DSC291: Machine Learning with Few Labels

Unsupervised Learning

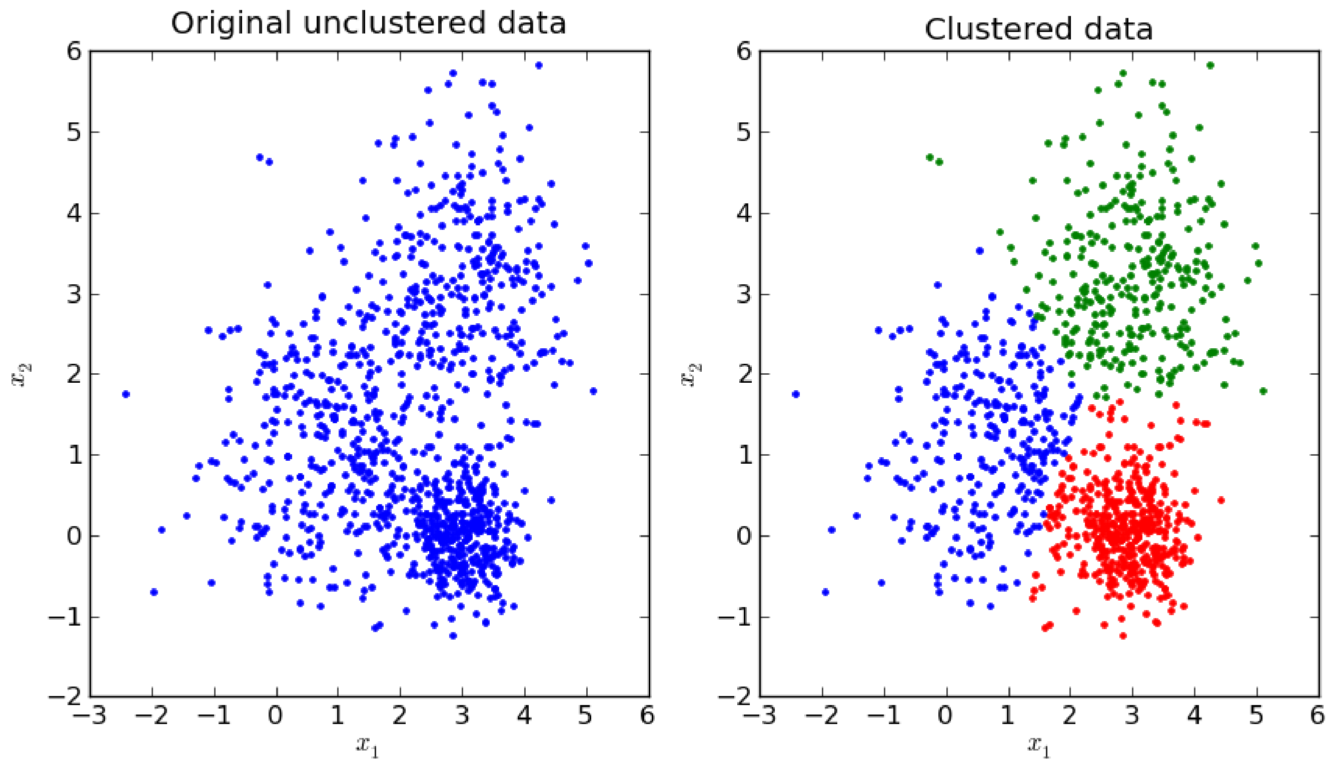
Zhiting Hu

Lecture 4, April 10, 2025

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

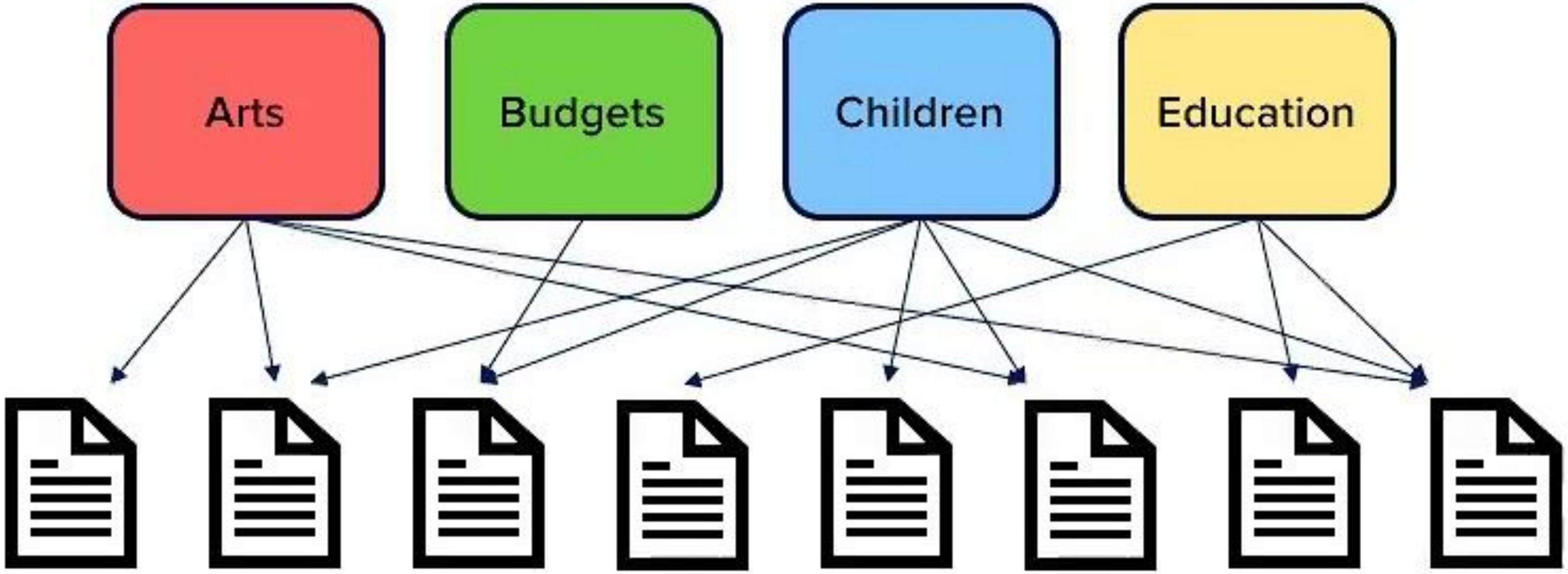
Unsupervised Learning

- Each data instance is partitioned into two parts:
  - observed variables $\boldsymbol{x}$
  - latent (unobserved) variables $\boldsymbol{z}$
- Want to learn a model $p_\theta(\boldsymbol{x}, \boldsymbol{z})$



Unsupervised Learning

- Each data instance is partitioned into two parts:
  - observed variables $\boldsymbol{x}$
  - latent (unobserved) variables $\boldsymbol{z}$
- Want to learn a model $p_\theta(\boldsymbol{x}, \boldsymbol{z})$

