

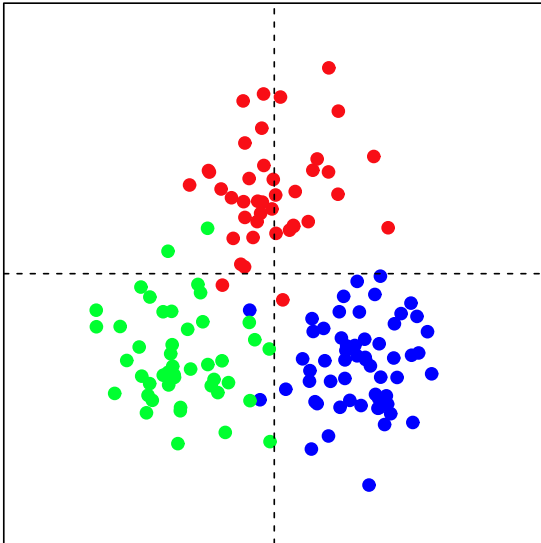
# Lecture 3 – Part 1

## Unsupervised Learning

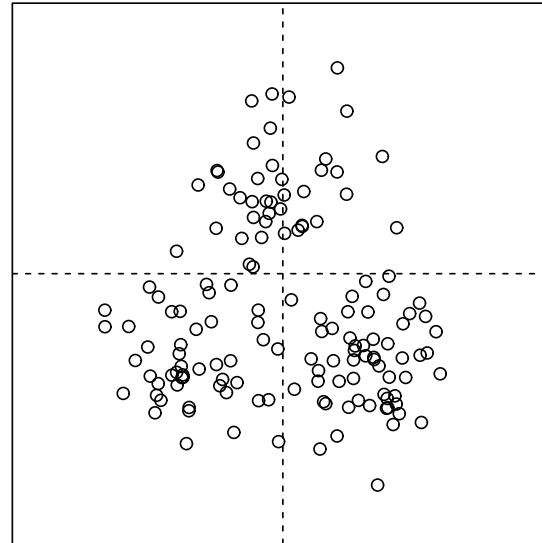
(January 23, 2015)

Mu Zhu  
University of Waterloo

**(a)**



**(b)**



(a) data from three groups; (b) group label unobserved

# Clustering

- **unsupervised** classification
- $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$
- given  $y_i = k$ ,  $\mathbf{x}_i \sim p_k(\mathbf{x})$ , for  $k = 1, 2, \dots, K$
- but class label  $y_i \in \{1, 2, \dots, K\}$  **unobserved**
- often, assumes  $p_k(\cdot) = p(\cdot; \boldsymbol{\theta}_k)$  — same family, different parameters

# Latent Variables

- suppose  $K$  is known
- introduce latent variables

$$z_{ik} = \begin{cases} 1, & \text{if } \mathbf{x}_i \text{ belongs to class } k; \\ 0, & \text{otherwise} \end{cases}$$

with  $\pi_k \equiv \mathbb{P}(z_{ik} = 1)$ , where  $\sum \pi_k = 1$

# Likelihood

- parameters

$$\Theta = \{\theta_k, \pi_k : k = 1, 2, \dots, K\}$$

- observed data

$$\mathbf{X} = \{\mathbf{x}_i : i = 1, 2, \dots, n\}$$

- unobserved “data”

$$\mathbf{Z} = \{z_{ik} : i = 1, 2, \dots, n; k = 1, 2, \dots, K\}$$

- likelihood

$$\mathcal{L}(\Theta; \mathbf{X}, \mathbf{Z}) \propto \prod_i \prod_k [p(\mathbf{x}_i; \theta_k)]^{z_{ik}} \times \pi_k^{z_{ik}}$$

# EM Algorithm

- log-likelihood (aside from constant)

$$\ell(\Theta; \mathbf{X}, \mathbf{Z}) = \sum_i \sum_k z_{ik} \log [p(\mathbf{x}_i; \theta_k)] + \sum_i \sum_k z_{ik} \log(\pi_k)$$

