Solution 1: Gaussian Mixture Models (Final Exam 2019, Problem 3)

The proposed Gaussian Mixture Model (GMM) predicts the probability that a lake is poisonous. The data with pH level, l_i, of each lake is given in ascending order, i.e., $\mathcal{D} = \{l_i\}_{i=1}^{N}$, such that $l_i \leq l_j$ for $i < j$. By the problem definition, we have $K = 2$ classes, with the first class representing the poisonous lakes, and they have their respective mean and variance values, μ_i, σ_i^2 for $i = 1, 2$, where it is hypothesized that $\mu_1 \geq \mu_2$. Furthermore, according to the overall split of the two classes amongst all lakes, their weights (or probability of selection are) are p_1 and p_2, respectively.

a) The probability that a random lake is poisonous given its pH level, l, can be written as

$$\begin{aligned} P(\text{poisonous} \mid l) &= \frac{f_{\text{pH}|\text{lake}}(l \mid \text{poisonous})P(\text{poisonous})}{f_{\text{pH}}(l)} \\ &= \frac{\mathcal{N}(\mu_1, \sigma_1^2)p_1}{\mathcal{N}(\mu_1, \sigma_1^2)p_1 + \mathcal{N}(\mu_2, \sigma_2^2)p_2} \end{aligned}$$

b) The pseudocode for EM algorithm for training our GMM with hard decisions is given below. The initial clusters are a split of the dataset at index k, i.e., $\mathcal{B}_1[0] = \{l_{k+1}, \dots, l_N\}$ and $\mathcal{B}_2[0] = \{l_1, \dots, l_k\}$.

- **Do**:
  - update p_i, μ_i, σ_i^2, for $i = 1, 2$
    - $p_1 = \frac{|\mathcal{B}_1|}{N}$ and $p_2 = \frac{|\mathcal{B}_2|}{N}$
    - $\mu_1 = \frac{1}{|\mathcal{B}_1[n]|} \sum_{l_i \in \mathcal{B}_1[n]} l_i$ and $\mu_2 = \frac{1}{|\mathcal{B}_2[n]|} \sum_{l_i \in \mathcal{B}_2[n]} l_i$
    - $\sigma_1^2 = \frac{1}{|\mathcal{B}_1[n]|} \sum_{l_i \in \mathcal{B}_1[n]} (l_i - \mu_1)^2$ and $\sigma_2^2 = \frac{1}{|\mathcal{B}_2[n]|} \sum_{l_i \in \mathcal{B}_2[n]} (l_i - \mu_2)^2$
  -
- **While**: $\mathcal{B}_1[n] \neq \mathcal{B}_1[n+1]$ and $\mathcal{B}_2[n] \neq \mathcal{B}_2[n+1]$.

c)

d)

Solution 2:

Hello again