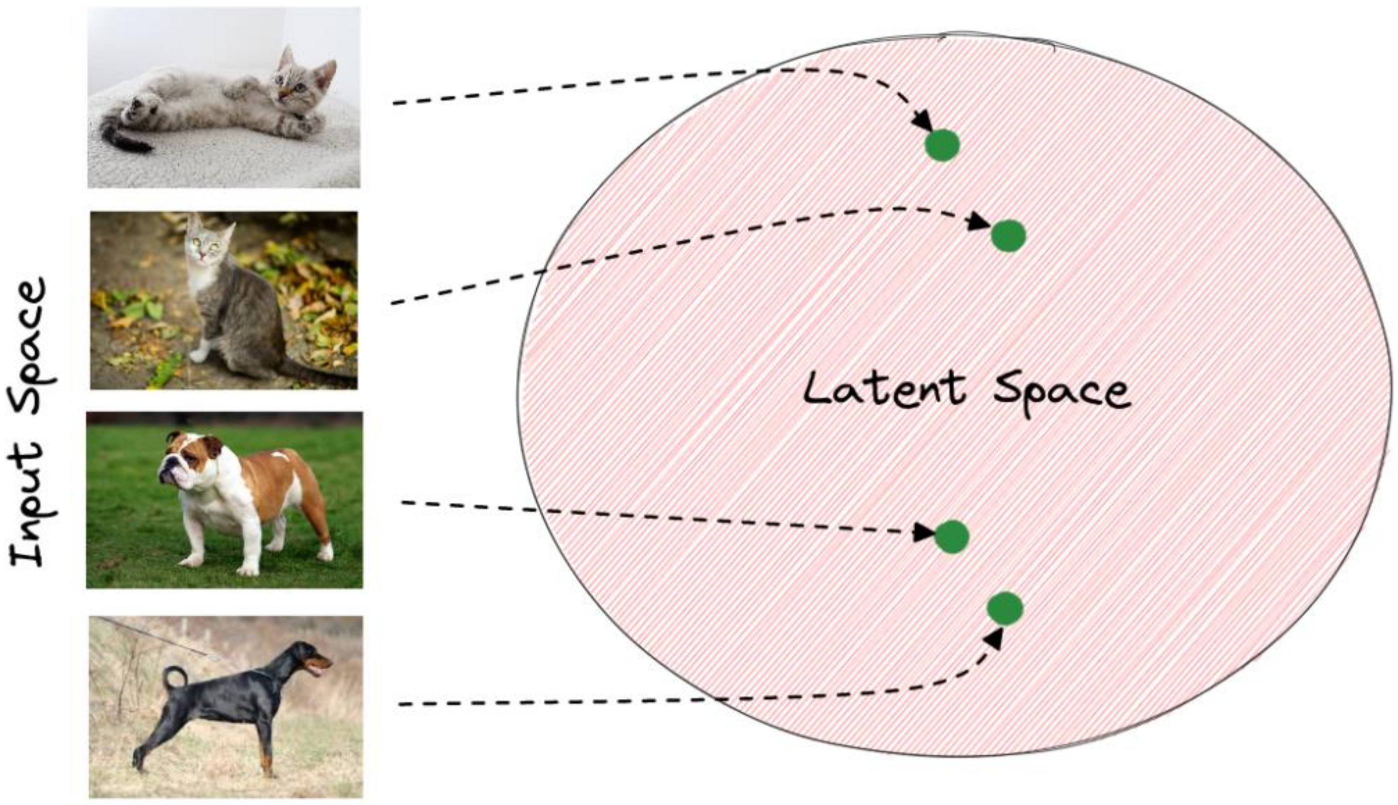


Unsupervised Learning

- Each data instance is partitioned into two parts:
  - observed variables $\boldsymbol{x}$
  - latent (unobserved) variables $\boldsymbol{z}$
- Want to learn a model $p_\theta(\boldsymbol{x}, \boldsymbol{z})$

Unsupervised Learning

- Each data instance is partitioned into two parts:
  - observed variables $\boldsymbol{x}$
  - latent (unobserved) variables $\boldsymbol{z}$
- Want to learn a model $p_\theta(\boldsymbol{x}, \boldsymbol{z})$



Why is Unsupervised Learning Harder?

- **Complete log likelihood:** if both $\boldsymbol{x}$ and $\boldsymbol{z}$ can be observed, then

$$\ell_c(\theta; \boldsymbol{x}, \boldsymbol{z}) = \log p(\boldsymbol{x}, \boldsymbol{z}|\theta) = \log p(\boldsymbol{z}|\theta_z) + \log p(\boldsymbol{x}|\boldsymbol{z}, \theta_x)$$

  - Decomposes into a sum of factors, the parameter for each factor can be estimated separately

Now $\boldsymbol{z}$ is not observed:

- **Incomplete (or marginal) log likelihood:** with $\boldsymbol{z}$ unobserved, our objective becomes the log of a marginal probability:

