

Lecture 16: 20 March, 2025

Madhavan Mukund

<https://www.cmi.ac.in/~madhavan>

Data Mining and Machine Learning
January–April 2025

Mixture models

- Probabilistic process — parameters Θ
 - Tossing a coin with $\Theta = \{Pr(H)\} = \{p\}$
- Perform an experiment
 - Toss the coin N times, $H\ T\ H\ H\ \dots\ T$
- Estimate parameters from observations
 - From h heads, estimate $p = h/N$
 - Maximum Likelihood Estimator (MLE)
- What if we have a **mixture** of two random processes
 - Two coins, c_1 and c_2 , with $Pr(H) = p_1$ and p_2 , respectively
 - Repeat N times: choose c_i with probability $1/2$ and toss it
 - Outcome: N_1 tosses of c_1 interleaved with N_2 tosses of c_2 , $N_1 + N_2 = N$
 - Can we estimate p_1 and p_2 ?

Mixture models ...

- Two coins, c_1 and c_2 , with $Pr(H) = p_1$ and p_2 , respectively
- Sequence of N interleaved coin tosses $H T H H \dots H H T$
- If the sequence is labelled, we can estimate p_1 , p_2 separately
 - $H T T H H T H T H H T H T H T H H T H T$
 - $p_1 = 8/12 = 2/3$, $p_2 = 3/8$
- What the observation is unlabelled?
 - $H T T H H T H T H H T H T H T H H T H T$
- Iterative algorithm to estimate the parameters
 - Make an initial guess for the parameters
 - Compute a (fractional) labelling of the outcomes
 - Re-estimate the parameters

Expectation Maximization (EM)

- Iterative algorithm to estimate the parameters

- Make an initial guess for the parameters
- Compute a (fractional) labelling of the outcomes
- Re-estimate the parameters

- $H\ T\ T\ H\ H\ T\ H\ T\ H\ H\ T\ H\ T\ H\ T\ H\ H\ T\ H\ T$

- Initial guess: $p_1 = 1/2$, $p_2 = 1/4$
- $Pr(c_1 = T) = q_1 = 1/2$, $Pr(c_2 = T) = q_2 = 3/4$
- For each H , likelihood it was c_i , $Pr(c_i | H)$, is $p_i/(p_1 + p_2)$
- For each T , likelihood it was c_i , $Pr(c_i | T)$, is $q_i/(q_1 + q_2)$
- Assign fractional count $Pr(c_i | H)$ to each H : $2/3 \times c_1$, $1/3 \times c_2$
- Likewise, assign fractional count $Pr(c_i | T)$ to each T : $2/5 \times c_1$, $3/5 \times c_2$

Expectation Maximization (EM)

- $H T T H H T H T H H T H T H H T H T$

- Initial guess: $p_1 = 1/2$, $p_2 = 1/4$

- Fractional counts: each H is $2/3 \times c_1$, $1/3 \times c_2$, each T : $2/5 \times c_1$, $3/5 \times c_2$

- Add up the fractional counts

- c_1 : $11 \cdot (2/3) = 22/3$ heads, $9 \cdot (2/5) = 18/5$ tails

- c_2 : $11 \cdot (1/3) = 11/3$ heads, $9 \cdot (3/5) = 27/5$ tails

- Re-estimate the parameters

- $p_1 = \frac{22/3}{22/3 + 18/5} = 110/164 = 0.67$, $q_1 = 1 - p_1 = 0.33$

- $p_2 = \frac{11/3}{11/3 + 27/5} = 55/136 = 0.40$, $q_2 = 1 - p_2 = 0.60$

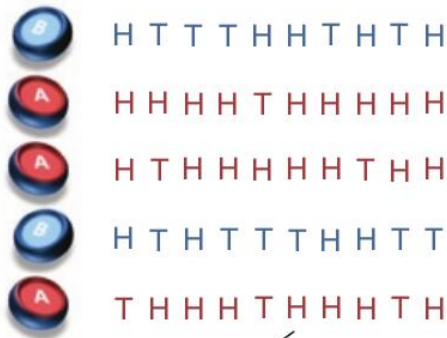
- Repeat until convergence

Expectation Maximization (EM)

