LECTURE PLAN

- Motivation: Observations with missing values
- Sketch of the algorithm, relation to K-means
- EM algorithm
Motivation. Example (1)

We measure lengths of vehicles. The observation space is two-dimensional, with \( x \in \{\text{car, truck}\} \) capturing vehicle type and \( y \in \mathbb{R} \) capturing length.

\[
p(x, y) : \text{distribution}, \quad x \in \{\text{car, truck}\}, \quad y \in \mathbb{R}
\]  

(1)

\[
p(\text{car}, y) = \pi_c \mathcal{N}(y|\mu_c, \sigma_c = 1) = \kappa_c \exp \left\{ -\frac{1}{2} (y - \mu_c)^2 \right\}, \quad (\kappa_c = \frac{\pi_c}{\sqrt{2\pi}})
\]  

(2)

\[
p(\text{truck}, y) = \pi_t \mathcal{N}(y|\mu_t, \sigma_t = 2) = \kappa_t \exp \left\{ -\frac{1}{8} (y - \mu_t)^2 \right\}, \quad (\kappa_t = \frac{\pi_c}{\sqrt{8\pi}})
\]  

(3)

Parameters \( \pi_c, \pi_t, \sigma_c, \sigma_t \) are assumed to be known. The only unknowns are \( \mu_c \) and \( \mu_t \). We want to recover \( \mu_c \) and \( \mu_t \) using Maximum Likelihood.

Example (\( \pi_c = 0.6, \pi_t = 0.4, \sigma_c = 1, \sigma_t = 2, \mu_c = 5, \mu_t = 10 \) )
Motivation. Example (2)

The observations are:

\[ \mathcal{T} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\} \]

\[ = \{(\text{car}, y_1^{(c)}), (\text{car}, y_2^{(c)}), \ldots, (\text{car}, y_C^{(c)})\}, (\text{truck}, y_1^{(t)}), (\text{truck}, y_2^{(t)}), \ldots, (\text{truck}, y_T^{(t)})\} \]

\( C \) car observations \hspace{1cm} \( T \) truck observations

Log-likelihood \( \ell(\mathcal{T}) = \ln p(\mathcal{T} | \mu_c, \mu_t) \):

\[
\ell(\mathcal{T}) = \sum_{i=1}^{N} \ln p(x_i, y_i | \mu_c, \mu_t) = C \ln \kappa_c - \frac{1}{2} \sum_{i=1}^{C} (y_i^{(c)} - \mu_c)^2 + T \ln \kappa_t - \frac{1}{8} \sum_{i=1}^{T} (y_i^{(t)} - \mu_t)^2
\]

Estimation of \( \mu_1, \mu_2 \) is very easy:

\[
\frac{\partial \ell(\mathcal{T})}{\partial \mu_c} = \sum_{i=1}^{C} (y_i^{(c)} - \mu_c) = 0 \Rightarrow \mu_c = \frac{1}{C} \sum_{i=1}^{C} y_i^{(c)}
\]

\[
\frac{\partial \ell(\mathcal{T})}{\partial \mu_t} = \frac{1}{4} \sum_{i=1}^{T} (y_i^{(t)} - \mu_t) = 0 \Rightarrow \mu_t = \frac{1}{T} \sum_{i=1}^{T} y_i^{(t)}
\]
Motivation. Missing Values (3)

Consider some observations to have the first coordinate missing ($\bullet$):

$$\mathcal{T} = \{(\text{car}, y_{1}^{(c)}), \ldots, (\text{car}, y_{C}^{(c)}), (\text{truck}, y_{1}^{(t)}), \ldots, (\text{truck}, y_{T}^{(t)}), (\bullet, y_{1}), \ldots, (\bullet, y_{M})\}$$  \hspace{1cm} (9)

data with unknown vehicle type

What is the probability of observing $y^\bullet$?

$$p(y^\bullet) = p(\text{car}, y^\bullet) + p(\text{truck}, y^\bullet)$$  \hspace{1cm} (marginalizing over unknown value)