- would like to maximize  $\ell(\Theta; \mathbf{X}, \mathbf{Z})$  over  $\Theta$  but  $\mathbf{Z}$  missing
- idea:
  - given  $\Theta$ , replace  $\mathbf{Z}$  with  $\tilde{\mathbf{Z}} = \mathbb{E}(\mathbf{Z} | \mathbf{X}, \Theta)$
  - given  $\tilde{\mathbf{Z}}$ , estimate  $\Theta$  by maximizing over  $\ell(\Theta; \mathbf{X}, \tilde{\mathbf{Z}})$
  - iterate

# EM Algorithm

- **E-Step**: compute  $\mathbb{E}(z_{ik} | \mathbf{X}, \Theta)$
- **M-Step**: maximize

$$Q(\Theta; \mathbf{X}) = \sum_i \sum_k \mathbb{E}(z_{ik} | \mathbf{X}, \Theta) \log [p(\mathbf{x}_i; \theta_k)] +$$
$$\sum_i \sum_k \mathbb{E}(z_{ik} | \mathbf{X}, \Theta) \log(\pi_k)$$

over  $\Theta$

# M-Step

- write

$$w_{ik} \equiv \mathbb{E}(z_{ik} | \mathbf{X}, \Theta)$$

- for the “ $\theta_k$  part”, just **weighted** maximum likelihood

$$\max_{\theta_k} \sum_i w_{ik} \log [p(\mathbf{x}_i; \theta_k)] \quad \forall k \quad (1)$$

- for the “ $\pi_k$  part”, usual **multinomial** likelihood with  $w_{ik}$

$$\max_{\pi_1, \dots, \pi_K} \sum_i \sum_k w_{ik} \log(\pi_k) \quad \text{s.t.} \quad \sum_k \pi_k = 1 \quad (2)$$



# M-Step

## Exercise

- (a) Solve (1) for the case of  $p_k \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ .
- (b) Solve (2).
- (c) What would happen if we modelled each  $p_k$  as  $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$  instead? (Hard)
- (d) What can we do in order to strike a compromise between (a) and (c)? [See, e.g., Fraley & Raftery (2002; *JASA*).]

# E-Step

Finally,

$$\begin{aligned}w_{ik} &\equiv \mathbb{E}(z_{ik} | \mathbf{X}, \Theta) \\ &= \mathbb{P}(z_{ik} = 1 | \mathbf{X}, \Theta) \\ &= \mathbb{P}(z_{ik} = 1 | \mathbf{x}_i, \Theta) \\ &= \frac{\pi_k p(\mathbf{x}_i; \boldsymbol{\theta}_k)}{\pi_1 p(\mathbf{x}_i; \boldsymbol{\theta}_1) + \dots + \pi_K p(\mathbf{x}_i; \boldsymbol{\theta}_K)}.\end{aligned}$$

**Remark** The denominator above shows the EM algorithm has just allowed us to fit a [mixture distribution](#),

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k p(\mathbf{x}; \boldsymbol{\theta}_k),$$

to the sample  $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ . Question: How to [cluster](#) each  $\mathbf{x}_i$ ?

## *K*-Means as a Special Case

- assume  $p_k(\mathbf{x}) \sim \mathcal{N}(\boldsymbol{\mu}_k, \sigma^2 \mathbf{I})$
- **E-Step**: in addition to computing the “soft” parameter  $w_{ik}$ , also make a “hard” decision

$$\hat{z}_{ik} = \begin{cases} 1, & \text{if } w_{ik} = \max_{k'} w_{ik'} \\ 0, & \text{otherwise} \end{cases}$$

- **M-Step**: do the same with  $\hat{z}_{ik}$  in place of  $w_{ik}$

**Exercise** The use of  $\hat{z}_{ik}$  in place of  $w_{ik}$  “happens automatically” if we let  $\sigma^2 \rightarrow 0$ . (Why is this algorithm called *K*-means?)