- Mixture of probabilistic models (M_1, M_2, \dots, M_k) with parameters $\Theta = (\theta_1, \theta_2, \dots, \theta_k)$
- Observation $O = o_1 o_2 \dots o_N$
- **Expectation** step
 - Compute likelihoods $Pr(M_i|o_j)$ for each M_i, o_j
- **Maximization** step
 - Recompute MLE for each M_i using fraction of O assigned using likelihood
- Repeat until convergence
 - Why should it converge?
 - If the value converges, what have we computed?

EM — another example

- Two biased coins, choose a coin and toss 10 times, repeat 5 times



- If we know the breakup, we can separately compute MLE for each coin

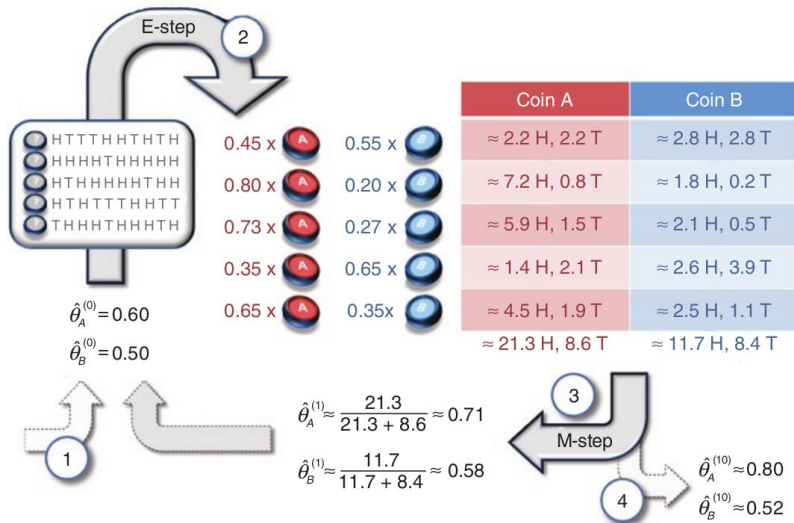
Coin A	Coin B
	5 H, 5 T
9 H, 1 T	
8 H, 2 T	
	4 H, 6 T
7 H, 3 T	
24 H, 6 T	9 H, 11 T

$$\hat{\theta}_A = \frac{24}{24 + 6} = 0.80$$

$$\hat{\theta}_B = \frac{9}{9 + 11} = 0.45$$

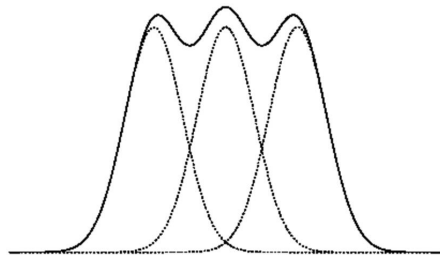
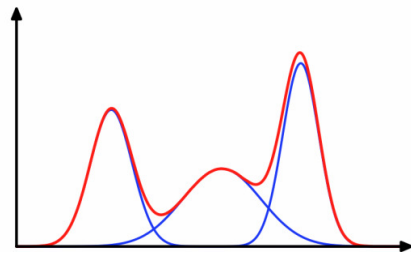
EM — another example

- Expectation-Maximization
- Initial estimates, $\theta_A = 0.6, \theta_B = 0.5$
- Compute likelihood of each sequence: $\theta^H (1 - \theta)^T$
- Assign each sequence proportionately
- Converge to $\theta_A = 0.8, \theta_B = 0.52$



