

Mixture Models and the EM Algorithm

Lecture Notes Consolidated with Claude AI — CS6140 Machine Learning

Instructor's lecture plan for the note of the same title ([EM_MIXTURES_claude_1015.pdf](#)). Section numbers (§) refer to that note.

This supports the Week 8 lecture and **HW4 Problems 3** (including **3(C)**), **4** and the theory problem **6**; also optional Problem 11. It continues directly from **GDA_NB** — specifically from that note's three options, of which this is option ② pushed one step further. Proofs live in **EM_ADVANCED**; the forward link is to the **VAE** (Week 10, HW5 Problem 3).

Lecture plan (100 minutes, mostly on the board)

Time	On the board	Worked example / figure
0–8	One bump is not enough. Draw a three-hump histogram and fit the best single Gaussian to it — its peak lands in a valley. Hence $p(\mathbf{x}) = \sum_k \pi_k p_k(\mathbf{x})$. (§1.1)	Figure <code>mixture_motivation.png</code> : adult weights, three sub-populations, recovered $\pi = 0.30/0.50/0.20$.
8–18	The generative recipe and the variable we threw away. Roll a K -sided die to get z_i , draw from component z_i , <i>discard</i> z_i . The one-hot z_{ik} and the two sum identities. Why the $\log \sum_k$ in the likelihood is the whole difficulty. (§1.2)	Write both likelihoods side by side — $\log \sum$ (hard) vs. $\sum \log$ (easy) — and leave them up all lecture.
18–30	On-ramp: k-means, then soft k-means. If you knew z , fitting is trivial; if you knew μ , guessing z is trivial; so alternate. Then replace the hard 0/1 by $\exp(-\text{dist}) / \sum \exp(-\text{dist})$ — and notice you have just written a Gaussian mixture with $\Sigma_k = \sigma^2 I$. (§2)	Figure <code>kmeans_vs_em.png</code> : the contested band of points is the entire difference.
30–42	The E step is Bayes' rule. $\langle z_{ik} \rangle = \pi_k \mathcal{N}(\mathbf{x}_i; \mu_k, \Sigma_k) / \sum_\ell (\cdots)$. Three readings: posterior; $\mathbb{E}[z_{ik}]$; soft k -means. Hammer the π_k -vs- $\langle z_{ik} \rangle$ distinction — prior per component vs. posterior per point. (§3)	Ask the class what a component with $\pi_k = 0.02$ has to do to win a point. (Fit it 50× better.)
42–55	Complete-data likelihood with the indicator exponent; it is <i>linear</i> in z_{ik} , so the expectation passes straight through and every z_{ik} becomes $\langle z_{ik} \rangle$. That is the ELL / Q . (§4.1)	The board's $\hat{\Pi}_{ik}$ route and the note's $\langle z_{ik} \rangle$ are the same object — point at the notation bridge (§10.1).
55–68	M step: three weighted averages. $N_k = \sum_i \langle z_{ik} \rangle$; then μ_k (differentiate, Σ_k^{-1} cancels), Σ_k (differentiate w.r.t. Σ_k^{-1} again, as in GDA_NB), π_k (Lagrangian, and $\lambda = N$ falls out). (§4.2)	Do the π_k Lagrangian in full — it is three lines and students like seeing $\lambda = N$ appear.
68–78	What EM actually optimizes: the ELBO. State $\log p(\mathbf{x}) \geq \mathbb{E}_q[\log p(\mathbf{x} z)] - \text{KL}(q \ p(z))$; map $q \rightarrow \langle z_{ik} \rangle$, $p(z) \rightarrow \pi_k$. Draw the staircase: E step makes the bound touch, M step climbs it. State monotonicity; do <i>not</i> prove it. (§6.1, §6.2)	The staircase diagram (§6.1) and the mapping table. This block is HW4 P3(C) and P6, and the prerequisite for the VAE in Week 10 — it is the one block that must not be cut.
78–88	Algorithm box; initialization and restarts; the three numerical issues (log-sum-exp, singular Σ_k , empty components). Then the hand example. (§5, §9.1)	$x = 1, 2, 3, 6$, $K = 2$, $\mu = (0, 4)$, $\sigma^2 = 4$: $r_{\cdot 1} = (0.731, 0.5, 0.269, 0.018)$, $\mu_1 \rightarrow 1.743$; $\ell = -9.133 \rightarrow -7.621 \rightarrow -7.514$.
88–96	Same machinery, new component: the coins. Binomial likelihood, the $\binom{n}{h}$ cancels, $p_k = S_k / (nR_k)$. Mention Poissons in one line to make the pattern obvious. (§7.1, §7.2)	Figure <code>em_iterations.png</code> . $n = 10$, $h = 9, 8, 2, 1$, start $(0.6, 0.4)$: $r = (0.962, 0.919, 0.081, 0.038)$, $R_1 = 2$ exactly, $p \rightarrow 0.811$, $q \rightarrow 0.189$.
96–100	Label switching and weak identifiability — why overlapping coins recover badly no matter what. HW4 map; forward pointer to HMMs and the VAE. (§8, §9.2, §10.2)	Figure <code>coins_recovery.png</code> : separated coins vs. overlapping ones.

Board layout suggestion. Top-left, all lecture: the two likelihoods from minute 8 ($\log \sum$ vs. $\sum \log$).

Every subsequent step is an answer to “how do we get from the first to the second”, and saying so out loud at minutes 42, 55 and 68 makes the lecture one argument instead of a sequence of formulas.

If you are running short. Compress 18–30 (the k -means on-ramp) to five minutes by drawing only the figure, or move 88–96 (the coins) to a five-minute sketch and let the note carry it — HW4 Problem 4 is well enough specified to follow on its own. Protect 68–78.

Questions that reliably come up.

- “*Why not just assign each point to its most likely component?*” That is k -means, and it is maximizing a different objective in which borderline points are declared certain. See the remark in §4.1.
- “*How do I pick K ?*” Not from the training likelihood — it rises forever. Held-out likelihood or BIC; the measurement is in **EM_ADVANCED** §3. For HW4, K is given.
- “*My likelihood went down.*” Then there is a bug, and §6.2 lists the six usual ones. Ask which of μ_k or Σ_k they updated first.
- “*Is this the same EM as in [some other course/paper]?*” Almost certainly yes — show the notation bridge (§10.1); w_k , $\gamma(z_{nk})$ and $\langle \Pi_{ik} \rangle$ are all objects on this board.