$$\ell(\theta; \boldsymbol{x}) = \log p(\boldsymbol{x}|\theta)$$

Why is Unsupervised Learning Harder?

- **Complete log likelihood:** if both $\boldsymbol{x}$ and $\boldsymbol{z}$ can be observed, then

$$\ell_c(\theta; \boldsymbol{x}, \boldsymbol{z}) = \log p(\boldsymbol{x}, \boldsymbol{z}|\theta) = \log p(\boldsymbol{z}|\theta_z) + \log p(\boldsymbol{x}|\boldsymbol{z}, \theta_x)$$

  - Decomposes into a sum of factors, the parameter for each factor can be estimated separately

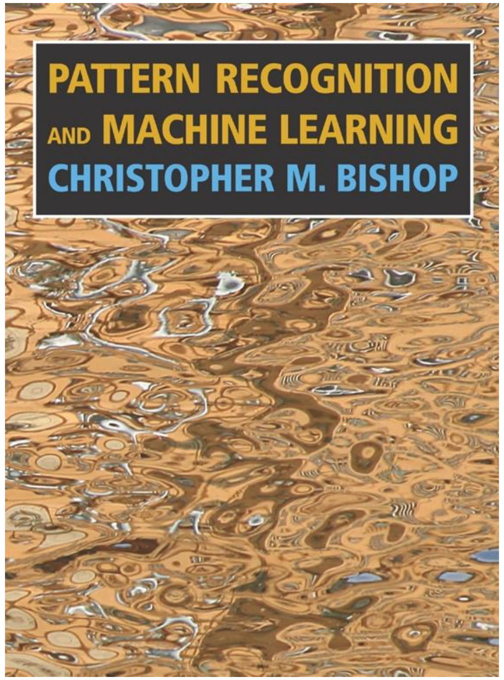
Now $\boldsymbol{z}$ is not observed:

- **Incomplete (or marginal) log likelihood:** with $\boldsymbol{z}$ unobserved, our objective becomes the log of a marginal probability:

$$\ell(\theta; \boldsymbol{x}) = \log p(\boldsymbol{x}|\theta) = \log \sum_z p(\boldsymbol{x}, \boldsymbol{z}|\theta)$$

  - All parameters become coupled together
  - In other models when $\boldsymbol{z}$ is complex (continuous) variables (as we'll see later), marginalization over $\boldsymbol{z}$ is intractable.

