PRINCIPLES OF PARAMETER ESTIMATION

CS6140

Predrag Radivojac

KHOURY COLLEGE OF COMPUTER SCIENCES

NORTHEASTERN UNIVERSITY

Spring 2021

PRELIMINARIES

Given: a set of observations $\mathcal{D} = \{x_i\}_{i=1}^n$, $x_i \in \mathcal{X}$

Objective: find a model $\hat{f} \in \mathcal{F}$ that models the phenomenon well

Requirements:

(i) the ability to generalize well

(ii) the ability to incorporate prior knowledge and assumptions

(iii) scalability

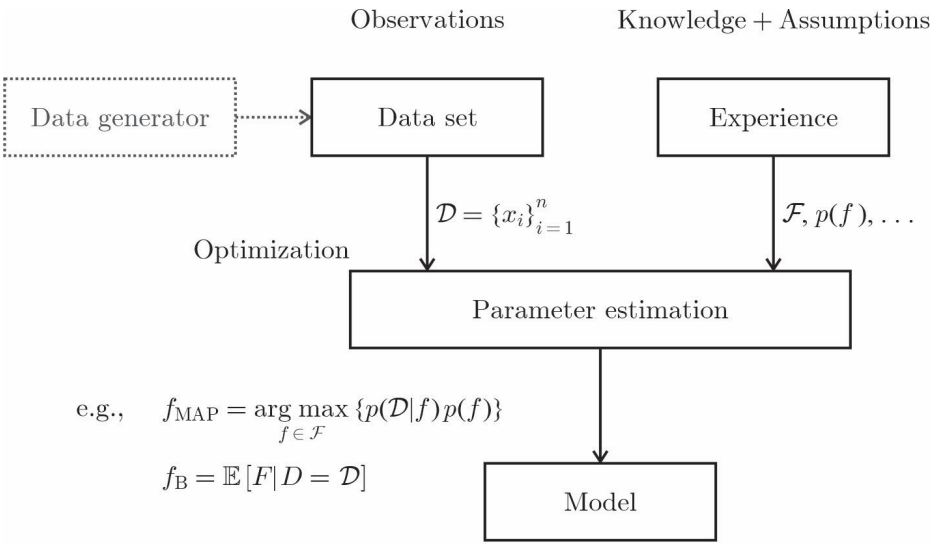
Terminology through an example: $\mathcal{D} = \{3.1, 2.4, -1.1, 0.1\}$

What is the data generator?

$\mathcal{F} = \text{Gaussian}(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$

Parameter estimation

Statistical Framework



Maximum A Posteriori (MAP) Inference

Idea:

$$f_{\text{MAP}} = \arg\max_{f \in \mathcal{F}} \{p(f|\mathcal{D})\},$$

where $p(f|\mathcal{D})$ is called the posterior distribution.

How do we calculate it?

$$p(f|\mathcal{D}) = \frac{p(\mathcal{D}|f) \cdot p(f)}{p(\mathcal{D})}$$

where $p(\mathcal{D}|f)$ = likelihood, $p(f)$ = prior, and $p(\mathcal{D})$ = data distribution.

Maximum A Posteriori (MAP) Inference

Finding the data distribution:

$$p(\mathcal{D}) = \begin{cases} \sum_{f \in \mathcal{F}} p(\mathcal{D}|f)p(f) & f : \text{discrete} \\ \\ \int_{\mathcal{F}} p(\mathcal{D}|f)p(f)df & f : \text{continuous} \end{cases}$$

We can now simplify the process if we observe that

$$\begin{aligned} p(f|\mathcal{D}) &= \frac{p(\mathcal{D}|f) \cdot p(f)}{p(\mathcal{D})} \\ &\propto p(\mathcal{D}|f) \cdot p(f) \end{aligned}$$

Maximum Likelihood (ML) Inference

Express the posterior distribution as

$$\begin{aligned} p(f|\mathcal{D}) &= \frac{p(\mathcal{D}|f) \cdot p(f)}{p(\mathcal{D})} \\ &\propto p(\mathcal{D}|f) \cdot p(f) \end{aligned}$$

Now, ignore $p(f)$ to get

$$f_{\mathrm{ML}} = \arg\max_{f \in \mathcal{F}} \{p(\mathcal{D}|f)\}$$

There are technical problems with this approach, but also reasons to use it.

MAP and ML estimates are called the point estimates.

Example: ML Inference

Example: $\mathcal{D} = \{2, 5, 9, 5, 4, 8\}$ is an i.i.d. sample from Poisson(λ), $\lambda \in \mathbb{R}^+$

Find λ

Solution: Poisson probability mass function is $p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$

$$\lambda_{\text{ML}} = \arg\max_{\lambda \in (0,\infty)} \{p(\mathcal{D}|\lambda)\}.$$

Likelihood:
$$\begin{aligned} p(\mathcal{D}|\lambda) &= p(\{x_i\}_{i=1}^n |\lambda) \\ &= \prod_{i=1}^n p(x_i|\lambda) \\ &= \frac{\lambda^{\sum_{i=1}^n x_i} \cdot e^{-n\lambda}}{\prod_{i=1}^n x_i!}. \end{aligned}$$

Example: ML Inference

Likelihood: $p(\mathcal{D}|\lambda) = \frac{\lambda^{\sum_{i=1}^{n} x_i} \cdot e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}$

Log-likelihood: $ll(\mathcal{D}, \lambda) = \ln \lambda \sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} \ln(x_i!)$

Optimization:

$$\begin{aligned} \frac{\partial ll(\mathcal{D}, \lambda)}{\partial \lambda} &= \frac{1}{\lambda} \sum_{i=1}^{n} x_i - n \\ &= 0 \end{aligned}$$

Solution:

$$\begin{aligned} \lambda_{\mathrm{ML}} &= \frac{1}{n} \sum_{i=1}^{n} x_i \\ &= 5.5 \end{aligned}$$

Example: MAP Inference

Example: $\mathcal{D} = \{2, 5, 9, 5, 4, 8\}$ is i.i.d. sample from Poisson(λ), $\lambda \in \mathbb{R}^+$.

Assume λ is taken from $\Gamma(x|k, \theta)$ with parameters $k = 3$ and $\theta = 1$.

Find λ.

Solution: Poisson: $p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$

Gamma: $\Gamma(x|k, \theta) = \frac{x^{k-1} e^{-\frac{x}{\theta}}}{\theta^k \Gamma(k)}$, where $x > 0$, $k > 0$, and $\theta > 0$.

Likelihood: $$p(\mathcal{D}|\lambda) = \frac{\lambda^{\sum_{i=1}^n x_i} \cdot e^{-n\lambda}}{\prod_{i=1}^n x_i!}$$

Prior: $$p(\lambda) = \frac{\lambda^{k-1} e^{-\frac{\lambda}{\theta}}}{\theta^k \Gamma(k)}.$$

Example: MAP Inference

Taking the logarithm

$$\begin{aligned}\ln p(\lambda|\mathcal{D}) &\propto \ln p(\mathcal{D}|\lambda) + \ln p(\lambda) \\ &= \ln \lambda(k-1+\sum_{i=1}^{n} x_i) - \lambda(n+\frac{1}{\theta}) - \sum_{i=1}^{n} \ln x_i! - k\ln\theta - \ln\Gamma(k)\end{aligned}$$

We now obtain

$$\begin{aligned}\lambda_{\text{MAP}} &= \frac{k-1+\sum_{i=1}^{n} x_i}{n+\frac{1}{\theta}} \\ &= 5\end{aligned}$$

Sufficient statistic: no other stiatistic calculated from the same sample provides any additional value to the parameter to be estimated (Fisher, 1922)

Sufficient Statistic

Statistic: funtion of the data, in the statistical sense

Sufficient statistic:

Let D be the data set random variable and $\mathcal{D}$ the observed data set.

T be random variable representing some function of the data.

$D \sim p(x_1, x_2, \ldots, x_n|\theta)$.

t is a realization of T, computed from $\mathcal{D}$.

