Homework 2
Computer Science
Fall 2016
B565

Professor Dalkilic

September 9, 2016

Directions

Please follow the syllabus guidelines in turning in your homework. I will provide the LaTeX of this document too. You may use it or create one of your own. This homework should be started quickly. Sometimes there are natural questions arising from code. Within a week, AIs can contact students to examine code; students must meet within three days. The session will last no longer than 5 minutes. If the code does not work, the grade for the program may be reduced. Lastly, source code cannot be modified post due date.

k-means Algorithm in Theory

This part of the problem asks you to reflect on k-means and work through its theoretical elements. I have written algorithm below. Answer the subsequent questions.

1: **ALGORITHM** `k-means`
2: **INPUT** (data Δ, distance $d : \Delta^2 \to \mathbb{R}_{\geq 0}$, centroid number k, threshold τ)
3: **OUTPUT** (Set of centoids $\{c_1, c_2, \ldots, c_k\}$)
4:
5: *** $Dom(\Delta)$ denotes domain of data.
6:
7: *** Assume centroid is structure $c = (v \in DOM(\Delta), B \subseteq \Delta)$
8: *** $c.v$ is the centroid value and $c.B$ is the set of nearest points.
9: *** c^i means centroid at i^{th} iteration.
10:
11: $i = 0$
12: *** Initialize Centroids
13: **for** $j = 1, k$ **do**
14: $\quad c_j^i.v \leftarrow random(Dom(\Delta))$
15: $\quad c_j^i.B \leftarrow \emptyset$
16: **end for**
17:
18: **repeat**
19: $\quad i \leftarrow i + 1$
20: $\quad$ *** Assign data point to *nearest* centroid
21: $\quad$ **for** $\delta \in \Delta$ **do**
22: $\quad\quad c_j^i.B \leftarrow c.B \cup \{\delta\}$, where $\min_{c_j^i}\{d(\delta, c_j^i.v)\}$
23: $\quad$ **end for**
24: $\quad$ **for** $j = 1, k$ **do**
25: $\quad\quad$ *** Get size of centroid

26: $n \leftarrow |c_j^i.B|$
27: *** Update centroid with average
28: $c_j^i.v \leftarrow (1/n) \sum_{\delta \in c_j^i.B} \delta$
29: *** Remove data from centroid
30: $c_j^i.B \leftarrow \emptyset$
31: **end for**
32: *** Calculate scalar product (abuse notation and structure slightly)
33: *** See notes
34: **until** $((1/k) \sum_{j=1}^{k} \ ||c_j^{i-1} - c_j^i||) < \tau$
35: **return** $(\{c_1^i, c_2^i, \ldots, c_k^i\})$

k-means on a tiny data set.

Here are the inputs:

$$\Delta = \{(2,5),(1,5),(22,55),(42,12),(15,16)\} \quad (1)$$
$$d((x_1,y_1),(x_2,y_2)) = [(x_1-x_2)^2+(y_1-y_2)^2)]^{1/2} \quad (2)$$
$$k = 2 \quad (3)$$
$$\tau = 10 \quad (4)$$

Observe that $Dom(\Delta) = \mathbb{R}^2$. We now work through k-means. We ignore the uninformative assignments. We remind the reader that T means transpose.

1: $i \leftarrow 0$
2: *** Randomly assign value to first centroid.
3: $c_1^0.v \leftarrow random(Dom(\Delta)) = (16,19)$
4: *** Randomly assign value to second centroid.
5: $c_2^0.v \leftarrow random(Dom(\Delta)) = (2,5)$
6: $i \leftarrow i+1$
7: *** Associate each datum with nearest centroid
8: $c_1^1.B = \{(22,55),(42,12),(15,16)\}$
9: $c_2^1.B = \{(2,5),(1,5)\}$
10: *** Update centroids
11: $c_1^1.v \leftarrow (26.3, 27.7) = (1/3)((22,55)+(42,12)+(15,16))$
12: $c_2^1.v \leftarrow (1.5,5) = (1/2)((2,5)+(1,5))$
13: *** The convergence condition is split over the next few lines to explicitly show the calculations
14: $(1/k)\sum_{j=1}^{k} \ ||c_j^{i-1} - c_j^i|| = (1/2)(||c_1^0 - c_1^1|| + ||c_2^0 - c_2^1||) = (1/2)(||\binom{2}{5} - \binom{1.5}{5}|| + ||\binom{16}{19} - \binom{26.3}{27.7}||)$
15: $= (1/2)[(\binom{.5}{0}^{\mathsf{T}}\binom{.5}{0})^{(1/2)} + (\binom{-9.7}{-8.7}^{\mathsf{T}}\binom{-9.7}{-8.7})^{(1/2)})] = (1/2)(\sqrt{.5} + \sqrt{169.7}) \sim (1/2)(13.7) = 6.9$
16: Since the threshold is met $(6.9 < 10)$, k-means stops, returning $\{(26.3,27.7),(1.5,5)\}$

