Statistical Machine Learning (STAT W4400)

Homework 4

Linan Qiu
lq2137

November 24, 2015

As usual, the R code is available at https://github.com/linanqiu/stat-w4400-homework

1 PCA

1.1 How many principal components

Since dimensionality of raw data is 10304, the number of principal components is 10304.

1.2 Approximately reconstruct a specific face image

$$\begin{aligned}\hat{x}_i &= \sum_{j=1}^{d} \langle \mathbf{x}_i, \xi_j \rangle \xi_j \\ &= \sum_{j=1}^{d} c_{i,j} \xi_j\end{aligned}$$

where c_i is the projection of $\mathbf{x_i}$ onto ξ_i and ξ_i is the eigenvector corresponding to the j-th largest eigenvalue. Dimensionality reduction occurs because instead of storing the complete $\mathbf{x}_i \in \mathbb{R}^{10304}$, we store only $\mathbf{c}_i \in \mathbb{R}^d$ and d eigenvectors $\xi_i \in \mathbb{R}^{10304}$, and usually $d \ll 10304$.

2 k-means

2.1 C_k intuition

C_k is the set of all (indices of) data points assigned to cluster k (ie. $m_i = k$.)

2.2 Describe the objective

Breaking down the objective,

$$||x_i - x_j||_2^2$$

is the squared distance between data point x_i and data point x_j.

$$\sum_{i \in C_k} \sum_{j \in C_k} ||x_i - x_j||_2^2$$

is the sum of squared distances between all possible pairs of data points of $x_i | i \in C_k$, ie. the sum of distances between all data points that are assigned to k.

$$\frac{1}{|C_k|} \sum_{i \in C_k} \sum_{j \in C_k} ||x_i - x_j||_2^2$$

scales the sum based on the number of elements in C_k, giving a measure, for each cluster, of how far apart the data points are from each other. This is also a measure of the mean distance of all data points in k to the cluster center (exact proof of this intuition is the next part of the question).

$$\sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i \in C_k} \sum_{j \in C_k} ||x_i - x_j||_2^2$$

sums this measure for all clusters $k \in K$, measuring the total mean distance of all data points to their respective cluster centers.

By minimizing this, we minimize the total mean distance of all data points to their cluster centers, effectively finding clustering assignments $\mathbf{m}$ that "bunches / clusters" data points closest together in the same group.

2.3 Prove that $\frac{1}{|C_k|}\sum_{i\in C_k}\sum_{j\in C_k}||x_i - x_j||_2^2 = 2\sum_{i\in C_k}||x_i - \mu_k||_2^2$

In this step, we are formalizing the intuition earlier that $\frac{1}{|C_k|}\sum_{i\in C_k}\sum_{j\in C_k}||x_i - x_j||_2^2$ is indeed a measure of the mean distance of all data points in k to the cluster center.

$$\begin{aligned}\frac{1}{|C_k|}\sum_{i\in C_k}\sum_{j\in C_k}||x_i - x_j||_2^2 &= \frac{1}{|C_k|}\sum_{i\in C_k}\sum_{j\in C_k}||x_i - \mu_k + \mu_k - x_j||_2^2 \\ &= \frac{1}{|C_k|}\sum_{i\in C_k}\sum_{j\in C_k}\left(||x_i - \mu_k||_2^2 + ||\mu_k - x_j||_2^2 + 2\left\langle x_i - \mu_k, \mu_k - x_j\right\rangle\right)\end{aligned}$$

For each of the three terms in the round brackets,

$$\begin{aligned}\frac{1}{|C_k|}\sum_{i\in C_k}\sum_{j\in C_k}||x_i - \mu_k||_2^2 &= \frac{1}{|C_k|}\sum_{i\in C_k}|C_k| * ||x_i - \mu_k||_2^2 \\ &= \sum_{i\in C_k}||x_i - \mu_k||_2^2 \\ \frac{1}{|C_k|}\sum_{i\in C_k}\sum_{j\in C_k}||\mu_k - x_j||_2^2 &= \frac{1}{|C_k|}\sum_{j\in C_k}|C_k| * ||\mu_k - x_j||_2^2 \\ &= \sum_{j\in C_k}||\mu_k - x_j||_2^2 \\ &= \sum_{i\in C_k}||\mu_k - x_i||_2^2 \\ &= \sum_{i\in C_k}||x_i - \mu_k||_2^2 \\ \frac{1}{|C_k|}\sum_{i\in C_k}\sum_{j\in C_k}2\left\langle x_i - \mu_k, \mu_k - x_j\right\rangle &= \frac{2}{|C_k|}\sum_{i\in C_k}\left\langle x_i - \mu_k, \sum_{j\in C_k}(\mu_k - x_j)\right\rangle \\ &= \frac{2}{|C_k|}\sum_{i\in C_k}\left\langle x_i - \mu_k, 0\right\rangle \\ &= 0\end{aligned}$$