T is sufficient if

$p(\mathcal{D}|t, \theta) = p(\mathcal{D}|t)$

Example:

$\mathcal{D} = \{x_i\}_{i=1}^n$, i.i.d., $T = \mathbb{E}[X]$, where $X \sim p(x|\theta)$

$t = \frac{1}{n}\sum_{i=1}^n x_i$

Another Example

Example: $\mathcal{D} = \{x_i\}_{i=1}^n$ is i.i.d. sample from Gaussian(μ, σ^2)

Find μ and σ

Solution: Gaussian: $p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

$$\mu_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad\qquad \sigma^2_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu_{\mathrm{ML}})^2 .$$

Relationship to Kullback-Leibler (KL) Divergence

The KL divergence between two probability distributions $p(x)$ and $q(x)$ is

$$D_{\mathrm{KL}}(p||q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx$$

Assume now the data is generated according to some $p(x|\theta_t)$. We estimated it as $p(x|\theta)$.

Let's look at the KL divergence

$$D_{\mathrm{KL}}(p(x|\theta_t)||p(x|\theta)) = \int_{-\infty}^{\infty} p(x|\theta_t) \log \frac{p(x|\theta_t)}{p(x|\theta)} dx \quad -\mathbb{E}\left[\log p(x|\theta)\right]$$

$$= \int_{-\infty}^{\infty} p(x|\theta_t) \log \frac{1}{p(x|\theta)} dx - \int_{-\infty}^{\infty} p(x|\theta_t) \log \frac{1}{p(x|\theta_t)} dx.$$

Relationship to Kullback-Leibler (KL) Divergence

$$\frac{1}{n}\sum_{i=1}^{n}\log p(x_i|\theta) \quad \overset{a.s.}{\to} \quad \mathbb{E}[\log p(x|\theta)]$$

when $n \to \infty$.

Conclusion:

When $n \to \infty$, ML estimation implies $p(x|\theta_{\mathrm{ML}}) = p(x|\theta_t)$

This usually implies $\theta_{\mathrm{ML}} = \theta_t$

Conditional Distributions

Given: a set of observations $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, $x_i, y_i \in \mathbb{R}$

Assumption: X and Y are random variables and $p(y|x) = \mathcal{N}(\mu = x, \sigma^2)$

Objective: find σ

Bayesian Approach

Idea: consider posterior risk R

$$R = \int_{\mathcal{F}} \ell(f, \hat{f}) \cdot p(f|\mathcal{D}) df$$

where $\ell(f, \hat{f})$ is some error function.

Assume $\ell(f, \hat{f}) = (f - \hat{f})^2$ and find best $\hat{f}$

$$\begin{aligned} \frac{\partial}{\partial \hat{f}} R &= 2\hat{f} - 2 \int_{\mathcal{F}} f \cdot p(f|\mathcal{D}) df \\ &= 0 \end{aligned}$$

Bayesian Approach

Solution:

$$\begin{aligned} f_{\mathrm{B}} &= \int_{\mathcal{F}} f \cdot p(f|\mathcal{D}) df \\ &= \mathbb{E}[F|D = \mathcal{D}] \end{aligned}$$

Example: $\mathcal{D} = \{2, 5, 9, 5, 4, 8\}$ is an i.i.d. sample from Poisson(λ_0), $\lambda \in \mathbb{R}^+$.

Assume λ_0 is taken from $\Gamma(x|k, \theta)$ with parameters $k = 3$ and $\theta = 1$.

Estimate λ_0.

A Note on Optimization

We have looked at these types of optimization

Likelihood $p(\mathcal{D}|\lambda)$

$$\hat{\lambda} = \underset{\lambda \in (0,\infty)}{\arg\max} \{p(x_1, x_2, \ldots, x_n|\lambda)\}$$

Was there a problem here?

Yes and no. We were lucky that $\hat{\lambda}$ was indeed a positive number.

In previous slides: $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i = 5.5 \in (0, \infty)$.

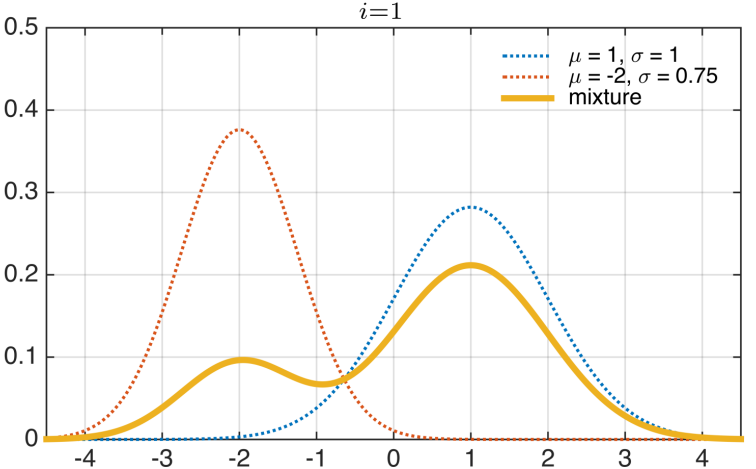
In the future, we might have to be more careful about enforcing constraints.

Revisiting Mixtures of Distributions

Mixture of $m = 2$ Gaussian distributions:

$w_1 = 0.75$, $w_2 = 0.25$

$$p(x) = \sum_{i=1}^{m} w_i p_i(x)$$



Parameter Estimation for Mixtures of Distributions

Given: a set of observations $\mathcal{D} = \{x_i\}_{i=1}^n$, $x_i \in \mathcal{X}$

$$p(x|\theta) = \sum_{j=1}^{m} w_j p(x|\theta_j). \qquad w_j \geq 0, \quad \sum_{j=1}^{m} w_j = 1.$$

where $\theta = (w_1, w_2, \ldots, w_m, \theta_1, \theta_2, \ldots, \theta_m)$

Example: Consider a mixture of $m = 2$ exponential distributions.

$p(x|\theta_j) = \lambda_j e^{-\lambda_j x}$, where $\lambda_j > 0$

$$p(x|\lambda_1, \lambda_2, w_1, w_2) = w_1 \cdot \lambda_1 e^{-\lambda_1 x} + w_2 \cdot \lambda_2 e^{-\lambda_2 x}$$

where $\lambda_1, \lambda_2 > 0$, $w_1, w_2 \geq 0$, and $w_1 = 1 - w_2$

Parameter Estimation for Mixtures of Distributions

Likelihood:

$$\begin{aligned} p(\mathcal{D}|\theta) &= \prod_{i=1}^{n} p(x_i|\theta) \\ &= \prod_{i=1}^{n} \left(\sum_{j=1}^{m} w_j p(x_i|\theta_j) \right) \end{aligned}$$

$p(\mathcal{D}|\theta)$ has $O(m^n)$ terms. It can be calculated in $O(mn)$ time as a log-likelihood.

How can we find θ? Is there a closed-form solution?

IDEA #1

Suppose we know what data point is generated by what mixing component.

That is, $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ is an i.i.d. sample from some distribution $p(x, y)$, where $y \in \mathcal{Y} = \{1, 2, \ldots, m\}$ specifies the mixing component.

$$\begin{aligned} p(\mathcal{D}|\theta) &= \prod_{i=1}^n p(x_i, y_i|\theta) \\ &= \prod_{i=1}^n p(x_i|y_i, \theta) p(y_i|\theta) \\ &= \prod_{i=1}^n w_{y_i} p(x_i|\theta_{y_i}), \end{aligned}$$

where $w_j = P(Y = j)$.

IDEA #1

Log-likelihood:

$$\begin{aligned}\log p(\mathcal{D}|\theta) &= \sum_{i=1}^{n} (\log w_{y_i} + \log p(x_i|\theta_{y_i})) \\ &= \sum_{j=1}^{m} n_j \log w_j + \sum_{i=1}^{n} \log p(x_i|\theta_{y_i}),\end{aligned}$$

where n_j is the number of data points in $\mathcal{D}$ generated by the j-th mixing component.