Expectation Maximization (EM)



Expectation Maximization (EM): Intuition

Expectation Maximization (EM): Intuition

- Supervised MLE is easy: $$\max_\theta \ \ell_c(\theta; \boldsymbol{x}, \boldsymbol{z}) = \log p(\boldsymbol{x}, \boldsymbol{z} | \theta)$$
  - Observe both $\boldsymbol{x}$ and $\boldsymbol{z}$
- Unsupervised MLE is hard: $$\max_\theta \ \ell(\theta; \boldsymbol{x}) = \log p(\boldsymbol{x} | \theta) = \log \sum_z p(\boldsymbol{x}, \boldsymbol{z} | \theta)$$
  - Observe only $\boldsymbol{x}$
- EM, intuitively:

E-step: $q(\boldsymbol{z}|\boldsymbol{x}) = p(\boldsymbol{z}|\boldsymbol{x}, \theta)$

We don't actually observe q, let's estimate it

M-step: $\max_\theta \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\, \log p(\boldsymbol{x}, \boldsymbol{z}|\theta)\,]$

Let's "pretend" we also observe $\boldsymbol{z}$ (its distribution)

Expectation Maximization (EM): Intuition

- Supervised MLE is easy: $$\max_{\theta} \ \ell_c(\theta; \boldsymbol{x}, \boldsymbol{z}) = \log p(\boldsymbol{x}, \boldsymbol{z}|\theta)$$
  - Observe both $\boldsymbol{x}$ and $\boldsymbol{z}$
- Unsupervised MLE is hard: $$\max_{\theta} \ \ell(\theta; \boldsymbol{x}) = \log p(\boldsymbol{x}|\theta) = \log \sum_{z} p(\boldsymbol{x}, \boldsymbol{z}|\theta)$$
  - Observe only $\boldsymbol{x}$
- EM, intuitively:

E-step: $$q^{t+1}(\boldsymbol{z}|\boldsymbol{x}) = p(\boldsymbol{z}|\boldsymbol{x}, \theta^t)$$

We don't actually observe q, let's estimate it

M-step: $$\max_{\theta} \mathbb{E}_{q^{t+1}(\boldsymbol{z}|\boldsymbol{x})}[\log p(\boldsymbol{x}, \boldsymbol{z}|\theta)]$$

Let's "pretend" we also observe $\boldsymbol{z}$ (its distribution)

This is an iterative process

Expectation Maximization (EM): Intuition

- Supervised MLE is easy: $\max_\theta \ \ell_c(\theta; \boldsymbol{x}, \boldsymbol{z}) = \log p(\boldsymbol{x}, \boldsymbol{z}|\theta)$
  - Observe both $\boldsymbol{x}$ and $\boldsymbol{z}$
- Unsupervised MLE is hard: $\max_\theta \ \ell(\theta; \boldsymbol{x}) = \log p(\boldsymbol{x}|\theta) = \log \sum_z p(\boldsymbol{x}, \boldsymbol{z}|\theta)$
  - Observe only $\boldsymbol{x}$
- EM, intuitively:

E-step: $q^{t+1}(\boldsymbol{z}|\boldsymbol{x}) = p(\boldsymbol{z}|\boldsymbol{x}, \theta^t)$

We don't actually observe q, let's estimate it

M-step: $\max_\theta \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p(\boldsymbol{x}, \boldsymbol{z}|\theta)]$

Let's "pretend" we also observe z (its distribution)

This is an iterative process

Expectation Maximization (EM): Intuition

- Supervised MLE is easy: $\max_{\theta} \ell_c(\theta; \boldsymbol{x}, \boldsymbol{z}) = \log p(\boldsymbol{x}, \boldsymbol{z}|\theta)$
  - Observe both $\boldsymbol{x}$ and $\boldsymbol{z}$
- Unsupervised MLE is hard: $\max_{\theta} \ell(\theta; \boldsymbol{x}) = \log p(\boldsymbol{x}|\theta) = \log \sum_{z} p(\boldsymbol{x}, \boldsymbol{z}|\theta)$
  - Observe only $\boldsymbol{x}$
- EM, intuiti

$$\ell(\theta; \boldsymbol{x}) \geq \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p(\boldsymbol{x}, \boldsymbol{z}|\theta)] + H(q)$$

E

ly observe q, let's estimate it

M-step: $\max_{\theta} \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p(\boldsymbol{x}, \boldsymbol{z}|\theta)]$

Let's "pretend" we also observe z (its distribution)

This is an iterative process

Expectation Maximization (EM): Intuition

- **Question**: show that $\ell(\theta; \boldsymbol{x}) \geq \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p(\boldsymbol{x}, \boldsymbol{z}|\theta)] + H(q)$

Expectation Maximization (EM): Intuition

- **Question**: show that

$$\ell(\theta; \boldsymbol{x}) \geq \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p(\boldsymbol{x}, \boldsymbol{z}|\theta)] + H(q)$$
$$= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z}|\theta)}{q(\boldsymbol{z}|\boldsymbol{x})}\right]$$