# Variation for Text Data

- each  $\mathbf{x}_i = (x_{i1}, \dots, x_{id})^\top$  a text document
- $x_{ij}$  = count of word  $j$  in document  $i$  (**bag of words**)
- want to cluster  $\mathbf{x}_1, \dots, \mathbf{x}_n$  into  $K$  groups
- can use the same framework, except choose

$$p_k(\mathbf{x}_i) \sim \text{multinomial}(m_i; \boldsymbol{\theta}_k),$$

where  $\boldsymbol{\theta}_k = (\theta_{k1}, \dots, \theta_{kd})^\top$ , and  $m_i$  = total number of words in document  $i$  (treated as fixed constants)

- many sophisticated extensions, e.g., **latent Dirichlet allocation** (Blei, Ng & Jordan, **2003**; *JMLR*)

# Latent Dirichlet Allocation

- instead of modelling each **document** ( $\mathbf{x}_i$ ) to be a mixture

$$\mathcal{L}(\Theta; \mathbf{X}) = \prod_i \left[ \sum_k \pi_k p(\mathbf{x}_i; \theta_k) \right],$$

model each **word** ( $x_{ij}$ ) to be a mixture, and allow each document to have different mixing proportions

$$\mathcal{L}(\Theta; \mathbf{X}) = \prod_i \prod_j \left[ \sum_k \pi_{ik} p(x_{ij}; \theta_k) \right]$$

- instead of EM, put (Dirichlet) **prior distributions** on

both  $(\pi_{i1}, \dots, \pi_{iK})^T$  and  $(\theta_{k1}, \dots, \theta_{kd})^T$

and use a **Bayesian** approach

# Empirical Bayes

- data:

$$x_1, x_2, \dots, x_n$$

- model:

$$x_i \sim f(\cdot | \theta_i), \quad x_1, \dots, x_n \text{ all independent}$$

- source of information for each  $\theta_i$ :

$x_i$  alone

# Empirical Bayes

- now, put (common) **prior** on each  $\theta_i$ :

$$\theta_i \sim \pi(\cdot|\psi),$$

where  $\psi$  is the parameter for the prior

- then, **posterior** of  $\theta_i$  given data is

$$p(\theta_i|x_i, \psi) = \frac{f(x_i|\theta_i)\pi(\theta_i|\psi)}{\int f(x_i|\theta_i)\pi(\theta_i|\psi)d\theta_i}$$

- source of information for each  $\theta_i$ :

both  $x_i$  and  $\psi$

- next, consider  $\psi$  unknown

# Empirical Bayes

- in Bayesian framework,  $f(x_i|\theta_i)$  is a **conditional** distribution (of  $x_i$  given  $\theta_i$ )

- can integrate out  $\theta_i$  to get the **marginal** distribution of  $x_i$ ,

$$m(x_i|\psi) = \int f(x_i|\theta_i)\pi(\theta_i|\psi)d\theta_i$$

- $x_1, \dots, x_n \stackrel{iid}{\sim} m(\cdot|\psi)$  all contain some information about  $\psi$

- can find an estimate  $\hat{\psi}$ , using all  $x_i$ , and use the “plug-in” posterior,  $p(\theta_i|x_i, \hat{\psi})$

- source of information for each  $\theta_i$  [based on  $p(\theta_i|x_i, \hat{\psi})$ ]:

all  $x_1, \dots, x_n$  [not just  $x_i$ ; **borrow strength**]



## Ex: James-Stein Estimator

**Set-up** Suppose

$$\begin{aligned}x_i &\sim \text{N}(\theta_i, \sigma^2) \quad \text{all independent;} \\ \theta_i &\sim \text{N}(\mu, \tau^2),\end{aligned}$$

where  $\sigma^2$  (but not  $\tau^2$ ) is assumed to be known.