Constrained optimization: Let's first find $\boldsymbol{w}$

$$L(\boldsymbol{w}, \alpha) = \sum_{j=1}^{m} n_j \log w_j + \alpha \left(\sum_{j=1}^{m} w_j - 1 \right)$$

where α is the Lagrange multiplier.

IDEA #1

Set $\frac{\partial}{\partial w_k} L(\boldsymbol{w}, \alpha) = 0$ for every $k \in \mathcal{Y}$ and $\frac{\partial}{\partial \alpha} L(\boldsymbol{w}, \alpha) = 0$. Solve it.

It follows that $w_k = -\frac{n_k}{\alpha}$ and $\alpha = -n$.

$$w_k = \frac{1}{n} \sum_{i=1}^{n} I(y_i = k),$$

where $I(\cdot)$ is the indicator function.

To find all θ_j, we need to get concrete; i.e., $p(x|\theta_j) = \lambda_j e^{-\lambda_j x}$.

$$\frac{\partial}{\partial \lambda_k} \sum_{i=1}^{n} \log p(x_i|\lambda_{y_i}) = 0,$$

for each $k \in \mathcal{Y}$.

IDEA #1

Thus, assuming an exponential distribution we obtain that

$$\lambda_k = \frac{n_k}{\sum_{i=1}^{n} I(y_i = k) \cdot x_i},$$

for each $k \in \mathcal{Y}$.

Recall that

$$w_k = \frac{1}{n} \sum_{i=1}^{n} I(y_i = k)$$

Conclusion:

If the mixing component designations $\boldsymbol{y}$ are known,
the parameter estimation is greatly simplified.

IDEA #2

Suppose we know the θ but not the mixing component designations.

We start with the posterior as

$$\begin{aligned} p(\boldsymbol{y}|\mathcal{D}, \theta) &= \prod_{i=1}^{n} p(y_i|x_i, \theta) \\ &= \prod_{i=1}^{n} \frac{w_{y_i} p(x_i|\theta_{y_i})}{\sum_{j=1}^{m} w_j p(x_i|\theta_j)} \end{aligned}$$

and subsequently find the best configuration out of m^n possibilities.

Data is i.i.d. so y_i's can be estimated separately. The MAP estimate for y_i

$$\hat{y}_i = \arg\max_{y_i \in \mathcal{Y}} \left\{ \frac{w_{y_i} p(x_i|\theta_{y_i})}{\sum_{j=1}^{m} w_j p(x_i|\theta_j)} \right\}$$

Conclusion: if θ is known, "cluster" assignments are simple.

Combine the Two Ideas (Iteratively)

1. Assume θ is known, call it $\theta^{(0)}$
2. Compute $\boldsymbol{y}^{(0)}$ using $\theta^{(0)}$ as known
3. Compute $\theta^{(1)}$ using $\boldsymbol{y}^{(0)}$ as known
4. Compute $\boldsymbol{y}^{(1)}$ using $\theta^{(1)}$ as known
5. ... (until convergence)

Classification Expectation Maximization (CEM)

1. Initialize $\lambda_k^{(0)}$ and $w_k^{(0)}$ for $\forall k \in \mathcal{Y}$
2. Calculate $y_i^{(0)} = \arg\max_{k \in \mathcal{Y}} \left\{ \frac{w_k^{(0)} p(x_i|\lambda_k^{(0)})}{\sum_{j=1}^m w_j^{(0)} p(x_i|\lambda_j^{(0)})} \right\}$ for $\forall i \in \{1, 2, \ldots, n\}$
3. Set $t = 0$
4. Repeat until convergence
   (a) $w_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^n I(y_i^{(t)} = k)$
   (b) $\lambda_k^{(t+1)} = \frac{\sum_{i=1}^n I(y_i^{(t)}=k)}{\sum_{i=1}^n I(y_i^{(t)}=k) \cdot x_i}$
   (c) $y_i^{(t+1)} = \arg\max_{k \in \mathcal{Y}} \left\{ \frac{w_k^{(t+1)} p(x_i|\lambda_k^{(t+1)})}{\sum_{j=1}^m w_j^{(t+1)} p(x_i|\lambda_j^{(t+1)})} \right\}$
   (d) $t = t + 1$
5. Report $\lambda_k^{(t)}$ and $w_k^{(t)}$ for $\forall k \in \mathcal{Y}$

Also called “hard EM”.

Expectation-Maximization (EM) Algorithm

Problems with the CEM formulation:

1. We want to estimate θ
2. We do not necessarily need to compute $\boldsymbol{y}$

Main idea for the EM algorithm:

1. Take step t and assume $\theta^{(t)}$ is known
2. Maximize $\mathbb{E}[p(\mathcal{D}, \boldsymbol{Y}|\theta)|\mathcal{D}, \theta^{(t)}]$ to calculate $\theta^{(t+1)}$

Expectation-Maximization (EM) Algorithm

Expected log-likelihood of the complete data over the posterior distribution for $\boldsymbol{y}$ assuming $\theta^{(t)}$ is true:

$$\mathbb{E}[\log p(\mathcal{D}, \boldsymbol{Y}|\theta)|\mathcal{D}, \theta^{(t)}] = \begin{cases} \sum_{\boldsymbol{y}} \log p(\mathcal{D}, \boldsymbol{y}|\theta) p(\boldsymbol{y}|\mathcal{D}, \theta^{(t)}) & \boldsymbol{y} : \text{discrete} \\ \int_{\boldsymbol{y}} \log p(\mathcal{D}, \boldsymbol{y}|\theta) p(\boldsymbol{y}|\mathcal{D}, \theta^{(t)}) d\boldsymbol{y} & \boldsymbol{y} : \text{continuous} \end{cases}$$

$$\theta^{(t+1)} = \arg\max_{\theta} \left\{ \mathbb{E}[\log p(\mathcal{D}, \boldsymbol{Y}|\theta)|\mathcal{D}, \theta^{(t)}] \right\}$$

Expectation-Maximization (EM)

1. Initialize $\lambda_k^{(0)}$ and $w_k^{(0)}$ for $\forall k \in \mathcal{Y}$
2. Set $t = 0$
3. Repeat until convergence
   (a) $p_{Y_i}(k|x_i, \theta^{(t)}) = \frac{w_k^{(t)} p(x_i|\lambda_k^{(t)})}{\sum_{j=1}^{m} w_j^{(t)} p(x_i|\lambda_j^{(t)})}$ for $\forall (i, k)$
   (b) $w_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} p_{Y_i}(k|x_i, \theta^{(t)})$
   (c) $\lambda_k^{(t+1)} = \frac{\sum_{i=1}^{n} p_{Y_i}(k|x_i, \theta^{(t)})}{\sum_{i=1}^{n} x_i p_{Y_i}(k|x_i, \theta^{(t)})}$
   (d) $t = t + 1$
4. Report $\lambda_k^{(t)}$ and $w_k^{(t)}$ for $\forall k \in \mathcal{Y}$

HARD VS. SOFT EM

1. Initialize $\lambda_k^{(0)}$ and $w_k^{(0)}$ for $\forall k \in \mathcal{Y}$
2. Calculate $y_i^{(0)} = \arg\max_{k \in \mathcal{Y}} \left\{ \frac{w_k^{(0)} p(x_i|\lambda_k^{(0)})}{\sum_{j=1}^m w_j^{(0)} p(x_i|\lambda_j^{(0)})} \right\}$ for $\forall i$
3. Set $t = 0$
4. Repeat until convergence
   (a) $w_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^n I(y_i^{(t)} = k)$
   (b) $\lambda_k^{(t+1)} = \frac{\sum_{i=1}^n I(y_i^{(t)}=k)}{\sum_{i=1}^n I(y_i^{(t)}=k) \cdot x_i}$
   (c) $y_i^{(t+1)} = \arg\max_{k \in \mathcal{Y}} \left\{ \frac{w_k^{(t+1)} p(x_i|\lambda_k^{(t+1)})}{\sum_{j=1}^m w_j^{(t+1)} p(x_i|\lambda_j^{(t+1)})} \right\}$
   (d) $t = t + 1$
5. Report $\lambda_k^{(t)}$ and $w_k^{(t)}$ for $\forall k \in \mathcal{Y}$