Log-likelihood:

$$\ell(\mathcal{T}) = \sum_{i=1}^{N} \ln p(x_{i}, y_{i} | \mu_{c}, \mu_{t}) = C \ln \kappa_{c} - \frac{1}{2} \sum_{i=1}^{C} (y_{i}^{(c)} - \mu_{c})^{2} + T \ln \kappa_{t} - \frac{1}{8} \sum_{i=1}^{T} (y_{i}^{(t)} - \mu_{t})^{2}$$

\hspace{1cm} (same term as before)

$$+ \sum_{i=1}^{M} \ln \left( \kappa_{c} \exp \left\{ -\frac{1}{2} (y_{i} - \mu_{c})^{2} \right\} + \kappa_{t} \exp \left\{ -\frac{1}{8} (y_{i} - \mu_{t})^{2} \right\} \right)$$  \hspace{1cm} (11)
Motivation. Missing Values (4)

Log-likelihood:

\[
\ell(T) = C \ln \kappa_c - \frac{1}{2} \sum_{i=1}^{C} (y_i^{(c)} - \mu_c)^2 + T \ln \kappa_t - \frac{1}{8} \sum_{i=1}^{T} (y_i^{(t)} - \mu_t)^2
\]

\[
+ \sum_{i=1}^{M} \ln \left( \kappa_c \exp \left\{ -\frac{1}{2} (y_i^{\cdot} - \mu_c)^2 \right\} + \kappa_t \exp \left\{ -\frac{1}{8} (y_i^{\cdot} - \mu_t)^2 \right\} \right)
\]

Optimality condition (shown for \( \mu_c \) only):

\[
0 = \frac{\partial \ell(T)}{\partial \mu_c} = \sum_{i=1}^{C} (y_i^{(c)} - \mu_c) +
\]

\[
+ \sum_{i=1}^{M} \frac{\kappa_c \exp \left\{ -\frac{1}{2} (y_i^{\cdot} - \mu_c)^2 \right\}}{\kappa_c \exp \left\{ -\frac{1}{2} (y_i^{\cdot} - \mu_c)^2 \right\} + \kappa_t \exp \left\{ -\frac{1}{8} (y_i^{\cdot} - \mu_t)^2 \right\}} (y_i^{\cdot} - \mu_c)
\]
Motivation. Missing Values (5)

Log-likelihood:

\[
\ell(T) = C \ln \kappa_c - \frac{1}{2} \sum_{i=1}^{C} (y_{i}^{(c)} - \mu_c)^2 + T \ln \kappa_t - \frac{1}{8} \sum_{i=1}^{T} (y_{i}^{(t)} - \mu_t)^2 \\
+ \sum_{i=1}^{M} \ln \left( \kappa_c \exp \left\{ -\frac{1}{2} (y_{i}^{\cdot} - \mu_c)^2 \right\} + \kappa_t \exp \left\{ -\frac{1}{8} (y_{i}^{\cdot} - \mu_t)^2 \right\} \right)
\]

Optimality condition (shown for \( \mu_c \) only):

\[
0 = \frac{\partial \ell(T)}{\partial \mu_c} = \sum_{i=1}^{C} (y_{i}^{(c)} - \mu_c) + \sum_{i=1}^{M} \frac{\kappa_c \exp \left\{ -\frac{1}{2} (y_{i}^{\cdot} - \mu_c)^2 \right\}}{p(\text{car}, y_{i}^{\cdot} | \mu_c, \mu_t)} (y_{i}^{\cdot} - \mu_c) \\
+ \sum_{i=1}^{M} \frac{\kappa_c \exp \left\{ -\frac{1}{2} (y_{i}^{\cdot} - \mu_c)^2 \right\}}{p(\text{car}, y_{i}^{\cdot} | \mu_c, \mu_t)} + \frac{\kappa_t \exp \left\{ -\frac{1}{8} (y_{i}^{\cdot} - \mu_t)^2 \right\}}{p(\text{truck}, y_{i}^{\cdot} | \mu_c, \mu_t)} (y_{i}^{\cdot} - \mu_t)
\]
Motivation. Missing Values (6)