Expectation Maximization (EM): Intuition

- **Question**: show that

$$\ell(\theta; \boldsymbol{x}) \geq \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p(\boldsymbol{x}, \boldsymbol{z}|\theta)] + H(q)$$
$$= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z}|\theta)}{q(\boldsymbol{z}|\boldsymbol{x})}\right]$$

- **Hint**: first show that

$$\ell(\theta; \boldsymbol{x}) = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z}|\theta)}{q(\boldsymbol{z}|\boldsymbol{x})}\right] + \mathrm{KL}\big(q(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}|\boldsymbol{x}, \theta)\big)$$

  - Since KL divergence is non-negative, we arrive at the conclusion

Expectation Maximization (EM): Intuition

- **Question**: show that

$$\ell(\theta;\boldsymbol{x}) \geq \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p(\boldsymbol{x},\boldsymbol{z}|\theta)] + H(q)$$
$$= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x},\boldsymbol{z}|\theta)}{q(\boldsymbol{z}|\boldsymbol{x})}\right]$$

- **Hint**: first show that

Evidence Lower Bound (ELBO)

$$\ell(\theta;\boldsymbol{x}) = \boxed{\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x},\boldsymbol{z}|\theta)}{q(\boldsymbol{z}|\boldsymbol{x})}\right]} + \mathrm{KL}\big(q(\boldsymbol{z}|\boldsymbol{x})\,||\,p(\boldsymbol{z}|\boldsymbol{x},\theta)\big)$$

$$= -\boxed{F(q,\theta)} + \mathrm{KL}\big(q(\boldsymbol{z}|\boldsymbol{x})\,||\,p(\boldsymbol{z}|\boldsymbol{x},\theta)\big)$$

Variational free energy

$$F(q,\theta) = -\,\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p(\boldsymbol{x},\boldsymbol{z}|\theta)] - H(q)$$

Lower Bound and Free Energy

- Variational free energy:

$$F(q,\theta) = -\,\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p(\boldsymbol{x},\boldsymbol{z}|\theta)] - H(q)$$

- The EM algorithm is coordinate-decent on F
  - At each step t:
    - E-step: $q^{t+1} = \arg\min_q F\left(q, \theta^t\right) \Rightarrow q^{t+1}(\boldsymbol{z}|\boldsymbol{x}) = p(\boldsymbol{z}|\boldsymbol{x},\theta^t)$
    - M-step: $\theta^{t+1} = \arg\min_\theta F\left(q^{t+1}, \theta^t\right) \Rightarrow \max_\theta \mathbb{E}_{q^{t+1}(\boldsymbol{z}|\boldsymbol{x})}[\log p(\boldsymbol{x},\boldsymbol{z}|\theta)]$

Lower Bound and Free Energy

- Variational free energy:

$$F(q,\theta) = -\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p(\boldsymbol{x},\boldsymbol{z}|\theta)] - H(q)$$

- The EM algorithm is coordinate-decent on F
  - At each step t:
    - E-step: $q^{t+1} = \arg\min_q F\left(q,\theta^t\right)$ ⇨ ? $q^{t+1}(\boldsymbol{z}|\boldsymbol{x}) = p(\boldsymbol{z}|\boldsymbol{x},\theta^t)$
    - M-step: $\theta^{t+1} = \arg\min_\theta F\left(q^{t+1},\theta^t\right)$ ⇨ $\max_\theta \mathbb{E}_{q^{t+1}(\boldsymbol{z}|\boldsymbol{x})}[\log p(\boldsymbol{x},\boldsymbol{z}|\theta)]$

E-step: minimization of $F(q,\theta)$ w.r.t q

- **Question**: show that that optimal solution of E-step is

$$q^{t+1} = \text{argmin}_q \, F(q,\theta^t) = p(\boldsymbol{z}|\boldsymbol{x},\theta^t)$$

  - I.e., the posterior distribution over the latent variables given the data and the current parameters.

- **Hint:** use the fact

$$\underbrace{\ell(\theta^t;\boldsymbol{x})}_{\text{Independent of } q} = -F(q,\theta^t) + \underbrace{\mathrm{KL}\big(q(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}|\boldsymbol{x},\theta)\big)}_{\geq 0}$$

  - $F(q,\theta^t)$ is minimized when $\mathrm{KL}\big(q(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}|\boldsymbol{x},\theta^t)\big) = 0$, which is achieved only when $q(\boldsymbol{z}|\boldsymbol{x}) = p(\boldsymbol{z}|\boldsymbol{x},\theta^t)$

Lower Bound and Free Energy

- Variational free energy:

$$F(q,\theta) = -\,\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p(\boldsymbol{x},\boldsymbol{z}|\theta)] - H(q)$$