Questions

1. Does k-means always converge? Given your answer, a bound on the iterate must be included. How is its value determined?

2. Lines 12-16 of the k-means algorithm describe initialization of the centroids. Why is this code problematic? What are some implications of using k-means?

3. What is the run-time of this algorithm (include your new parameter from Question 1).

4. We describe two problems that arise when using k-means in practice. Assume the datum is $\delta \in \Delta$, the centroids are c_i, c_j for $i \neq j$ and distance d.

   - *Ties* occur when $d(c_i, \delta) = d(c_j, \delta)$. Of course, there can be threeway, fourway, ..., k-way ties. One solution is to randomly assign the datum to one of the two centroids. What are two other solutions to this problem?

- *Centroid collapse* occurs when $d(c_i, c_j) \sim 0$. Like ties, this can include more than two. One is to find the median m of the union of the two centroids and then assign values less than the median to one and values greater than the median to the other, taking into account an odd number will be the problem above. What are two other solutions? Observe that an additional threshold on centroids, $\tau_c > 0$, is needed, to determine whether $d(c_i, c_j) \leq \tau_c$ is true. First, how would τ_c be determined? Second, where in the algorithm should this be checked?
- Modify the k-means algorithm to address ties and collapsing centroids. Explicitly add pseudo-code to the algorithm and call this k-meansr.

Integration

We will look at the problem of integrating two pieces of data through a metric. The data are described by $([X : t], d_x), ([Y : u], d_u)$ where $X : t$ means it is type t, $Y : u$ is type u, and d_x, d_y distance metrics. We integrate the data and now need a metric $([X : t] \times [Y : u], d)$. Is this possible? We need to prove that d is a metric. To make notation easier, assume $Z = [X : t] \times [Y : u]$. For $(a, b) \in Z^2$, we write a_0 to mean the t type leftside of the product and b_0 for the t type rightside. For example, $Z = [N : \mathsf{int}] \times [S : \mathsf{string}]$. $(a, b) = ((34, \texttt{two}), (100, \texttt{three}))$, then $a_0 = 34, b_0 = 100$ and $a_1 = \texttt{two}, b_1 = \texttt{three}$.

Let's define one of the simplest metrics. $d : Z^2 \to \mathbb{R}_{\geq 0}$ where:

$$d(a, b) \quad = \quad d_x(a_0, b_0) + d_y(a_1, b_1)$$

Now we show reflexivity, symmetry, and transitivity.

- $(\forall, a \in Z)\ d(a, a) = 0$. Then $d(a, a) = d_x(a_0, a_0) + d_y(a_1, a_1) = 0$
- $(\forall\, a, b)\ d(a, b) \to d(b, a)$.

$$d(a, b) = d_x(a_0, b_0) + d_y(a_1, b_1) = d_x(b_0, a_0) + d_x(b_1, a_1) = d(b, a)$$

- $(\forall\, a, b, c)\ d(a, b) + d(b, c) \geq d(a, c)$

$$\begin{aligned} d(a, b) + d(b, c) &= d_x(a_0, b_0) + d_x(b_0, c_0) + d_y(a_1, b_1) + d_y(b_1, c_1) \\ &\geq d_x(a_0, c_0) + d_y(a_1, c_1) = d(a, c) \end{aligned}$$