**Exercise** The posterior of  $\theta_i$  is normal with mean

$$\left[ \frac{\tau^2}{\sigma^2 + \tau^2} \right] x_i + \left[ \frac{\sigma^2}{\sigma^2 + \tau^2} \right] \mu = \mu + \left[ 1 - \frac{\sigma^2}{\sigma^2 + \tau^2} \right] (x_i - \mu),$$

which can be used as a (Bayesian) point estimate of  $\theta_i$  (if  $\mu, \tau^2$  are known).

## Ex: James-Stein Estimator

**Empirical estimate of  $\mu$  and  $1/(\sigma^2 + \tau^2)$**  Marginally,

$$x_i | \mu \stackrel{iid}{\sim} N(\mu, \sigma^2 + \tau^2),$$

so we can estimate  $\mu$  by

$$\hat{\mu} = \bar{x} = \frac{x_1 + \dots + x_n}{n}.$$

**Exercise** Show that, if  $x_1, \dots, x_n \stackrel{iid}{\sim} N(\mu, \sigma^2 + \tau^2)$ , then

$$\mathbb{E} \left[ \frac{n-3}{\sum (x_i - \bar{x})^2} \right] = \frac{1}{\sigma^2 + \tau^2}.$$

## Ex: James-Stein Estimator

The J-S estimator

$$\hat{\theta}_i^{JS} = \bar{x} + \left[ 1 - \frac{(n-3)\sigma^2}{\sum (x_i - \bar{x})^2} \right] (x_i - \bar{x})$$

has been shown to have smaller **mean-squared error** than  $\hat{\theta}_i = x_i$  for  $n > 3$ .

# Hierarchical Bayes

- alternatively, put prior on unknown  $\psi$  (another layer), e.g.,

$$x_i \sim f(\cdot|\theta_i);$$

$$\theta_i \sim \pi(\cdot|\psi);$$

$$\psi \sim \xi(\cdot)$$

- here, we have suppressed any parameters for  $\xi$  — eventually, *something* must be treated as fixed
- instead of point estimate  $\hat{\psi}$ , rely on posterior of  $\psi$
- posterior of  $\psi$  depends on all  $x_i$  ... still **borrow strength**

# Some Challenges

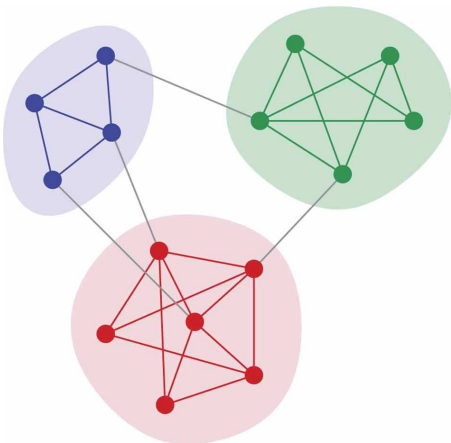
- for EM
  - many local solutions in practice
- in reality,  $K$  unknown
  - use a different prior for  $(\pi_1, \dots, \pi_K)^T$
  - e.g., [Chinese restaurant process](#), etc
- for Bayesian approach
  - effective computational techniques
  - e.g., [collapsed Gibbs sampler](#), [variational inference](#), etc

# Summary

- key ideas:
  - unsupervised learning; clustering
  - latent variables; mixture models
  - empirical Bayes; hierarchical Bayes; borrow strength
- specific methods:
  - EM algorithm
  - $K$ -means
  - bag of words; latent Dirichlet allocation
  - James-Stein estimator

## Next ...

- lecture @ 2 pm by Professor E. Kolaczyk on [network data](#)
- a short, 10-minute break
- current research with collaborators on clustering nodes in a **transactional network** (e.g., an NBA basketball game)



Community detection in networks. [Image source: Newman (2012), *Nature Physics* **8**, pp. 25–31.]