Log-likelihood:

\[
\ell(T) = C \ln \kappa_c - \frac{1}{2} \sum_{i=1}^{C} (y_i^{(c)} - \mu_c)^2 + T \ln \kappa_t - \frac{1}{8} \sum_{i=1}^{T} (y_i^{(t)} - \mu_t)^2 
\]

\[
+ \sum_{i=1}^{M} \ln \left( \kappa_c \exp \left\{ -\frac{1}{2} (y_i^{\bullet} - \mu_c)^2 \right\} + \kappa_t \exp \left\{ -\frac{1}{8} (y_i^{\bullet} - \mu_t)^2 \right\} \right) 
\]

Optimality condition (shown for \( \mu_c \) only):

\[
0 = \frac{\partial \ell(T)}{\partial \mu_c} = \sum_{i=1}^{C} (y_i^{(c)} - \mu_c) + \frac{\kappa_c \exp \left\{ -\frac{1}{2} (y_i^{\bullet} - \mu_c)^2 \right\}}{\kappa_c \exp \left\{ -\frac{1}{2} (y_i^{\bullet} - \mu_c)^2 \right\} + \kappa_t \exp \left\{ -\frac{1}{8} (y_i^{\bullet} - \mu_t)^2 \right\}} (y_i^{\bullet} - \mu_c) 
\]
Motivation. Missing Values (7)

Optimality conditions (shown for both $\mu_c$ and $\mu_t$):

\[
0 = \frac{\partial \ell(T)}{\partial \mu_c} = \sum_{i=1}^{C} (y_i^{(c)} - \mu_c) + p(\text{car}|y_i^\bullet, \mu_c, \mu_t) \left( \frac{\kappa_c \exp \left\{ -\frac{1}{2} (y_i^\bullet - \mu_c)^2 \right\}}{\kappa_c \exp \left\{ -\frac{1}{2} (y_i^\bullet - \mu_c)^2 \right\} + \kappa_t \exp \left\{ -\frac{1}{8} (y_i^\bullet - \mu_t)^2 \right\}} \right) (y_i^\bullet - \mu_c) \quad (24)
\]

\[
0 = 4\frac{\partial \ell(T)}{\partial \mu_t} = \sum_{i=1}^{T} (y_i^{(t)} - \mu_t) + \sum_{i=1}^{M} p(\text{truck}|y_i^\bullet, \mu_c, \mu_t) (y_i^\bullet - \mu_t) \quad (25)
\]

Note:

- Complicated equations for the unknowns $\mu_c$, $\mu_t$
- Both equations contain $\mu_c$ and $\mu_t$ (cf. case with no missing variables)
Motivation. Missing Values (8)

Optimality conditions (shown for both $\mu_c$ and $\mu_t$):

$$C \sum_{i=1}^{C} (y_i^{(c)} - \mu_c) + \sum_{i=1}^{M} p(\text{car}|y_i^\bullet, \mu_c, \mu_t) (y_i^\bullet - \mu_c) = 0 \quad (27)$$

$$T \sum_{i=1}^{T} (y_i^{(t)} - \mu_t) + \sum_{i=1}^{M} p(\text{truck}|y_i^\bullet, \mu_c, \mu_t) (y_i^\bullet - \mu_t) = 0 \quad (28)$$

If $p(\text{car}|y_i^\bullet, \mu_c, \mu_t)$ and $p(\text{truck}|y_i^\bullet, \mu_c, \mu_t)$ were known, the estimation would’ve been easy:

- Let $z_i \ (i = 1, 2, ..., M)$, $z_i \in \{\text{car, truck}\}$ denote the missing values. Define $q(z_i) = p(z_i|y_i^\bullet, \mu_c, \mu_t)$
- The equations lead to

$$C \sum_{i=1}^{C} (y_i^{(c)} - \mu_c) + \sum_{i=1}^{M} q(z_i = \text{car}) (y_i^\bullet - \mu_c) = 0 \quad (29)$$

$$\Rightarrow \mu_c = \frac{\sum_{i=1}^{C} y_i^{(c)} + \sum_{i=1}^{M} q(z_i = \text{car}) y_i^\bullet}{C + \sum_{i=1}^{M} q(z_i = \text{car})} \quad (30)$$