Suppose we have $[X : \mathsf{int}]$ are the number of cable subscription cancelations (say, *per* hour). We find data $[Y : \mathsf{char}]$ that indicates whether there was "good" programming at that time (we're purposely being vague). The ordering is $\texttt{n} < \texttt{o} < \texttt{g} < \texttt{e}$, $\texttt{e}$ being the best. We integrate this and get:

X	Y
14	g
45	o
54	g
21	n
60	o

Although we didn't need to use the type information explicitly, its presence shows that we can build metrics over disparate kinds of integrated data. Design a simple metric, different from the one above, for this integrated data. Prove it is a metric.

1. We can combine multiple metrics to built more sophisticated measures of dissimilarity. This problem has to do with different metrics over the same data. Let $x = \{a, b, c, d\}, y = \{a, b, e\}, z = \{b, f\}, w = \{a, d, f, e\}$. Here are several metrics:

$$
\begin{aligned}
d_1(x,y) &= \begin{cases} 0, & x = y \\ 1, & x \neq y \end{cases} \quad \text{For objects } x, y. \\
J(x,y) &= |x \cap y| / |x \cup y| \quad \text{For sets } x, y. \\
d_2(x,y) &= 1 - J(x,y) \quad \text{For sets } x, y. \\
c(x,y) &= \begin{cases} 0, & x = y \\ 1, & otherwise \end{cases} \quad \text{for individual characters, } e.g., \texttt{a} = \texttt{b} \\
d_3(\mathbf{x},\mathbf{y}) &= \Sigma_{i=0}^{n-1} c(\mathbf{x}[i], \mathbf{y}[i]) \quad n = ||\mathbf{x}||, \text{the length of the string.} \\
d_4(\mathbf{x},\mathbf{y}) &= \left| \frac{\mathbf{x}^\mathsf{T}\mathbf{y}}{||\mathbf{x}||\,||\mathbf{y}||} \right| \quad \text{for vectors } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n
\end{aligned}
$$

Calculate the following:

(a) For every i, find $d_i(x, w)$

(b) Find the d_i that has the minimum value for x, z.

(c) Which distance gives the the maximum value for any pairs?

(d) True or False. For any set v, $d_1(v, v) = d_2(v, v) = d_3(v, v) = d_4(v, v)$.

2. We have shown that metrics can be combined. Why is the important to integration? Prove or disprove the following are metrics (using d_i from above):

(a) $d_{i'}(x, y) = \frac{d_i(x,y)}{1+d_i(x,y)}$ for every i.

(b) $d_{i'}(x, y) = \alpha d_i(x, y)$ for $\alpha \in \mathbb{R}_{>0}$

(c) $d_5(x, y) = d_1(x, y) + 3d_2(x, y)$

(d) $d_6(x, y) = d_2(y, x)$

(e) $d_7(x, y) = d_3(x, y)d_2(x, y)$

(f) $d_8(x, y) = \sum_{i=1}^{4} d(x, y)$

3. Read the paper, "A Survey on Tree Edit Distance and Related Problems," by Bille[**?**]. In no more than two paragraphs, discuss what is *most* relevant to either datamining or data science.

Application of k-means and Data Prepartion to Medical Data

This problem examines Wolberg's breast cancer data[**?**] that we will denote by Δ. This set, though tiny, provides a good start for k-means and preprocessing. Δ is found at
`http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/`

data	`breast-cancer-wisconsin.data`
description	`breast-cancer-wisconsin.names`

While you will read the data description to more fully understand the format, we create some attribute names to make discussion easier.