1. Initialize $\lambda_k^{(0)}$ and $w_k^{(0)}$ for $\forall k \in \mathcal{Y}$
2. Set $t = 0$
3. Repeat until convergence
   (a) $p_{Y_i}(k|x_i, \theta^{(t)}) = \frac{w_k^{(t)} p(x_i|\lambda_k^{(t)})}{\sum_{j=1}^m w_j^{(t)} p(x_i|\lambda_j^{(t)})}$ for $\forall (i, k)$
   (b) $w_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^n p_{Y_i}(k|x_i, \theta^{(t)})$
   (c) $\lambda_k^{(t+1)} = \frac{\sum_{i=1}^n p_{Y_i}(k|x_i, \theta^{(t)})}{\sum_{i=1}^n x_i p_{Y_i}(k|x_i, \theta^{(t)})}$
   (d) $t = t + 1$
4. Report $\lambda_k^{(t)}$ and $w_k^{(t)}$ for $\forall k \in \mathcal{Y}$

Expectation-Maximization (EM) Algorithm

E-step: evaluate $p(\boldsymbol{y}|\mathcal{D}, \theta^{(t)})$

M-step: $\theta^{(t+1)} = \arg\max_{\theta} \left\{ \mathbb{E}[\log p(\mathcal{D}, \boldsymbol{Y}|\theta)|\mathcal{D}, \theta^{(t)}] \right\}$

How did we Arrive at this Solution?

1. Try to maximize likelihood $p(\mathcal{D}|\theta)$
   (a) can be difficult even as log-likelihood; e.g., we get log of a sum of some function of parameters θ that is unfriendly to differentiation
2. Recognize we have some unobserved or "hidden" variables
   (a) mixture case: we figured out that there is a "class label" vector $\boldsymbol{y}$ so we can see the *complete data* as set of pairs $\{(x_i, y_i)\}_{i=1}^n$
3. Attempt to maximize the likelihood of complete data $p(\mathcal{D}, \boldsymbol{y}|\theta)$
   (a) cannot do because vector $\boldsymbol{y}$ is unobserved

How did we Arrive at this Solution?

4. Think of an iterative process and assume we have $\theta^{(t)}$ as an approximation of θ_{ML} in step t. New goal: find $\theta^{(t+1)}$ of the next step $(t+1)$ that is a little better than $\theta^{(t)}$ from step t.

   (a) good news: we can compute the posterior of unobserved data $p(\boldsymbol{y}|\mathcal{D}, \theta^{(t)})$ since $\mathcal{D}$ and $\theta^{(t)}$ are given.

   (b) This will become the E-step.

5. To find $\theta^{(t+1)}$, try to maximize the expected likelihood of the complete data $\mathbb{E}[p(\mathcal{D}, \boldsymbol{Y}|\theta)|\mathcal{D}, \theta^{(t)}]$, where we integrate over $p(\boldsymbol{y}|\mathcal{D}, \theta^{(t)})$

   (a) this is still hard as we have to work with products instead of sums

6. Try to maximize the expected log-likelihood of the complete data $\mathbb{E}[\log p(\mathcal{D}, \boldsymbol{Y}|\theta)|\mathcal{D}, \theta^{(t)}]$

   (a) good news: we get expressions that can be simplified so we can compute $\theta^{(t+1)}$ by maximizing $\mathbb{E}[\log p(\mathcal{D}, \boldsymbol{Y}|\theta)|\mathcal{D}, \theta^{(t)}]$

   (b) This will become the M-step.

How did we Arrive at this Solution?

7. The EM algorithm iterates the E-step with the M-step.
8. Prove that maximizing $\mathbb{E}[\log p(\mathcal{D}, \boldsymbol{Y}|\theta)|\mathcal{D}, \theta^{(t)}]$ maximizes $p(\mathcal{D}|\theta)$
   (a) good news: it can be done, but it is not obvious so it has to be done.
   (b) bad news: we can only prove local maximization as the likelihood function is not convex.

Recap of Reasoning

1. Try to maximize likelihood $p(\mathcal{D}|\theta)$
2. Recognize we have some unobserved or "hidden" variables
3. Attempt to maximize the likelihood of complete data $p(\mathcal{D}, \boldsymbol{y}|\theta)$
4. Think of an iterative process and assume we have $\theta^{(t)}$ as an approximation of θ_{ML} in step t. New goal: find $\theta^{(t+1)}$ of the next step $(t+1)$ that is a little better than $\theta^{(t)}$ from step t.
5. To find $\theta^{(t+1)}$, try to maximize the expected likelihood of the complete data $\mathbb{E}[p(\mathcal{D}, \boldsymbol{Y}|\theta)|\mathcal{D}, \theta^{(t)}]$, where we integrate over $p(\boldsymbol{y}|\mathcal{D}, \theta^{(t)})$
6. Try to maximize the expected log-likelihood of the complete data $\mathbb{E}[\log p(\mathcal{D}, \boldsymbol{Y}|\theta)|\mathcal{D}, \theta^{(t)}]$
7. The EM algorithm iterates the E-step with the M-step.
8. Prove that maximizing $\mathbb{E}[\log p(\mathcal{D}, \boldsymbol{Y}|\theta)|\mathcal{D}, \theta^{(t)}]$ maximizes $p(\mathcal{D}|\theta)$

Why EM Works

$$\log p(\mathcal{D}|\theta^{(t+1)}) - \log p(\mathcal{D}|\theta^{(t)}) = \log \frac{p(\mathcal{D}|\theta^{(t+1)})}{p(\mathcal{D}|\theta^{(t)})}$$

Why EM Works

$$\log p(\mathcal{D}|\theta^{(t+1)}) - \log p(\mathcal{D}|\theta^{(t)}) = \log \frac{p(\mathcal{D}|\theta^{(t+1)})}{p(\mathcal{D}|\theta^{(t)})}$$

Marginalize

$$= \log \int_{\boldsymbol{y}} \frac{p(\mathcal{D}, \boldsymbol{y}|\theta^{(t+1)})}{p(\mathcal{D}|\theta^{(t)})} d\boldsymbol{y}$$

Why EM Works

$$\log p(\mathcal{D}|\theta^{(t+1)}) - \log p(\mathcal{D}|\theta^{(t)}) = \log \frac{p(\mathcal{D}|\theta^{(t+1)})}{p(\mathcal{D}|\theta^{(t)})}$$

$$= \log \int_{\boldsymbol{y}} \frac{p(\mathcal{D}, \boldsymbol{y}|\theta^{(t+1)})}{p(\mathcal{D}|\theta^{(t)})} d\boldsymbol{y}$$

Apply product rule

$$= \log \int_{\boldsymbol{y}} \frac{p(\mathcal{D}, \boldsymbol{y}|\theta^{(t+1)})}{p(\mathcal{D}, \boldsymbol{y}|\theta^{(t)})} p(\boldsymbol{y}|\mathcal{D}, \theta^{(t)}) d\boldsymbol{y}$$

Why EM Works

$$
\begin{aligned}
\log p(\mathcal{D}|\theta^{(t+1)}) - \log p(\mathcal{D}|\theta^{(t)}) &= \log \frac{p(\mathcal{D}|\theta^{(t+1)})}{p(\mathcal{D}|\theta^{(t)})} \\
&= \log \int_{\boldsymbol{y}} \frac{p(\mathcal{D}, \boldsymbol{y}|\theta^{(t+1)})}{p(\mathcal{D}|\theta^{(t)})} d\boldsymbol{y} \\
&= \log \int_{\boldsymbol{y}} \frac{p(\mathcal{D}, \boldsymbol{y}|\theta^{(t+1)})}{p(\mathcal{D}, \boldsymbol{y}|\theta^{(t)})} p(\boldsymbol{y}|\mathcal{D}, \theta^{(t)}) d\boldsymbol{y} \\
\text{Apply Jensen's inequality} \qquad &\geq \int_{\boldsymbol{y}} \log \frac{p(\mathcal{D}, \boldsymbol{y}|\theta^{(t+1)})}{p(\mathcal{D}, \boldsymbol{y}|\theta^{(t)})} p(\boldsymbol{y}|\mathcal{D}, \theta^{(t)}) d\boldsymbol{y}
\end{aligned}
$$

Jensen's inequality: $\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]$, where φ is a convex function. $\log(\cdot)$ above is a concave function.