- The EM algorithm is coordinate-decent on F
  - At each step t:
    - E-step: $q^{t+1} = \arg\min_{q} F\left(q,\theta^{t}\right)$ $\Rightarrow$ $q^{t+1}(\boldsymbol{z}|\boldsymbol{x}) = p(\boldsymbol{z}|\boldsymbol{x},\theta^{t})$ ✓
    - M-step: $\theta^{t+1} = \arg\min_{\theta} F\left(q^{t+1},\theta^{t}\right)$? $\Rightarrow$ $\max_{\theta} \mathbb{E}_{q^{t+1}(\boldsymbol{z}|\boldsymbol{x})}[\,\log p(\boldsymbol{x},\boldsymbol{z}|\theta)]$

EM Algorithm: Quick Summary

- Observed variables $\boldsymbol{x}$, latent variables $\boldsymbol{z}$
- To learn a model $p(\boldsymbol{x}, \boldsymbol{z}|\theta)$, we want to maximize the marginal log-likelihood

$$\ell(\theta; \boldsymbol{x}) = \log p(\boldsymbol{x}|\theta) = \log \sum_{z} p(\boldsymbol{x}, \boldsymbol{z}|\theta)$$

  - But it's too difficult
- EM algorithm:
  - maximize a lower bound of $\ell(\theta; \boldsymbol{x})$
  - Or equivalently, minimize an upper bound of $-\ell(\theta; \boldsymbol{x})$
- Key equation:

Evidence Lower Bound (ELBO)

$$\ell(\theta; \boldsymbol{x}) = \boxed{\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z}|\theta)}{q(\boldsymbol{z}|\boldsymbol{x})}\right]} + \mathrm{KL}\big(q(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}|\boldsymbol{x}, \theta)\big)$$

$$= -\boxed{F(q, \theta)} + \mathrm{KL}\big(q(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}|\boldsymbol{x}, \theta)\big)$$

Variational free energy

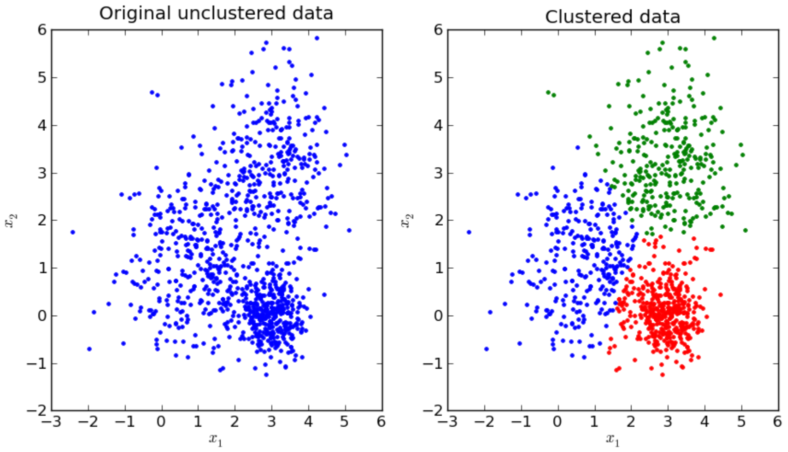
EM Algorithm: Quick Summary

- The EM algorithm is coordinate-decent on $F(q,\theta)$
  - E-step: $$q^{t+1} = \arg\min_q F\left(q, \theta^t\right) = p(\boldsymbol{z}|\boldsymbol{x}, \theta^t)$$
    - the posterior distribution over the latent variables given the data and the current parameters
  - M-step: $$\theta^{t+1} = \arg\min_\theta F\left(q^{t+1}, \theta^t\right) = \operatorname{argmax}_\theta \sum_z q^{t+1}(\boldsymbol{z}|\boldsymbol{x}) \log p(\boldsymbol{x}, \boldsymbol{z}|\theta)$$

$$\ell(\theta; \boldsymbol{x}) = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z}|\theta)}{q(\boldsymbol{z}|\boldsymbol{x})}\right] + \mathrm{KL}\big(q(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}|\boldsymbol{x}, \theta)\big)$$

$$= -F(q, \theta) + \mathrm{KL}\big(q(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}|\boldsymbol{x}, \theta)\big)$$

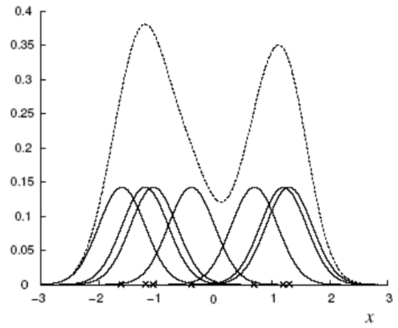
Example: Gaussian Mixture Models

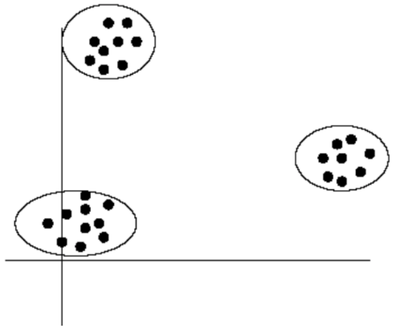


- Consider a mixture of K Gaussian components:

