CSCI-B 505 Applied Algorithms (3 cr.) № 6

Dr. H. Kurban

Computer Science

School of Informatics, Computing, and Engineering

Indiana University, Bloomington, IN, USA

November 15, 2021

Contents

Problem 1: Expectation-Maximization Algorithm for Clustering [30 pt.]

Implement expectation-maximization algorithm for Gaussian mixture models (see the EM algorithm below) in *Python* and call this program G_k. As you present your code explain your protocol for

1. initializing each Gaussian
2. deciding ties
3. stopping criteria

Problem 2: Analysis of the EM over Real-world Data Sets [20 pt.]

Run your EM program, G_k, against the Ringnorm and Ionosphere data sets. Discuss your results.

- Ringnorm Data Set
- Ionosphere Data Set

Run G_k with $k = 2, \ldots, 5$ (20 runs each for each k). Report error rates and iteration counts for each k using whisker plots. An example of whisker can be found in the appendix.

A simple error rate can be calculated as follows:

- After G_k converges, combine the clusters to ended up with two clusters for any k as follows: Since the true clusters are known for a given arbitrary blocks number, final clusters are determined by measuring the Euclidean (this is the easiest choice) distances between true cluster centers and predicted cluster centers.

In other words, you will always calculate the error for $k = 2$ since there are only 2 clusters in the given data sets. Below is an example of error calculation for Ionosphere data set. You can similarly calculate an error rate for Ringnorm data set. After G_k converges, partition the data points to k-clusters, $C_1, \ldots, C_k$ using likelihoods (hard assignment). A data point in a cluster is classified as good if $P(g_i) > P(b_i)$ and bad otherwise.

For each cluster C_i form two counts (over Ionosphere Data Set):

$$\begin{aligned} g_i &\leftarrow \sum_{\delta_j \in C_i} [\delta_j = \mathsf{g}], \quad \text{good} \\ b_i &\leftarrow \sum_{\delta_j \in c_i.B} [\delta_j == \mathsf{b}], \quad \text{bad} \end{aligned}$$

where $[x = y]$ returns 1 if True, 0 otherwise. For example, $[2 = 3] + [0 = 0] + [34 = 34] = 2$

We can now calculate a simple error rate. Assume G_i is good. Then the error is:

$$error(C_i) = \frac{b_i}{b_i + g_i}$$

We can find the total error rate easily:

$$Error(\{C_1, C_2\}) = \sum_{i=1}^{2} error(C_i)$$

Problem 3: Algorithm Design [30pt]

Problem 3.1

Given a text $\mathcal{D}$ and a pattern $\mathcal{P}$, describe an $\Omega(\mathsf{d+p})$ time method for finding the longest prefix of $\mathcal{P}$ that is a substring of $\mathcal{D}$. The lengths of $\mathcal{D}$ and $\mathcal{P}$ are d and p, respectively.

Problem 3.2

X, Y, and Z are three arrays and each has m elements. For an arbitrary integer t, describe $O(m^2 logm)$-time algorithm to determine if there exist numbers, x in X, y in Y, and z in Z, such that t = x+y+z.

Problem 3.3

Describe an efficient algorithm for deleting a string from a compressed trie and analyze its running time.

Directions

Please follow the syllabus guidelines in turning in your homework. While testing your programs, run them with a variety of inputs and discuss your findings. This homework is due Sunday, Nov 23, 2021 10:00pm. **OBSERVE THE TIME**. Absolutely no homework will be accepted after that time. All the work should be your own.

Appendix

The EM algorithm

This part is provided to help you implement the EM algorithm.

Let **D** $= \{\mathbf{x_j} \mid j = 1, \ldots, n\}$ be the data set where each $\mathbf{x_j} \in \Re^d$ ($\Re$: Reals) and **D** is a mixture of a Gaussian. Given **D**, the number of blocks k, and convergence threshold ϵ, the EM-T algorithm partition data into k clusters, C $= \{C_1, C_2, \ldots, C_k\}$, where each $C_i \in$ C can be characterized as a Gaussian distribution. If each cluster $C_i \in$ C can be represented by a multivariate normal distribution (MVN):

$$f(\mathbf{x_j}|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} * \exp\Big\{ -\frac{1}{2}(\mathbf{x_j} - \mu_i)^T \Sigma_i^{-1}(\mathbf{x_j} - \mu_i) \Big\}$$

where $\mu_i \in \Re^d$ and $\Sigma_i \in \Re^{d \times d}$ denote cluster mean and covariance matrix for cluster C_i, respectively, and they are unknown parameters. $f(\mathbf{x_j}|\mu_i, \Sigma_i)$ represents the probability density at $\mathbf{x_j}$ for cluster C_i. Lastly, **D** is a mixture of $C_1, C_2, \ldots, C_k$.