Why EM Works

$$
\begin{aligned}
\log p(\mathcal{D}|\theta^{(t+1)}) - \log p(\mathcal{D}|\theta^{(t)}) &= \log \frac{p(\mathcal{D}|\theta^{(t+1)})}{p(\mathcal{D}|\theta^{(t)})} \\
&= \log \int_{\boldsymbol{y}} \frac{p(\mathcal{D}, \boldsymbol{y}|\theta^{(t+1)})}{p(\mathcal{D}|\theta^{(t)})} d\boldsymbol{y} \\
&= \log \int_{\boldsymbol{y}} \frac{p(\mathcal{D}, \boldsymbol{y}|\theta^{(t+1)})}{p(\mathcal{D}, \boldsymbol{y}|\theta^{(t)})} p(\boldsymbol{y}|\mathcal{D}, \theta^{(t)}) d\boldsymbol{y} \\
&\geq \int_{\boldsymbol{y}} \log \frac{p(\mathcal{D}, \boldsymbol{y}|\theta^{(t+1)})}{p(\mathcal{D}, \boldsymbol{y}|\theta^{(t)})} p(\boldsymbol{y}|\mathcal{D}, \theta^{(t)}) d\boldsymbol{y} \\
\text{Rewrite} \qquad &= \mathbb{E}[\log p(\mathcal{D}, \boldsymbol{Y}|\theta^{(t+1)})|\mathcal{D}, \theta^{(t)}] - \mathbb{E}[\log p(\mathcal{D}, \boldsymbol{Y}|\theta^{(t)})|\mathcal{D}, \theta^{(t)}]
\end{aligned}
$$

So,

$$
\log p(\mathcal{D}|\theta^{(t+1)}) \geq \log p(\mathcal{D}|\theta^{(t)}) + \mathbb{E}[\log p(\mathcal{D}, \boldsymbol{Y}|\theta^{(t+1)})|\mathcal{D}, \theta^{(t)}] - \mathbb{E}[\log p(\mathcal{D}, \boldsymbol{Y}|\theta^{(t)})|\mathcal{D}, \theta^{(t)}]
$$

Why EM Works

$$
\begin{aligned}
\log p(\mathcal{D}|\theta^{(t+1)}) - \log p(\mathcal{D}|\theta^{(t)}) &= \log \frac{p(\mathcal{D}|\theta^{(t+1)})}{p(\mathcal{D}|\theta^{(t)})} \\
&= \log \int_{\boldsymbol{y}} \frac{p(\mathcal{D}, \boldsymbol{y}|\theta^{(t+1)})}{p(\mathcal{D}|\theta^{(t)})} d\boldsymbol{y} \\
&= \log \int_{\boldsymbol{y}} \frac{p(\mathcal{D}, \boldsymbol{y}|\theta^{(t+1)})}{p(\mathcal{D}, \boldsymbol{y}|\theta^{(t)})} p(\boldsymbol{y}|\mathcal{D}, \theta^{(t)}) d\boldsymbol{y} \\
&\geq \int_{\boldsymbol{y}} \log \frac{p(\mathcal{D}, \boldsymbol{y}|\theta^{(t+1)})}{p(\mathcal{D}, \boldsymbol{y}|\theta^{(t)})} p(\boldsymbol{y}|\mathcal{D}, \theta^{(t)}) d\boldsymbol{y} \\
\text{Rewrite} \qquad &= \mathbb{E}[\log p(\mathcal{D}, \boldsymbol{Y}|\theta^{(t+1)})|\mathcal{D}, \theta^{(t)}] - \mathbb{E}[\log p(\mathcal{D}, \boldsymbol{Y}|\theta^{(t)})|\mathcal{D}, \theta^{(t)}]
\end{aligned}
$$

So,

$$
\log p(\mathcal{D}|\theta^{(t+1)}) \geq \log p(\mathcal{D}|\theta^{(t)}) + \overbrace{\mathbb{E}[\log p(\mathcal{D}, \boldsymbol{Y}|\theta^{(t+1)})|\mathcal{D}, \theta^{(t)}] - \mathbb{E}[\log p(\mathcal{D}, \boldsymbol{Y}|\theta^{(t)})|\mathcal{D}, \theta^{(t)}]}^{\text{Non-negative}}
$$

BISHOP'S OPENING

First paragraph of Chapter 9, Bishop 2006

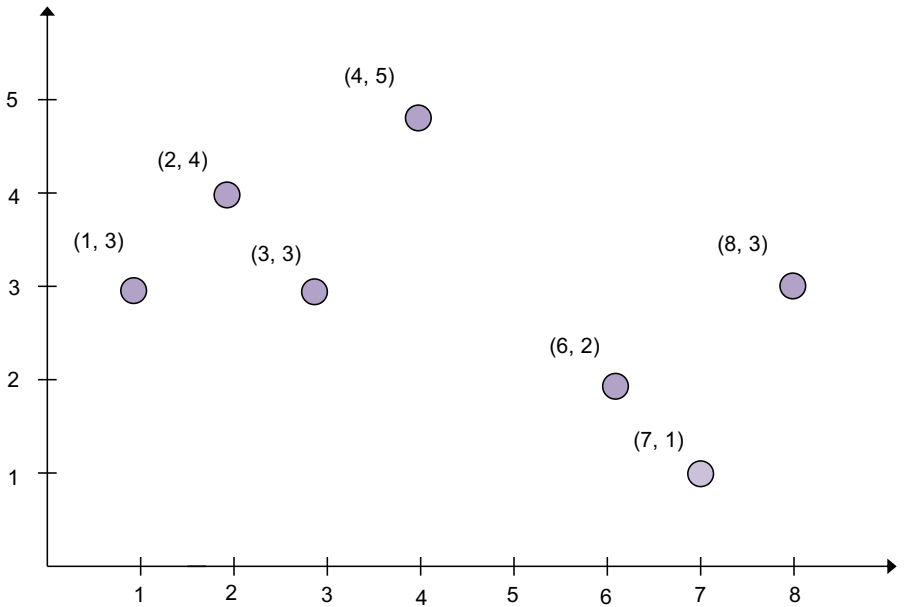
If we define a joint distribution over observed and latent variables, the corresponding distribution of the observed variables alone is obtained by marginalization. This allows relatively complex marginal distributions over observed variables to be expressed in terms of more tractable joint distributions over the expanded space of observed and latent variables. The introduction of latent variables thereby allows complicated distributions to be formed from simpler components. In this chapter, we shall see that mixture distributions, such as the Gaussian mixture discussed in Section 2.3.9, can be interpreted in terms of discrete latent variables. Continuous latent variables will form the subject of Chapter 12.

