ECE421 - Winter 2022
Homework Problems - Tutorial #9

Theme: EM Algorithm and SVMs

Due: March 27, 2022 11:59 PM

Question 1 (Problem 3 - Final 2019)

There was a village surrounded by hundreds of lakes. Each lake was either **poisonous** or **healthy**. Anyone who ate fish from a poisonous lake would die immediately while anyone who ate fish from a healthy lake would survive. All fish looked and tasted identical and villagers knew no other way of knowing whether a lake was poisonous or healthy, which of course was a huge problem for the villagers.

Fortunately, a famous chemist visited the village and was told of this dilemma. The chemist suggested using the pH level of water to determine whether a given lake was poisonous or healthy, and hypothesized that lakes with poisonous fish would have higher pH value than healthy lakes. Accordingly, the chemist visited each lake and collected the pH value of the water in each lake. The data set is denoted by $\mathcal{D} = \{l_1, l_2, \ldots, l_N\}$, where l_i denotes the pH level of the lake $i \in \{1, 2, \ldots, N\}$. Assume that $l_1 \leq l_2 \leq l_3 \leq \ldots \leq l_N$.

You are hired as a machine learning scientist to help determine the probability that a randomly selected lake is poisonous given its pH value. In order to do this you propose to use a Gaussian Mixture Model (GMM) as follows.

- $\Pr(\text{lake is poisonous}) = p_1$
- $\Pr(\text{lake is healthy}) = p_2$
- $f(\text{pH} = l | \text{ lake is poisonous}) \sim \mathcal{N}(\mu_1, \sigma_1^2)$
- $f(\text{pH} = l | \text{ lake is healthy}) \sim \mathcal{N}(\mu_2, \sigma_2^2)$
- $\mu_1 \geq \mu_2$ as the pH value for poisonous lakes will be higher on average.

Here $f(\cdot)$ denotes the conditional density function for the pH value. You decide to use the EM algorithm to train the above GMM on the dataset $\mathcal{D}$.

(a) Using the above GMM, provide an expression of the probability that a randomly selected lake is poisonous given that its pH level is measured to be l.

(b) Write down the pseudocode for the EM algorithm with hard decisions that finds the parameters of the GMM. Assume that we initialize the algorithm in such a way that $\mathcal{B}_2 = \{l_1, l_2, \ldots, l_K\}$ denotes one cluster of lakes and $\mathcal{B}_1 = \{l_{K+1}, l_{K+2}, \ldots, l_N\}$ denotes its complement.

(c) Write down the pseudocode for the EM algorithm with soft decisions that finds the parameters of the GMM. Do state the initialization of your algorithm explicitly.

(d) Suppose that $N = 5$ and we observe $\mathcal{D} = \{1, 2, 3, 6, 10\}$. Let $\mathcal{B}_2 = \{1, 2, 3\}$ and $\mathcal{B}_1 = \{6, 10\}$. Execute the hard decision EM algorithm and compute the parameters of the GMM.

Question 2 (Problem 8.2 from LFD)

Consider the dataset X with three data points in $\mathbb{R}^2$, and the label vector y consisting of the labels of the three data points:

$$X = \begin{bmatrix} 0 & 0 \\ 0 & -1 \\ -2 & 0 \end{bmatrix}, \quad y = \begin{bmatrix} -1 \\ -1 \\ +1 \end{bmatrix}. \tag{1}$$

Use X and y to train a SVM by solving the following optimization problem. Obtain the optimal hyperplane $(b^\star, w^\star)$ and its margin.

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2} w^\top w \\ \text{subject to} \quad & y_n(w^\top x_n + b) \geq 1, n = 1, \ldots, N \end{aligned} \tag{2}$$