and similarly,

$$\mu_t = \frac{\sum_{i=1}^{T} y_i^{(t)} + \sum_{i=1}^{M} q(z_i = \text{truck}) y_i^\bullet}{T + \sum_{i=1}^{M} q(z_i = \text{truck})} \quad (31)$$
Motivation. Missing Values (9)

\[
\mu_c = \frac{\sum_{i=1}^{C} y_i^{(c)} + \sum_{i=1}^{M} q(z_i = \text{car}) y_i^*}{C + \sum_{i=1}^{M} q(z_i = \text{car})}
\]

\[
\mu_t = \frac{\sum_{i=1}^{T} y_i^{(t)} + \sum_{i=1}^{M} q(z_i = \text{truck}) y_i^*}{T + \sum_{i=1}^{M} q(z_i = \text{truck})}
\]

These expressions are weighted averages of the observed \(y\)'s. Data with non-missing \(x\) have weight 1, the data with missing \(x\) have weight \(q(z_i)\). How about trying the following procedure for finding the ML estimate of \(\mu_c\) and \(\mu_t\):

1. Initialize \(\mu_c, \mu_t\)
2. Compute \(q(z_i) = p(z_i \mid y^*_i, \mu_c, \mu_t)\) for all \(i = 1, 2, ..., M\)
3. Recompute \(\mu_c, \mu_t\) according to Eqs. (32, 33)
4. If termination condition is met, finish. Otherwise goto 2.

This is the essence of the **EM algorithm**, with Step 2 called the **Expectation** (E) step and Step 3 called the **Maximization** (M) step.
An extreme of the previous example is that no data have the $x$-coordinate value (car/truck vehicle type). Everything works just as well:

$$
\mu_c = \frac{\sum_{i=1}^{M} q(z_i = \text{car}) y_i \cdot}{\sum_{i=1}^{M} q(z_i = \text{car})}
$$

(34)

$$
\mu_t = \frac{\sum_{i=1}^{M} q(z_i = \text{truck}) y_i \cdot}{\sum_{i=1}^{M} q(z_i = \text{truck})}
$$

(35)

1. Initialize $\mu_c$, $\mu_t$

2. Compute $q(z_i) = p(z_i|y_i \cdot, \mu_c, \mu_t)$ for all $i = 1, 2, ..., M$

3. Recompute $\mu_c$, $\mu_t$ according to Eqs.(36, 37)

4. If termination condition is met, finish. Otherwise goto 2.

Note: Can you imagine this algorithm to end up at a local maximum?
Clustering, Soft Assignment, Relation to K-means (2)

An extreme of the previous example is that no data have the $x$-coordinate (car/truck).

$$\mu_c = \frac{\sum_{i=1}^{M} q(z_i = \text{car}) y_i}{\sum_{i=1}^{M} q(z_i = \text{car})}$$  \hspace{1cm} (36)

$$\mu_t = \frac{\sum_{i=1}^{M} q(z_i = \text{truck}) y_i}{\sum_{i=1}^{M} q(z_i = \text{truck})}$$  \hspace{1cm} (37)

**EM** algorithm:

1. Initialize $\mu_c$, $\mu_t$
2. Compute $q(z_i) = p(z_i|y^*_i, \mu_c, \mu_t)$ for all $i = 1, 2, ..., M$
3. Recompute $\mu_c$, $\mu_t$ according to Eqs.\(36, 37\)
4. If termination condition is met, finish. Otherwise goto 2.

**K-means:**

1. ditto
2. $q(z_i = \text{car}) = [|y^*_i - \mu_c| < |y^*_i - \mu_t|]$  
   $q(z_i = \text{truck}) = [|y^*_i - \mu_t| \leq |y^*_i - \mu_c|]$ for all $i = 1, 2, ..., M$
3. ditto
4. ditto

EM-based clustering uses soft assignment. K-means can be interpreted as an EM-based clustering with hard assignment.
**EM algorithm**

- $\mathcal{T}$: training set
- $\mathbf{o}$: all observed values (no essential difference between $\mathcal{T}$ and $\mathbf{o}$, just notational convenience)
- $\mathbf{z}$: all unobserved values
- $\theta$: model parameters to be estimated.