K-Means Clustering

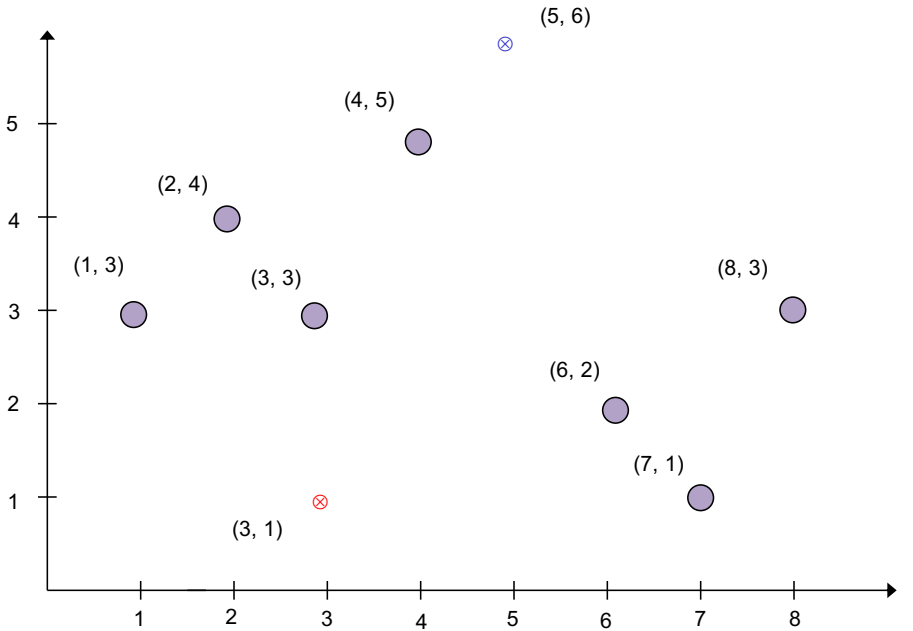
1. Partitional approach to clustering
2. Each cluster is associated with a centroid
3. Each point is assigned to the cluster with closest centroid
4. The number of clusters, K, must be specified

