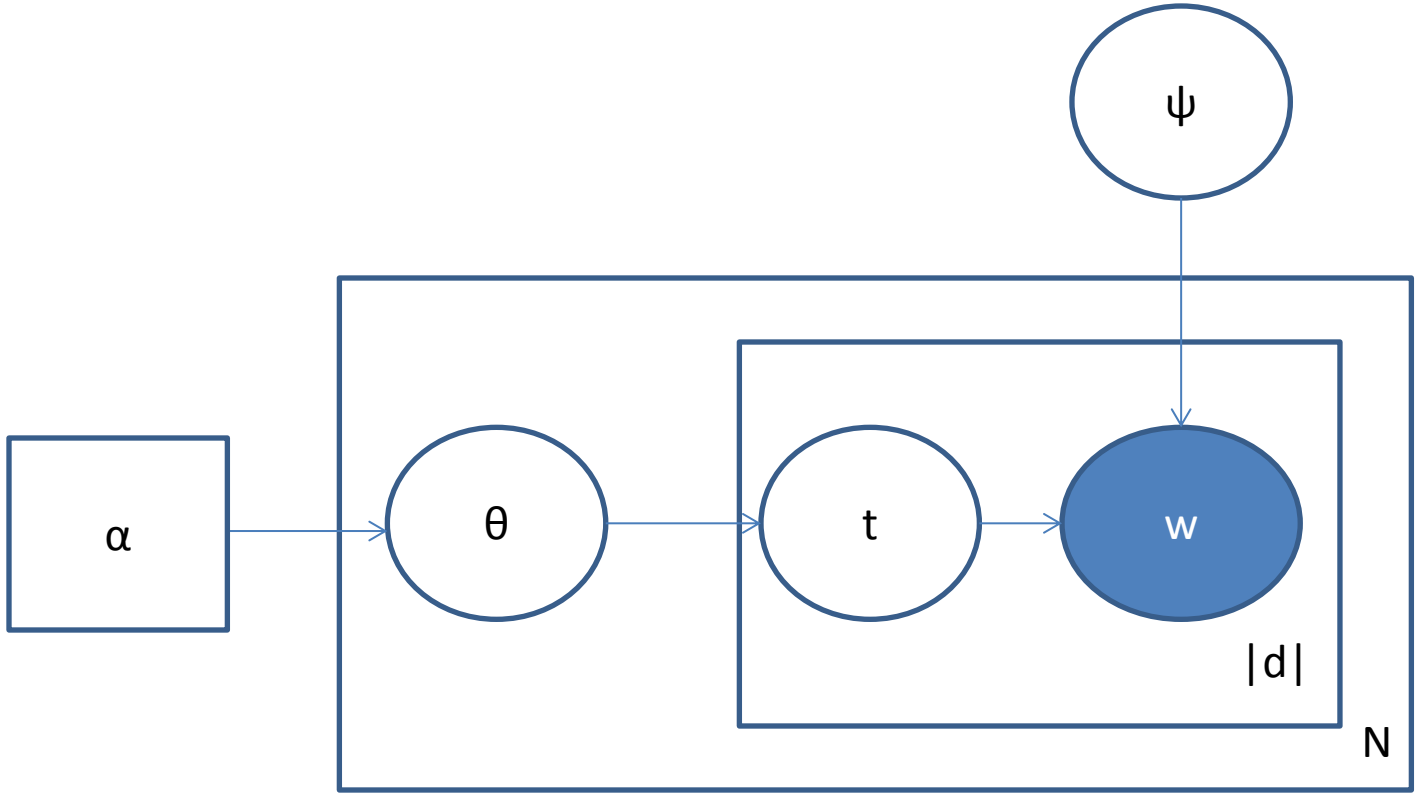
LDA in plate notation



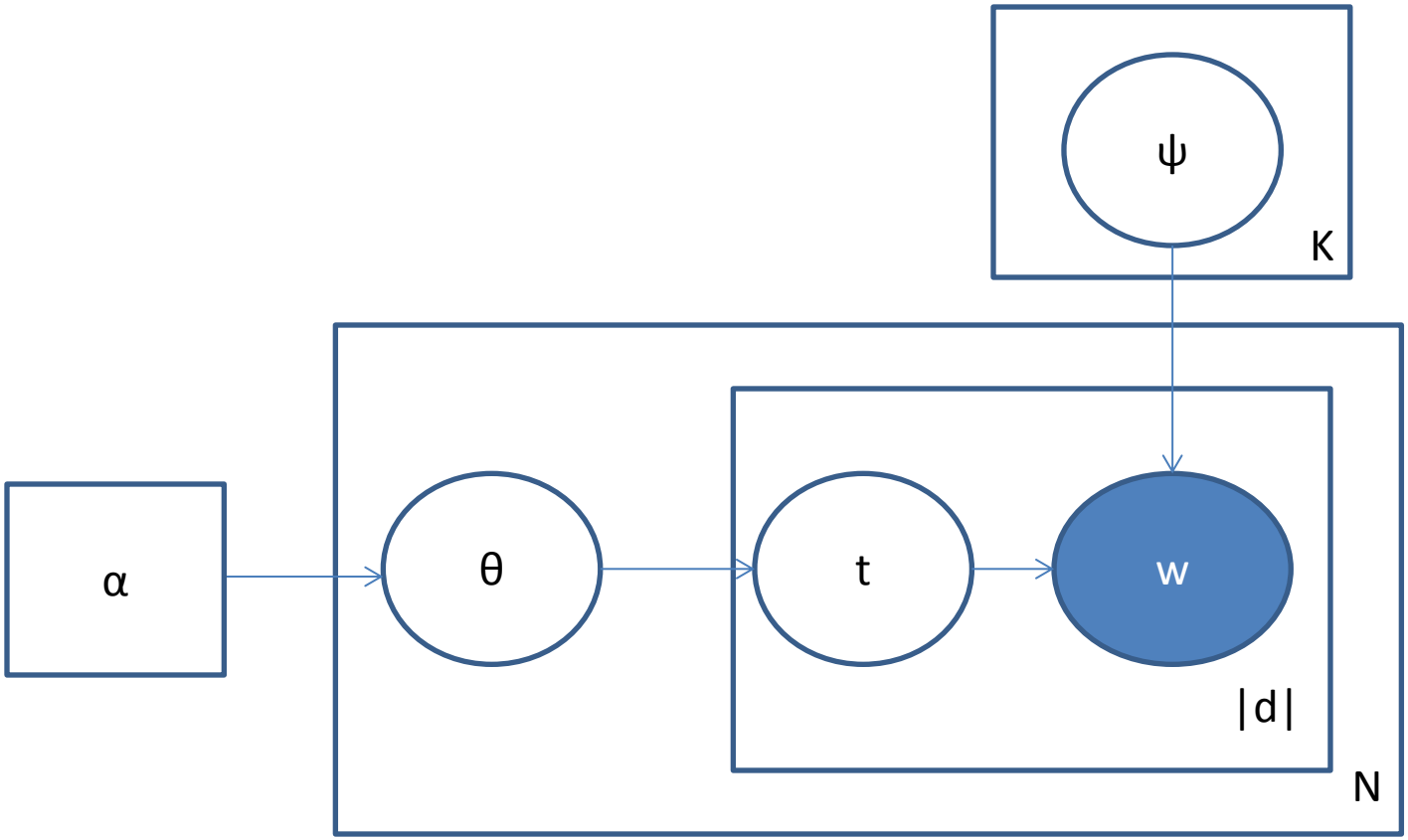
LDA in plate notation



LDA in plate notation



LDA in plate notation



LDA in plate notation