The algorithm iteratively alternates between (1) computing log-likelihood of each data point being from each Gaussian (E-step) (2) recalculating the parameters (M-step). Iteration continues until a set of means is stable.

- **Initialization:**
  - μ_i is randomly selected from **D** for each cluster.
  - $\Sigma_i \leftarrow I$. For each cluster, the covariance matrix is a $d \times d$ identity matrix.
  - $P(C_i) = \frac{1}{k}$. The priors are uniformly initialized.

EXPECTATION-MAXIMIZATION $(\mathbf{D}, k, \epsilon)$:

$t \leftarrow 0$

// Initialization

Randomly initialize $\boldsymbol{\mu}_1^t, \ldots, \boldsymbol{\mu}_k^t$

$\boldsymbol{\Sigma}_i^t \leftarrow \mathbf{I},\ \forall i = 1, \ldots, k$

$P^t(C_i) \leftarrow \frac{1}{k},\ \forall i = 1, \ldots, k$

repeat

  $t \leftarrow t + 1$

  // Expectation Step

  for $i = 1, \ldots, k$ *and* $j = 1, \ldots, n$ **do**

    $w_{ij} \leftarrow \frac{f(\mathbf{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \cdot P(C_i)}{\sum_{a=1}^{k} f(\mathbf{x}_j | \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a) \cdot P(C_a)}$ // posterior probability $P^t(C_i | \mathbf{x}_j)$

  // Maximization Step

  for $i = 1, \ldots, k$ **do**

    $\boldsymbol{\mu}_i^t \leftarrow \frac{\sum_{j=1}^{n} w_{ij} \cdot \mathbf{x}_j}{\sum_{j=1}^{n} w_{ij}}$ // re-estimate mean

    $\boldsymbol{\Sigma}_i^t \leftarrow \frac{\sum_{j=1}^{n} w_{ij} (\mathbf{x}_j - \boldsymbol{\mu}_i)(\mathbf{x}_j - \boldsymbol{\mu}_i)^T}{\sum_{j=1}^{n} w_{ij}}$ // re-estimate covariance matrix

    $P^t(C_i) \leftarrow \frac{\sum_{j=1}^{n} w_{ij}}{n}$ // re-estimate priors

until $\sum_{i=1}^{k} \left\| \boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^{t-1} \right\|^2 \leq \epsilon$

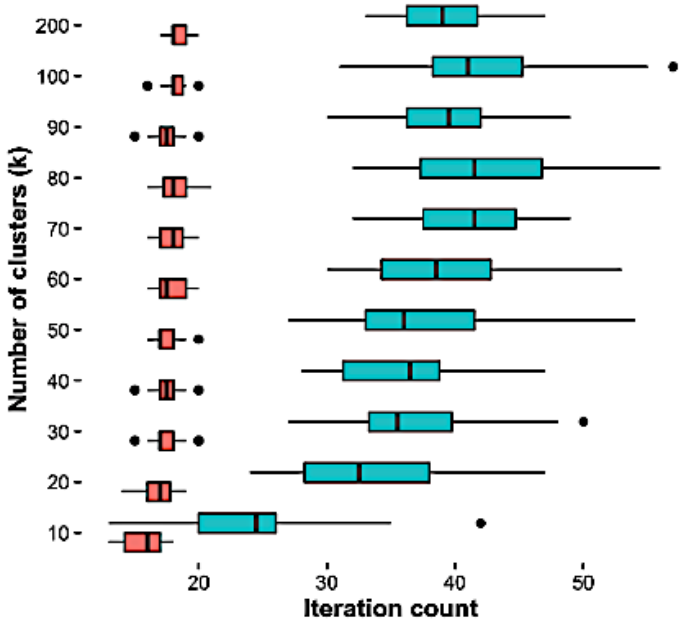

Figure 1: An example of whisker plot