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

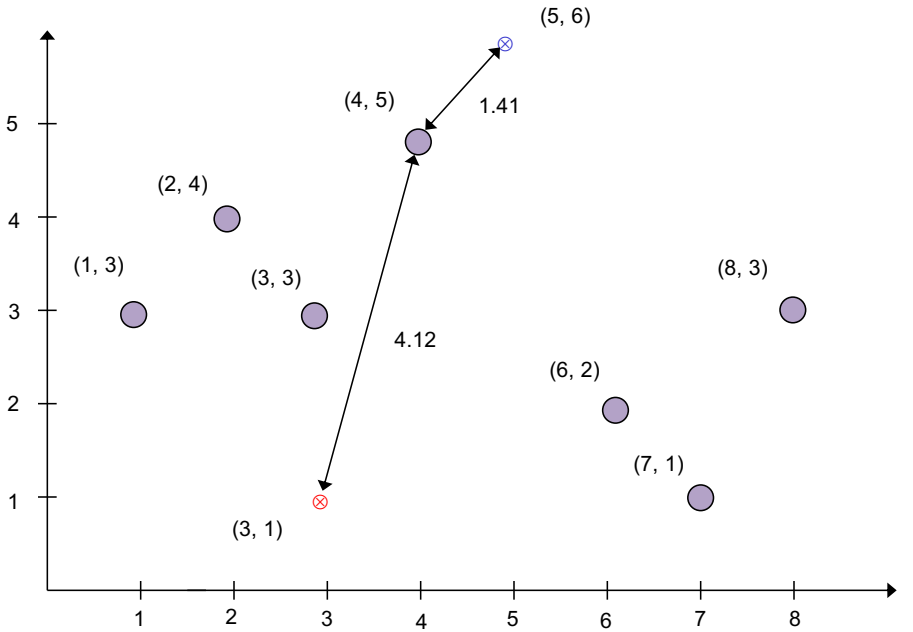
EXAMPLE: K-MEANS CLUSTERING



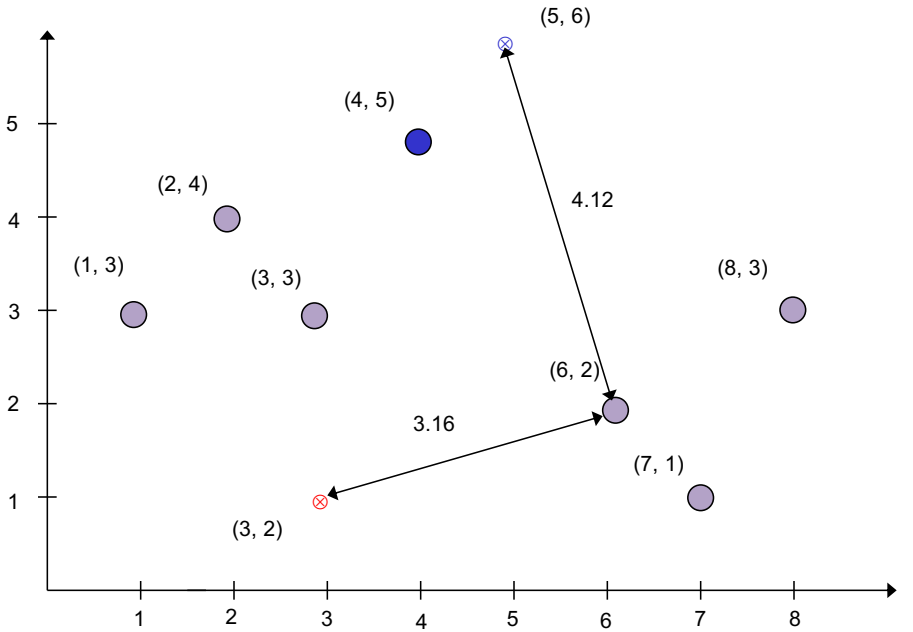
EXAMPLE: K-MEANS CLUSTERING



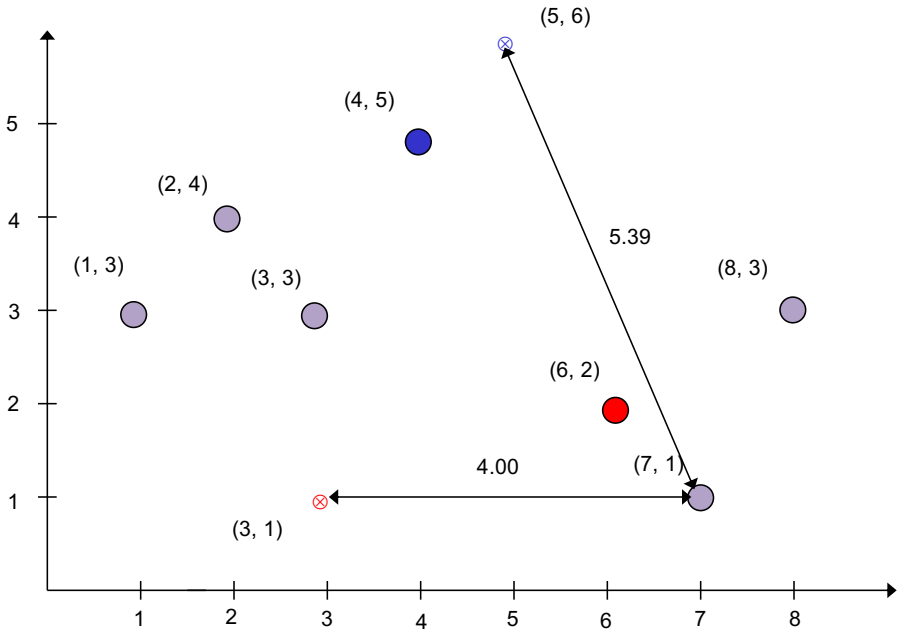
EXAMPLE: K-MEANS CLUSTERING



EXAMPLE: K-MEANS CLUSTERING



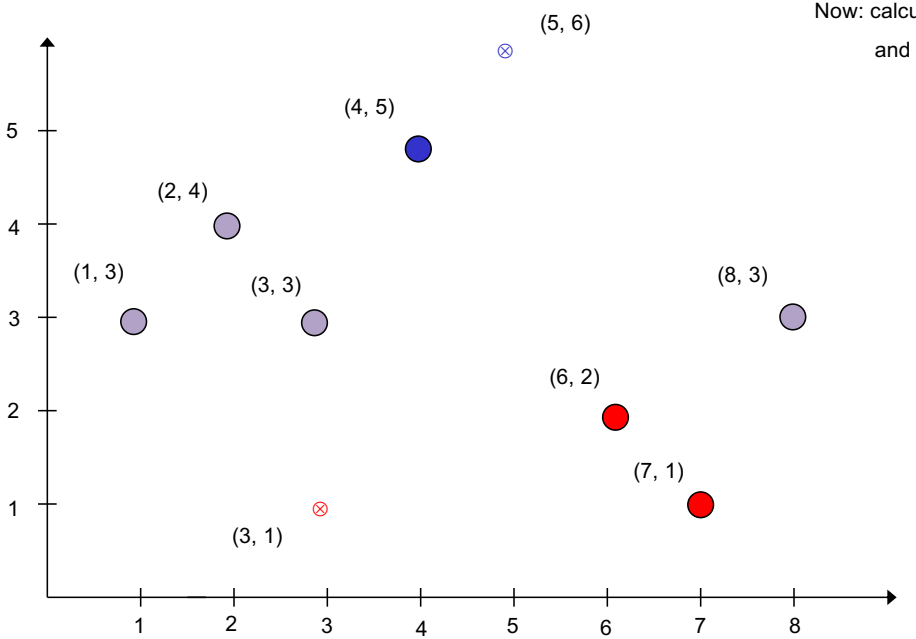
EXAMPLE: K-MEANS CLUSTERING



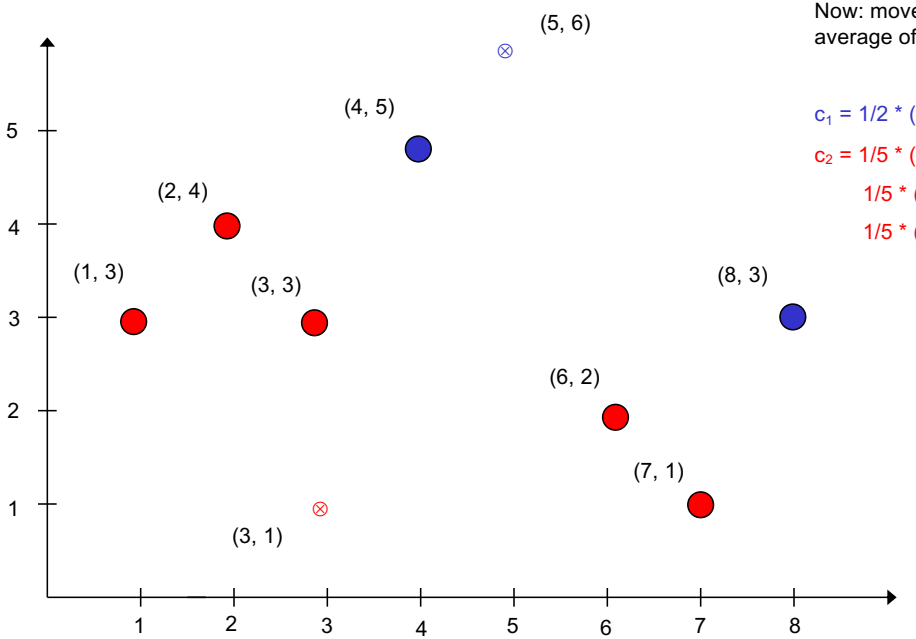
EXAMPLE: K-MEANS CLUSTERING

Now: calculate all distances...

and color all data points.



EXAMPLE: K-MEANS CLUSTERING



Now: move cluster centers to be the average of data points.

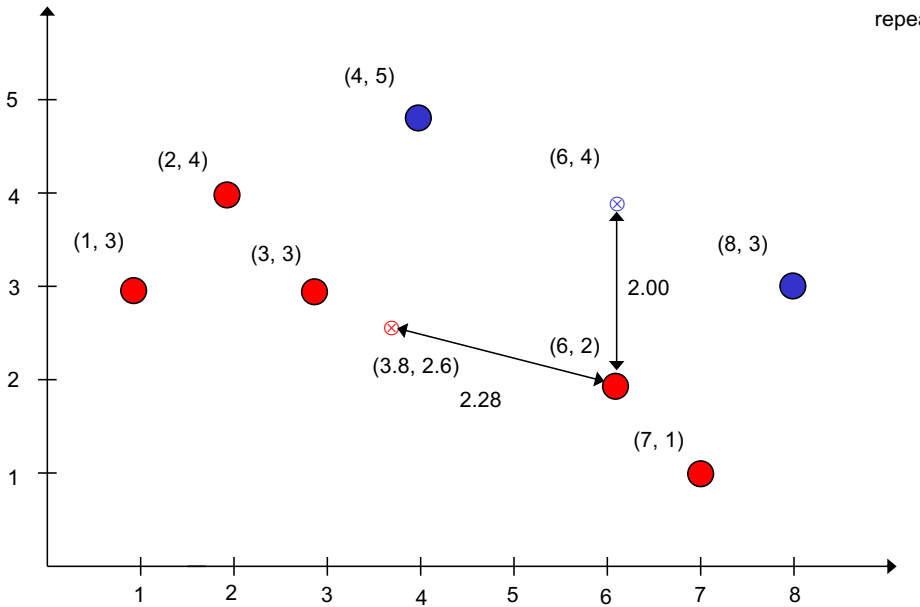
c_1 = 1/2 * (4 5) + 1/2 * (8 3) = (6 4)

c_2 = 1/5 * (1 3) + 1/5 * (2 4) +
1/5 * (3 3) + 1/5 * (6 2) +
1/5 * (7 1) = (3.8 2.6)

EXAMPLE: K-MEANS CLUSTERING

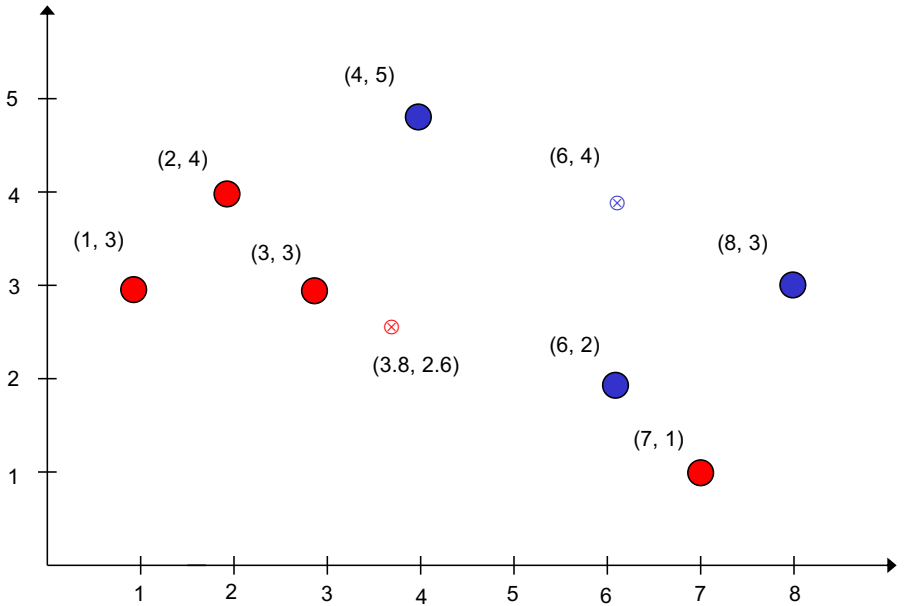
Now: start next iteration

repeat distance calculation.