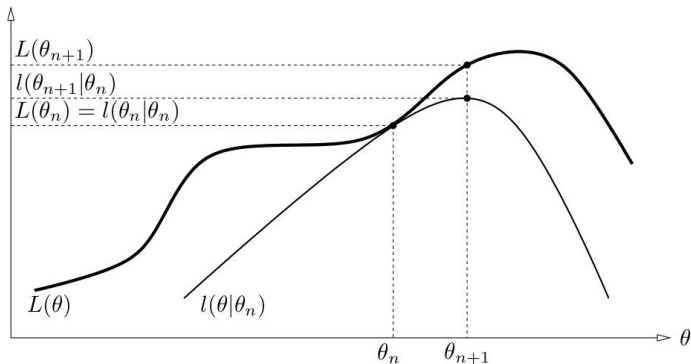
EM — mixture of Gaussians

- Sample uniformly from multiple Gaussians, $\mathcal{N}(\mu_i, \sigma_i)$
- For simplicity, assume all $\sigma_i = \sigma$
- N sample points z_1, z_2, \dots, z_N
- Make an initial guess for each μ_j
- $Pr(z_i | \mu_j) = \exp(-\frac{1}{2\sigma^2}(z_i - \mu_j)^2)$
- $Pr(\mu_j | z_i) = c_{ij} = \frac{Pr(z_i | \mu_j)}{\sum_k Pr(z_i | \mu_k)}$
- MLE of μ_j is sample mean, $\frac{\sum_i c_{ij} z_i}{\sum_i c_{ij}}$
- Update estimates for μ_j and repeat



Theoretical foundations of EM

- Mixture of probabilistic models (M_1, M_2, \dots, M_k) with parameters $\theta = (p_1, p_2, \dots, p_k)$
- Observation $O = o_1 o_2 \dots o_N$
- EM builds a sequence of estimates $\theta_1, \theta_2, \dots, \theta_n$
- $L(\theta_j)$ — log-likelihood function, $\ln \Pr(O \mid \theta_j)$
- Want to extend the sequence with θ_{n+1} such that $L(\theta_{n+1}) > L(\theta_n)$



- EM performs a form of gradient descent
- If we update θ_n to θ' we get a new likelihood $L(\theta_n) + \Delta(\theta', \theta_n)$ which we call $\ell(\theta' \mid \theta_n)$
- Choose θ_{n+1} to maximize $\ell(\theta' \mid \theta_n)$

Semi-supervised learning

- Supervised learning requires labelled training data
- What if we don't have enough labelled data?
- For a probabilistic classifier we can apply EM
 - Use available training data to assign initial probabilities
 - Label the rest of the data using this model — fractional labels
 - Add up counts and re-estimate the parameters

Semi-supervised topic classification

- Each document is a **multiset** or **bag** of words over a vocabulary $V = \{w_1, w_2, \dots, w_m\}$
- Each topic c has probability $Pr(c)$
- Each word $w_i \in V$ has conditional probability $Pr(w_i | c_j)$, for $c_j \in C$
 - Note that $\sum_{i=1}^m Pr(w_i | c_j) = 1$
- Assume document length is independent of the class
- Only a small subset of documents is labelled
 - Use this subset for initial estimate of $Pr(c)$, $Pr(w_i | c_j)$

Semi-supervised topic classification

- Current model $Pr(c)$, $Pr(w_i | c_j)$
- Compute $Pr(c_j | d)$ for each unlabelled document d
 - Normally we assign the maximum among these as the class for d
 - Here we keep fractional values
- Recompute $Pr(c_j) = \frac{\sum_{d \in D} Pr(c_j | d)}{|D|}$
 - For labelled d , $Pr(c_j | d) \in \{0, 1\}$
 - For unlabelled d , $Pr(c_j | d)$ is fractional value computed from current parameters
- Recompute $Pr(w_i | c_j)$ — fraction of occurrences of w_i in documents labelled c_j
 - n_{id} — occurrences of w_i in d
 - $Pr(w_i | c_j) = \frac{\sum_{d \in D} n_{id} Pr(c_j | d)}{\sum_{t=1}^m \sum_{d \in D} n_{td} Pr(c_j | d)}$

Clustering

- Data points from a mixture of Gaussian distributions
- Use EM to estimate the parameters of each Gaussian distribution
- Assign each point to “best” Gaussian
- Can tweak the shape of the clusters by constraining the covariance matrix
- Outliers are those that are outside $k\sigma$ for all the Gaussians