ID	Description	Domain	Attribute Name
1.	Sample code number	string	SCN
2.	Clump Thickness	$\mathbb{N}$	A_2
3.	Uniformity of Cell Size	$\mathbb{N}$	A_3
4.	Uniformity of Cell Shape	$\mathbb{N}$	A_4
5.	Marginal Adhesion	$\mathbb{N}$	A_5
6.	Single Epithelial Cell Size	$\mathbb{N}$	A_6
7.	Bare Nuclei	$\mathbb{N}$	A_7
8.	Bland Chromatin	$\mathbb{N}$	A_8
9.	Normal Nucleoli	$\mathbb{N}$	A_9
10.	Mitoses	$\mathbb{N}$	A_{10}
11.	Class:	char	C

1. **Datamining Problem** Suppose you're working to help a clinic serve a community that has limited resources to identify and treat breast cancer. The cost of a biopsy is from \$1000 to \$5000, since it requires a pathologist. The cost of a masectomy is \$15,000 to \$55,000 (these are representative costs in 2016). The cost of a computer program, ignoring the modest fixed cost of machine *etc.*, is \$10.

   (a) What is the total cost of the biopsies in Δ when done by a pathologist? Assume the computer can identify 90% of the cases to nearly 100% accuracy. What is the cost of the computer program?

   (b) What would have been the likely total cost of masectomies?

   (c) Assuming a 70% mortality rate for untreated in year five, how many deaths does the data suggest in five years?

   (d) Compose a succint problem statement that you imagine is pertinent to this scenario.

2. **Data Preparation** Ignoring the `Sample code number` (SCN),

   (a) Ignoring the SCN and C columns, how many attributes (or features) does Δ have?

   (b) Let $\Delta^{miss} \subset \Delta$ be the data that has missing values. How many missing values exist (total)? What is the size of Δ^{miss}?

   (c) How many patients have missing values?

   (d) Give the SCNs for that have missing values.

   (e) Of these data, would you have recommended re-examination for the women? What would be the costs both for the pathologist and computer program?

   (f) Is the amount of missing data significant from an algorithmic perspective?

   (g) Assess the significance of either keeping or removing the tuples with unknown data. You should consider the human element too.

   (h) Repair Δ^{miss} by replacing unknown data using one of the techniques we discussed in class. This will be presented as $(\texttt{SCN}, A_i, v)$ where `SCN` is the tuple key, A_i is the attribute, and v is the new value. Create a `CSV` file `DeltaFix.csv` for this data. Call the entire data set, including the values that have been replaced, as Δ_1^{clean}.

3. **Data Analysis**

   (a) Using either `MySQL`, `SQL Server` or `PostgreSQL`, built a table and load the fixed data set. Connect to `R` so that you can quickly and easily perform analysis. Using `R`,

   (b) Plot histograms for each attribute and C.

   (c) Find the mean, median, mode, and variance of each attribute.

(d) For each pair A_i, A_j, $i \neq j$, find the Pearson's correlation coefficient. This provides an insight to the linearity of the attributes. To remind you,

$$\rho_{X,Y} = \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

σ is the standard deviation

μ is the mean

Eis the expectation

How is ρ related to $cos\theta = \frac{\mathbf{xy}}{||\mathbf{x}||||\mathbf{y}||}$? Remove one of the pairs of attributes that are strongly linearly related for every pair of attributes. Call this Δ_2^{clean}. What is the purpose of this step?

4. Implement k-means so that you can cluster Δ_2^{clean} without using C. Upon stopping, you will calculate the quality of the centroids and of the partition. For each centroid c_i, form two counts:

$$b_i \leftarrow \sum_{\delta \in c_i.B} [\delta.C = 2], \quad \text{benign}$$

$$m_i \leftarrow \sum_{\delta \in c_i.B} [\delta.C = 4], \quad \text{malignant}$$

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

The centroid c_i is classified as benign if $b_i > m_i$ and malignant otherwise. We can now calculate a simple error rate. Assume c_i is benign. Then the error is:

$$error(c_i) = \frac{m_i}{m_i + b_i}$$

We can find the total error rate easily:

$$Error(\{c_1, c_2, \ldots, c_k\}) = \sum_{i=1}^{k} error(c_i)$$

Report the total error rates for $k = 2, \ldots 5$ for 20 runs each, presenting the results that are easily understandable. Plots are generally a good way to convey complex ideas quickly. Discuss your results and include your initial problem statement.

What to Turn-in

- The *pdf of the written answers to this document.
- The code for k-means, `R`.
- The AIs can schedule a time to verify your codes works. If there is a subsequent time-stamp to the due date of the source code, the grade may be reduced.