$$p(x_n|\mu,\Sigma) = \sum_k \pi_k N(x,|\,\mu_k,\Sigma_k)$$

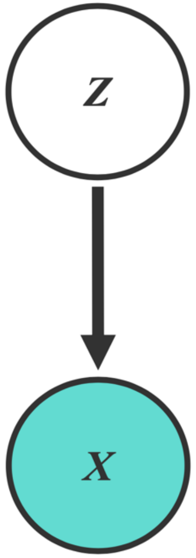
mixture proportion   mixture component





- This model can be used for unsupervised clustering.
  - This model (fit by AutoClass) has been used to discover new kinds of stars in astronomical data, etc.

Example: Gaussian Mixture Models (GMMs)



- Consider a mixture of K Gaussian components:
  - Z is a latent class indicator vector:

$$p(z_n) = \mathrm{multi}(z_n : \pi) = \prod_k (\pi_k)^{z_n^k}$$

  - X is a conditional Gaussian variable with a class-specific mean/covariance

$$p(x_n \mid z_n^k = 1, \mu, \Sigma) = \frac{1}{(2\pi)^{m/2} |\Sigma_k|^{1/2}} \exp\left\{ -\tfrac{1}{2} (x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k) \right\}$$

  - The likelihood of a sample:

Parameters to be learned:

mixture component

mixture proportion

$$\begin{aligned} p(x_n | \mu, \Sigma) &= \sum_k p(z^k = 1 \mid \pi) p(x, \mid z^k = 1, \mu, \Sigma) \\ &= \sum_{z_n} \prod_k \left((\pi_k)^{z_n^k} N(x_n : \mu_k, \Sigma_k)^{z_n^k} \right) = \sum_k \pi_k N(x, \mid \mu_k, \Sigma_k) \end{aligned}$$

Example: Gaussian Mixture Models (GMMs)

- Consider a mixture of K Gaussian components
- The expected complete log likelihood

$$\mathbb{E}_q\left[\ell_c(\boldsymbol{\theta}; x, z)\right] = \sum_n \mathbb{E}_q\left[\log p\left(z_n \mid \pi\right)\right] + \sum_n \mathbb{E}_q\left[\log p\left(x_n \mid z_n, \mu, \Sigma\right)\right]$$

$$= \sum_n \sum_k \mathbb{E}_q\left[z_n^k\right] \log \pi_k - \frac{1}{2} \sum_n \sum_k \mathbb{E}_q\left[z_n^k\right] \left((x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k) + \log |\Sigma_k| + C \right)$$

- E-step: computing the posterior of z_n given the current estimate of the parameters (i.e., π , μ, Σ)

$$p(z_n^k = 1 \mid x, \mu^{(t)}, \Sigma^{(t)}) = \frac{\pi_k^{(t)} N(x_n, \mid \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_i \pi_i^{(t)} N(x_n, \mid \mu_i^{(t)}, \Sigma_i^{(t)})}$$

$p(z_n^k = 1, x, \mu^{(t)}, \Sigma^{(t)})$

$p(x, \mu^{(t)}, \Sigma^{(t)})$

Example: Gaussian Mixture Models (GMMs)

- M-step: computing the parameters given the current estimate of z_n

$$\pi_k^* = \arg\max \langle l_c(\boldsymbol{\theta}) \rangle, \qquad \Rightarrow \frac{\partial}{\partial \pi_k} \langle l_c(\boldsymbol{\theta}) \rangle = 0, \forall k, \quad \text{s.t.} \sum_{\mathrm{k}} \pi_k = 1$$

$$\Rightarrow \quad \pi_k^* = \sum_n \langle z_n^k \rangle_{q^{(t)}} \Big/ N = \sum_n \tau_n^{k(t)} \Big/ N = \langle n_k \rangle \Big/ N$$

$$\mu_k^* = \arg\max \langle l(\boldsymbol{\theta}) \rangle, \qquad \Rightarrow \quad \mu_k^{(t+1)} = \frac{\sum_n \tau_n^{k(t)} x_n}{\sum_n \tau_n^{k(t)}}$$

$$\Sigma_k^* = \arg\max \langle l(\boldsymbol{\theta}) \rangle, \qquad \Rightarrow \quad \Sigma_k^{(t+1)} = \frac{\sum_n \tau_n^{k(t)} (x_n - \mu_k^{(t+1)})(x_n - \mu_k^{(t+1)})^T}{\sum_n \tau_n^{k(t)}}$$