**Goal:** Find $\theta^*$ using the Maximum Likelihood approach:

$$\theta^* = \arg\max_{\theta} \ell(\theta) = \arg\max_{\theta} \ln p(\mathbf{o}|\theta) \tag{38}$$

**Line of thought**

Assume that solving this:

$$\arg\max_{\theta} \ln p(\mathbf{o}, \mathbf{z}|\theta) \tag{39}$$

is easy (optimal parameters had $\mathbf{z}$ been known.)

Our goal will be to rewrite Eq. (38) in a way which will involve optimization terms of kind as in Eq. (39).
Lower Bound on the Log Likelihood

\[ \ln p(o|\theta) = \ln \sum_z p(o, z|\theta) \]

Marginalizing over missing values \( (40) \)

\[ = \ln \sum_z q(z) \frac{p(o, z|\theta)}{q(z)} \]

Introduction of distribution \( q(z) \) \( (41) \)

As \( \forall z: 0 \leq q(z) \leq 1 \) and \( \sum_z q(z) = 1 \), the sum is now a convex combination of \( \frac{p(o, z|\theta)}{q(z)} \).

\[ \geq \sum_z q(z) \ln \frac{p(o, z|\theta)}{q(z)} \]

Jensen's inequality. Here inequality holds because logarithm is a concave function. \( (42) \)

Define

\[ \mathcal{L}(q, \theta) = \sum_z q(z) \ln \frac{p(o, z|\theta)}{q(z)}. \] \( (43) \)

This \( \mathcal{L}(q, \theta) \) is the lower bound for \( \ln p(o|\theta) \) due to Eq. \( (42) \), for any distribution \( q \).

Maximizing \( \mathcal{L}(q, \theta) \) will also push the log likelihood upwards.
How Tight Is This Bound? (1)

\[ \ln p(o|\theta) - \mathcal{L}(q, \theta) = \ln p(o|\theta) - \sum_z q(z) \ln \frac{p(o, z|\theta)}{q(z)} \] (44)

\[ = \ln p(o|\theta) - \sum_z q(z) \{ \ln p(o, z|\theta) - \ln q(z) \} \] (45)

\[ = \ln p(o|\theta) - \sum_z q(z) \{ \ln p(z|o, \theta) + \ln p(o|\theta) - \ln q(z) \} \] (46)

\[ = \ln p(o|\theta) - \sum_z q(z) \ln p(o|\theta) - \sum_z q(z) \{ \ln p(z|o, \theta) - \ln q(z) \} \]

\[ + \sum_z q(z) \ln \frac{p(z|o, \theta)}{q(z)} \] (47)

\[ = - \sum_z q(z) \ln \frac{p(z|o, \theta)}{q(z)} \] (48)

This is the Kullback Leibler divergence between the two distributions \( q(z) \) and \( p(z|o, \theta) \):

\[ D_{KL}(q||p) = \sum_z q(z) \ln \frac{q(z)}{p(z|o, \theta)} = - \sum_z q(z) \ln \frac{p(z|o, \theta)}{q(z)} \] (49)
How Tight Is This Bound? (2)

\[
\ln p(o|\theta) = \mathcal{L}(q, \theta) + D_{KL}(q||p) \tag{50}
\]

\[\uparrow \quad \uparrow \quad \uparrow \]
log likelihood  lower bound  gap

We already know that due to Jensen’s inequality, \(\mathcal{L}(q, \theta)\) is indeed the lower bound. This is confirmed by the fact that \(D_{KL}(q||p) \geq 0\) for any \(q, p\). Additionally,

\[
D_{KL}(q||p) = 0 \quad \Leftrightarrow \quad p = q. \tag{51}
\]

When \(q = p\), the bound is tight.
EM algorithm

\[ \ln p(o|\theta) = \mathcal{L}(q, \theta) + D_{KL}(q||p) \]  
\[ \uparrow \quad \uparrow \quad \uparrow \]
log likelihood lower bound gap

EM algorithm attempts to maximize the log-likelihood by instead maximizing the lower bound (why 'attempts'? Because it may end up in local maximum).

