CS 189 Fall 2017 — Introduction to Machine Learning — HW14

This homework is due **Thursday, December 7 at 10pm.**

1 Getting Started

You may typeset your homework in latex or submit neatly handwritten and scanned solutions. Please make sure to start each question on a new page, as grading (with Gradescope) is much easier that way! Deliverables:

1. Submit a PDF of your writeup to assignment on Gradescope, "HW[n] Write-Up"
2. Submit all code needed to reproduce your results, "HW[n] Code".
3. Submit your test set evaluation results, "HW[n] Test Set".

After you've submitted your homework, be sure to watch out for the self-grade form.

(a) Before you start your homework, write down your team. Who else did you work with on this homework? List names and email addresses. In case of course events, just describe the group. How did you work on this homework? Any comments about the homework?

None. Alone
Comment: terrible hw timing

(b) Please copy the following statement and sign next to it:

I certify that all solutions are entirely in my words and that I have not looked at another student's solutions. I have credited all external sources in this write up.

I certify that all solutions are entirely in my words and that I have not looked at another student's solutions. I have credited all external sources in my write up

Problem # 2

(a) K-means:

$$F = \min_{\pi, \mu \in \mathbb{R}^{K \times d}} \sum_{k=1}^{K} \sum_{i \in \pi_k} \| x_i - \mu_k \|_2^2$$

if we stack N vectors of x_i's, we have:

$$F = \min_{\pi, \mu} \left\| \underbrace{\begin{bmatrix} x_1^T \\ \vdots \\ x_{\pi_1}^T \\ \vdots \\ x_N^T \end{bmatrix}}_{X} - \begin{bmatrix} \mu_1^T \\ \mu_1^T \\ \vdots \\ \mu_K^T \end{bmatrix} \right\|_F^2$$

The stack of μ_i^T's has some repetition, we can decompose this one into $Y\mu$ where μ is also the stack of μ_i's without repetition:

$$\begin{array}{c} \pi_1 \\ \\ \pi_2 \\ \vdots \end{array} \begin{bmatrix} \mu_1^T \\ \vdots \\ \mu_1^T \\ \mu_2^T \\ \vdots \\ \mu_2^T \\ \vdots \\ \mu_K^T \end{bmatrix} = \begin{array}{c} \pi_1 \\ \\ \pi_2 \\ \vdots \end{array} \underbrace{\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & 1 \\ 0 & 0 & \cdots & \end{bmatrix}}_{K} \begin{bmatrix} \mu_1^T \\ \mu_2^T \\ \vdots \\ \mu_K^T \end{bmatrix} = Y\mu$$

It's like:

$$\begin{bmatrix} a_1 & a_2 \\ a_1 & a_2 \\ b_1 & b_2 \\ b_1 & b_2 \\ b_1 & b_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a_1 & a_2 \\ b_1 & b_2 \end{bmatrix}$$

we can see: Y satisfies $\|y_i^T\|_0 = 1 \quad \forall i$

thus
$$F = \min_{\mu \in \mathbb{R}^{K \times d}, Y \in \mathbb{R}^{N \times K}} \| X - Y\mu \|_F^2$$

and $y_i \in \{0,1\}^K \quad \|y_i^T\|_0 = 1 \quad \forall i$

which is a special case of K-SVD w/ $s=1$

(b) The regression in Algorithm 3 is:

$$\hat{\beta}, k_l = \underset{\beta \in R^l, k}{\operatorname{argmin}} \| x - \sum_{D_j \in B_l \cup \{D_k\}} \beta_j D_j \|_2^2$$

the initial residue error is $R_1 = x$, after each loop β & k are choosen so that $\| x - \sum \beta_j D_j \|_2^2$ min

we can see that if we can not choose any β, k s.t

$$\| x \|_2^2 > \| x - \sum \beta_j D_j \|_2^2$$

then we can choose $\hat{\beta} = 0$, by this way:

$$\| x \|_2^2 = \| x - \sum \beta_i D_j \|_2^2$$

i.e. the residue error does not increase

Since $\hat{\beta} = 0$ can always be chosen, Alg 3 cannot increase the residue error of linear regression (but, in the worst case, it remains unchanged).

c.) The objective function: $F = \| X - ZD \|_F^2 = \sum_1^N \| x_n^T - z_n^T D \|_F$

consisting of N Froberius terms of $\| x_n^T - z_n^T D \|_F$

In algorithm 2:

for $n = 1 \to N$:

$z_n' = \cdots$

if $\| x_n^T - z_n^T D \|_F > \| x_n^T - z_n'^T D \|_F$ then

update $z_n = z_n'$

endif

end for

That is we replace a Froberius term of F only the new term is strictly less than the old term, thereby F either reduces or remains unchanged

Thus, over N iterations, the objective function F can not increase (also it is lower-bounded by 0)

(d) The error matrix $E_K \in R^{n \times d}$ come from:

$$F = X - \sum_{j=1}^{K} z_j D_j^T = \left(X - \sum_{j \neq k} z_i D_j^T \right) - z_k^T d_k = E_k - z_k^T d_k$$

we want: $\min \| F \| = \min \| E_K - z_K^T d_K \|$

we can shrink the matrix F down and the min value does not change

In algorithm 5, the way to shrink F down is to obtain $E_K^R \in R^{|w_K| \times d}$ by choosing the rows of E_K s.t the index of that row is in w_K. It is nothing but

$$\min \| F \| = \min \| \Omega_K F \| = \min \| \Omega^K E^K - \Omega^K z_K^T d_K \|$$
$$= \min \| E_R^K - z_{R_K}^T d_K \|$$

where $\Omega_K \in R^{|w_K| \times n}$ w/ 1 on $(i, w_K(i))$-th, and 0 otherwise

by this way, we enforce the sparsity constrain on Z since Z is shrunk by discarding the 0's entries of each row

(i.e 0's entries are not filled → sparsity is not changed)

=) $\min \| F \| = \min \| E_R^K - z_{R_K}^T d_K \|$

According to the Eckart-Young theorem, if we can decompose $E_R^K = U \Lambda V^T$, then we can choose d_K to be the first column of V^T and $z_{R_K}^T$ to be the the first column of U multiplied by Λ_{11} → algorithm 5

(e) this comes straightforward from (c):

1.) the objective function can not increases

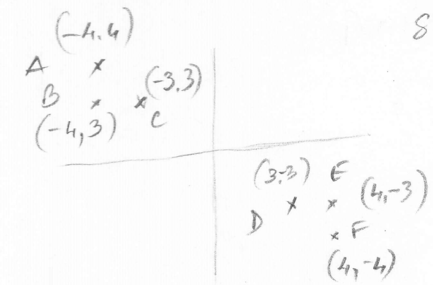
2.) It is either reduced or unchanged after each update

3.) It is lower-bounded by 0

$\Rightarrow$ it must converged.

Problem # 4

Conduct 2-means according to the following diagram for 2 scenario of clusterings.



1: Group 1 : A, B

Group 2 : C, D, E, F

2. Group 1 : A, B, C

Group 2 : D, E, F.

Which scenario is closer to 2-means ?

Solution :

1.)

$$\mu_1 = \frac{A+B}{2} = \frac{(-4,4)+(-4,3)}{2} = (-4,3.5)$$

$$\mu_2 = \frac{C+D+E+F}{4} = \frac{(-3,3)+(3,-3)+(4,-3)+(4,-4)}{4}$$

$$= (1,-1.75)$$

$$F_1 = \|A-\mu_1\|^2 + \|B-\mu_1\|^2 + \|C-\mu_2\|^2 + \|D-\mu_2\|^2$$
$$+ \|E-\mu_2\|^2 + \|F-\mu_2\|^2$$
$$= 0^2 + 0.5^2 + 0^2 + (-0.5)^2 + (-4)^2 + 4.75^2 + 2^2 + (-1.25)^2$$
$$+ 3^2 + 4.75^2 + 3^2 + (2.25)^2 = 84.25$$

2.)

$$\mu_1 = \frac{A+B+C}{3} = \frac{(-4;3)+(-4;4)+(-3;3)}{3} = \left(-\frac{11}{3};\frac{10}{3}\right)$$

$$\mu_2 = \frac{C+D+E}{3} = \frac{(3,-3)+(4;-3)+(4;-4)}{3} = \left(\frac{11}{3};-\frac{10}{3}\right)$$

$$\Rightarrow F_2 = \frac{1}{9}+\frac{1}{9}+\frac{1}{9}+\frac{4}{9}+\frac{4}{9}+\frac{1}{9}+\frac{4}{9}+\frac{1}{9}+\frac{1}{9}+\frac{1}{9}+\frac{1}{9}+\frac{4}{9} = \frac{8}{3}$$

$F_1 > F_2$

$\Rightarrow$ F_2 is closer to 2-means.