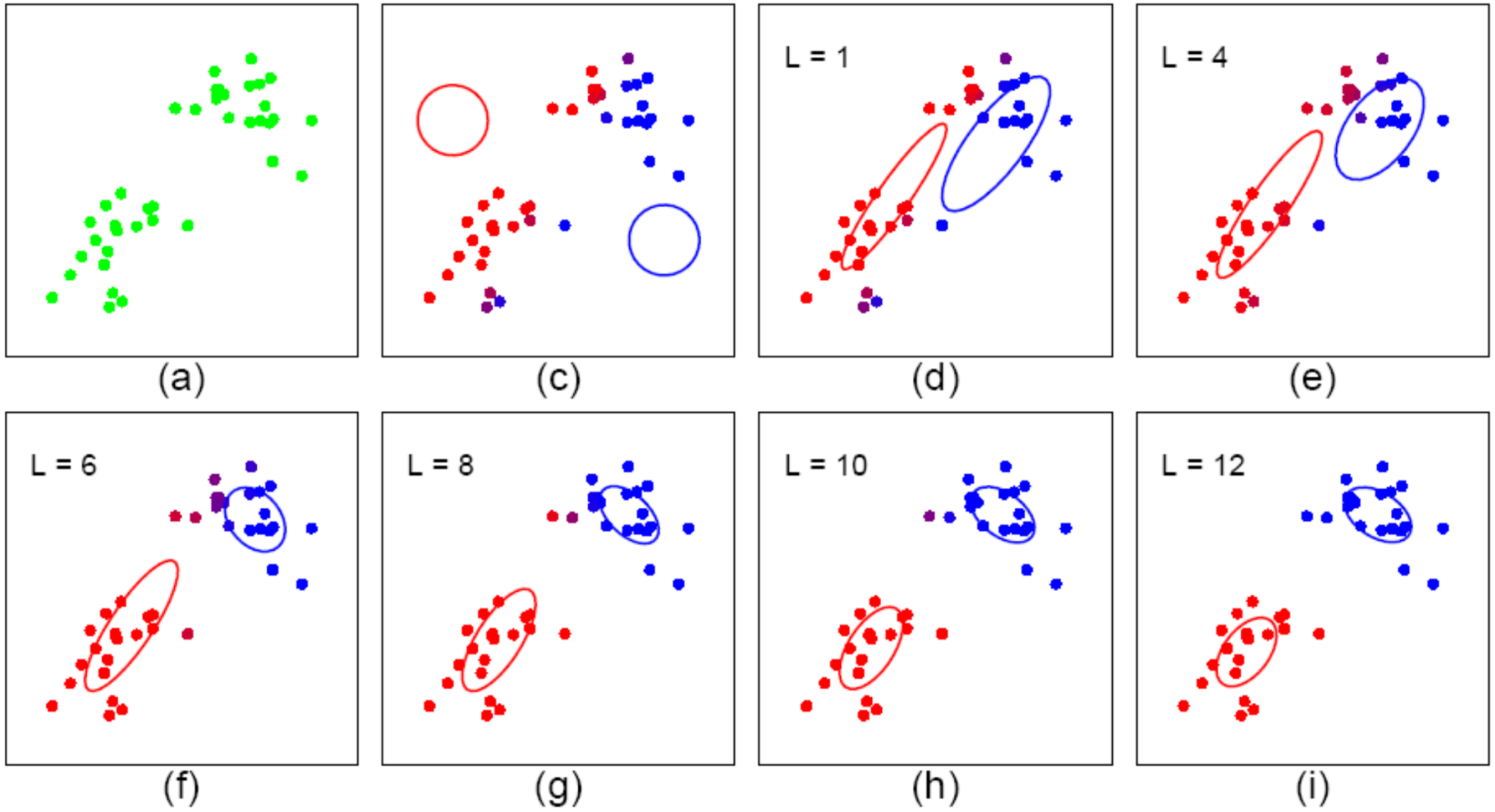
Fact:

$$\frac{\partial \log |\mathrm{A}^{-1}|}{\partial \mathrm{A}^{-1}} = \mathrm{A}^T$$

$$\frac{\partial \mathbf{x}^T \mathrm{A} \mathbf{x}}{\partial \mathrm{A}} = \mathbf{x}\mathbf{x}^T$$

Example: Gaussian Mixture Models (GMMs)

- Start: "guess" the centroid μ_k and covariance Σ_k of each of the K clusters
- Loop:



Summary: EM Algorithm

- A way of maximizing likelihood function for latent variable models. Finds MLE of parameters when the original (hard) problem can be broken up into two (easy) pieces
  - Estimate some "missing" or "unobserved" data from observed data and current parameters.
  - Using this "complete" data, find the maximum likelihood parameter estimates.

- Alternate between filling in the latent variables using the best guess (posterior) and updating the parameters based on this guess:
  - E-step:
  - M-step:

$$q^{t+1} = \arg\min_q F\left(q, \theta^t\right)$$

$$\theta^{t+1} = \arg\min_\theta F\left(q^{t+1}, \theta^t\right)$$

Questions?