Since by definition, $\sum_{j\in C_k}(\mu_k - x_j) = |C_k|\mu_k - \sum_{j\in C_k}x_j = |C - k|\mu_k - |C_k|\mu_k = 0$. Then,

$$\frac{1}{|C_k|}\sum_{i\in C_k}\sum_{j\in C_k}||x_i - x_j||_2^2 = 2\sum_{i\in C_k}||x_i - \mu_k||_2^2$$

2.4 Prove result implies that the k-meansalgorithm decreases the objective with each step of the algorithm

Let $\sum_{i\in C_k}||x_i - \mu_k||_2^2 = \text{RSS}$

$$\begin{aligned}
\frac{\delta \text{RSS}}{\delta \mu_k} &= \frac{\delta}{\delta \mu_k}\sum_{i\in C_k}||x_i - \mu_k||_2^2 \\
&= \frac{\delta}{\delta \mu_k}\sum_{i\in C_k}\sum_{d=1}^{D}(x_{i,d} - \mu_{k,d})^2 \\
&= -2\sum_{i\in C_k}\sum_{d=1}^{D}(x_{i,d} - \mu_{k,d}) \\
&= 2\left[\sum_{i\in C_k}\sum_{d=1}^{D}x_{i,d} - \sum_{i\in C_k}\sum_{d=1}^{D}\mu_{k,d}\right] \\
&= 2\left[\sum_{i\in C_k}x_i - \sum_{i\in C_k}\mu_k\right] = 0
\end{aligned}$$

Then,

$$\begin{aligned}
\sum_{i\in C_k}x_i &= \sum_{i\in C_k}\mu_k \\
\sum_{i\in C_k}x_i &= |C_k|\mu_k \\
\mu_k &= \frac{1}{|C_k|}\sum_{i\in C_k}x_i
\end{aligned}$$

This is the definition of μ_k.

This makes sense, since there are two steps per iteration: reassignment and recomputation. RSS decreases in the reassignment step since each data point is assigned to the closest centroid, so the distance it contributes to the total RSS decreases. Furthermore, it decreases in the recomputation step because the new centroid is the k for which RSS_k is

minimum. As shown above, we minimize RSS_k when the old centroid is replaced with the new centroid.

3 Multinomial Clustering

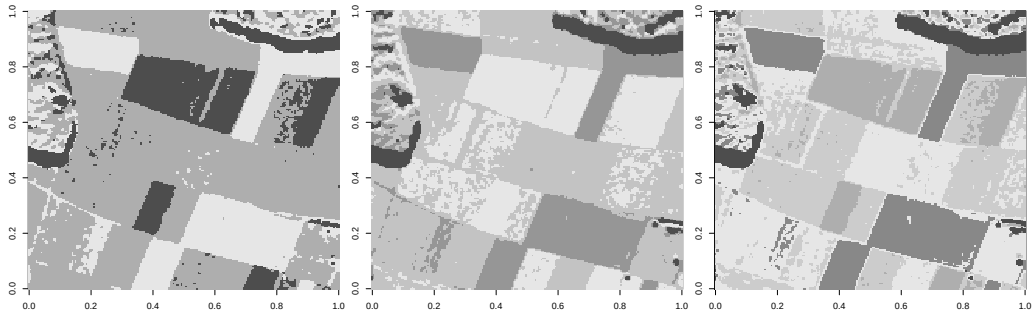
3.1 Implement EM algoithm in `R`

Done in `em.R`.

3.2 Run algorithm on $K \in \{3, 4, 5\}$

Done in `run-seq.R`. Produces `plot3.pdf`, `plot4.pdf`, `plot5.pdf`. A parallel version that does the algorithm on different Ks in parallel is in `run-par.R`. This was fun `:D`

3.3 Visualize the results



(a) Results for $K = 3$ (b) Results for $K = 4$ (c) Results for $K = 5$

Figure 1: Results for histogram cluster for $K \in \{3, 4, 5\}$