1. Initialize \( \theta = \theta^{(0)} \) \( (t = 0) \)

2. **E-step** (Expectation):
\[ q^{(t+1)} = \arg\max_q \mathcal{L}(q, \theta^{(t)}) \]  
(53)

3. **M-step** (Maximization):
\[ \theta^{(t+1)} = \arg\max_{\theta} \mathcal{L}(q^{(t+1)}, \theta) \]  
(54)

4. If termination condition is not met, goto 2.
Expectation step

E-step: $\theta^{(t)}$ is fixed

$$q^{(t+1)} = \underset{q}{\operatorname{argmax}} \mathcal{L}(q, \theta^{(t)})$$  \hspace{1cm} (55)

$$\mathcal{L}(q, \theta^{(t)}) = \ln p(o | \theta^{(t)}) - D_{KL}(q || p)$$ \hspace{1cm} (56)

**Note:** The distribution $q$ maximizing this term is the one which minimizes the KL divergence. KL divergence is minimized when the two distributions are the same. Thus, the distribution maximizing Eq. (55) is

$$q^{(t+1)}(z) = p(z | o, \theta^{(t)})$$ \hspace{1cm} (57)

**Recall:**

$$D_{KL}(q || p) = - \sum_z q(z) \ln \frac{p(z | o, \theta)}{q(z)}$$ \hspace{1cm} (58)
Maximization step

M-step: \( q^{(t+1)} \) is fixed

\[
\theta^{(t+1)} = \underset{\theta}{\text{argmax}} \ L(q^{(t+1)}, \theta) 
\]  

(59)

\[
L(q^{(t+1)}, \theta) = \sum_z q^{(t+1)}(z) \ln \frac{p(o, z|\theta)}{q^{(t+1)}(z)} 
\]

(60)

\[
= \sum_z q^{(t+1)}(z) \ln p(o, z|\theta) - \sum_z q^{(t+1)}(z) \ln q^{(t+1)}(z) 
\]

(61)

\text{const.}

\textbf{Result:} The parameters \( \theta \) maximizing Eq. (59) are

\[
\theta^{(t+1)} = \underset{\theta}{\text{argmax}} \sum_z q^{(t+1)}(z) \ln p(o, z|\theta) .
\]

(62)
Example 1 - Setting

\[ \pi_c = 0.6, \quad \pi_t = 0.4, \quad \sigma_c = 1, \quad \sigma_t = 2, \quad \mu_c = 5, \quad \mu_t = 10 \]

Data:
- 50 points from car distribution,
- 50 points from truck distribution,
- 1000 points from mixed distribution (car/truck coordinate unknown)

Experiment:
Employ EM algorithm for estimating \( \mu_1, \mu_2 \). Use different initializations.
Log-likelihood $\ell$ after 10 iterations of EM, depending on initialization $(\mu_1^{\text{init}}, \mu_2^{\text{init}})$.

Convergence in this case is quite fast (3 iterations are enough for most of the initialization values.)

Value of $(\mu_1, \mu_2)$ after 10 iterations, depending on initialization $(\mu_1^{\text{init}}, \mu_2^{\text{init}})$. The first point of convergence corresponds to the ground truth values $(\mu_1, \mu_2) = (5, 10)$. The second point is a only a local maximum of log-likelihood. It corresponds to car distribution approximating truck sample points, and vice versa.
Mixture Models

Generalization of the Motivation example with missing values.

\[
\mu_c = \frac{\sum_{i=1}^{M} q(z_i = \text{car}) y^*_i}{\sum_{i=1}^{M} q(z_i = \text{car})} \tag{63}
\]

\[
\sigma^2_c = \frac{\sum_{i=1}^{M} q(z_i = \text{car}) (y^*_i - \mu_c)^2}{\sum_{i=1}^{M} q(z_i = \text{car})} \tag{64}
\]

\[
\pi_c = \frac{\sum_{i=1}^{M} q(z_i = \text{car})}{M} \tag{65}
\]
Example: Mixture of Gaussians

Figure 7.10  a) Initial model. b) E-step. For each data point the posterior probability that it was generated from each Gaussian is calculated (indicated by color of point). c) M-step. The mean, variance and weight of each Gaussian is updated based on these posterior probabilities. Ellipse shows Mahalanobis distance of two. Weight (thickness) of ellipse indicates weight of Gaussian. d-t) Further E-step and M-step iterations.