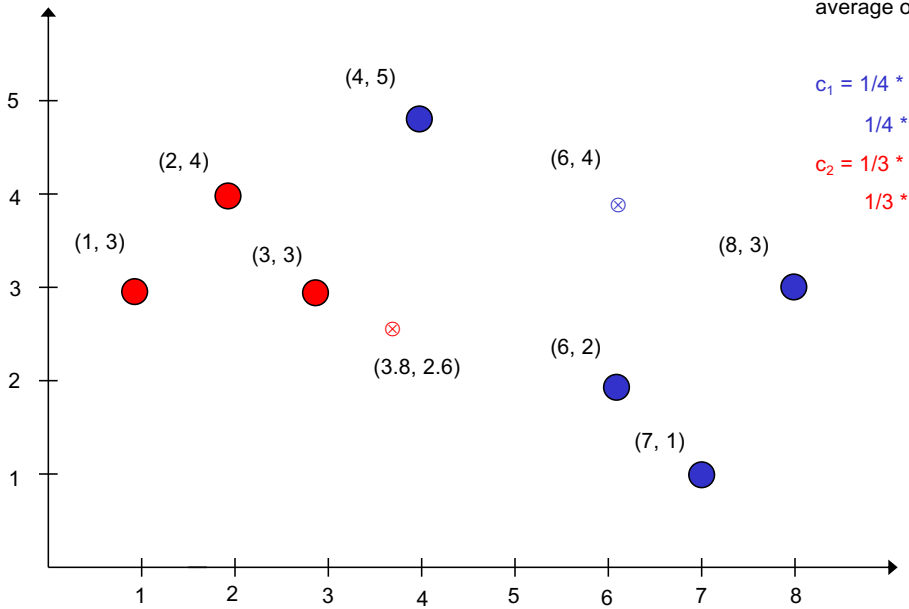


EXAMPLE: K-MEANS CLUSTERING

Now: calculate all other distances...



EXAMPLE: K-MEANS CLUSTERING



Now: move cluster centers to be the average of data points.

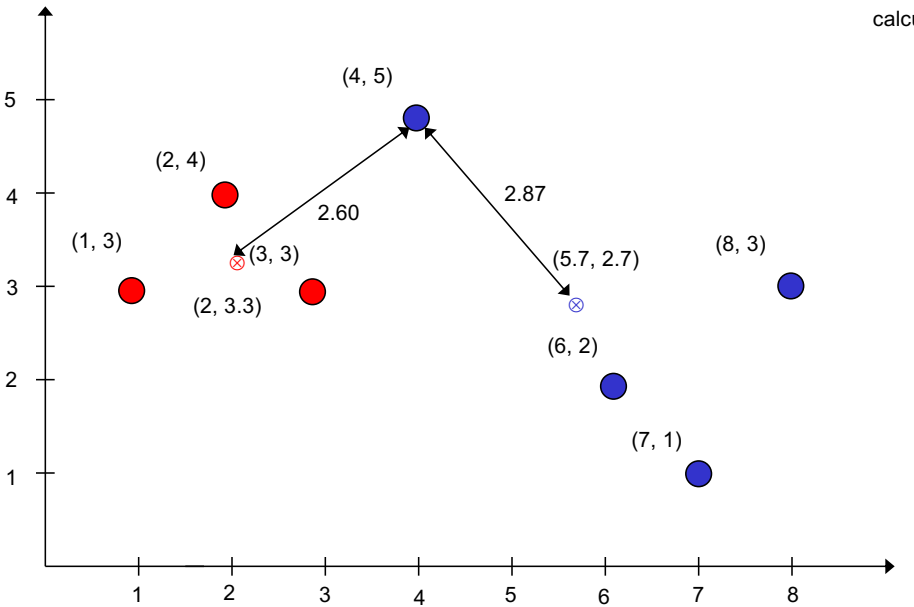
c_1 = 1/4 * (4 5) + 1/4 * (8 3) +
1/4 * (6 2) + 1/4 * (7 1) = (5.67 2.67)

c_2 = 1/3 * (1 3) + 1/3 * (2 4) +
1/3 * (3 3) = (2.00 3.33)

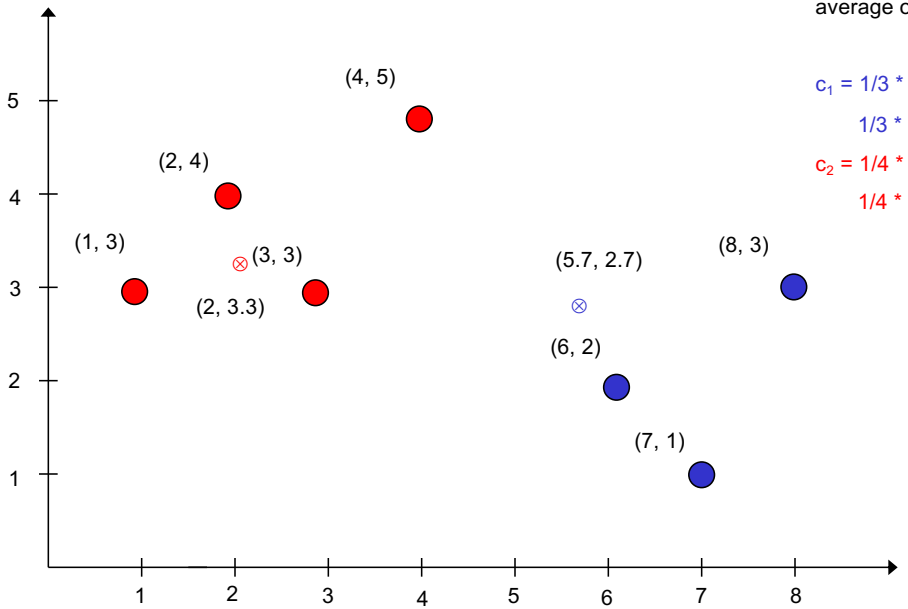
EXAMPLE: K-MEANS CLUSTERING

Now: start next iteration

calculate distances again...



EXAMPLE: K-MEANS CLUSTERING

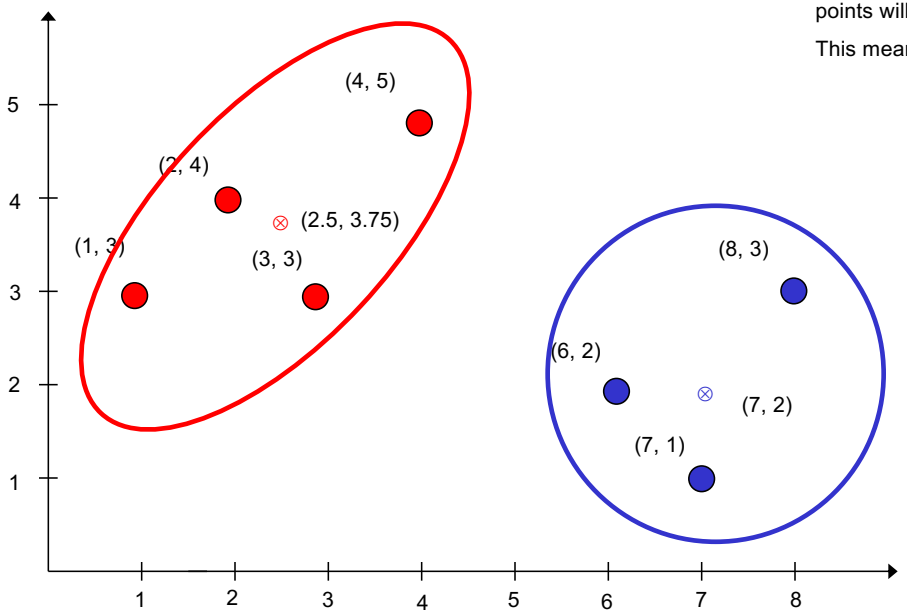


Now: move cluster centers to be the average of data points.

c_1 = 1/3 * (6 2) + 1/3 * (8 3) +
1/3 * (7 1) = (7 2)

c_2 = 1/4 * (1 3) + 1/4 * (2 4) +
1/4 * (3 3) + 1/4 * (4 5) = (2.5 3.75)

EXAMPLE: K-MEANS CLUSTERING



Now: if we calculate all distances, no data points will change color.

This means, we can stop!