Homework 2
Computer Science
Fall 2015
B565

Professor Dalkilic
Sunday, October 18, 9:00 p.m.

September 30, 2015

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.

1 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.

```
ALGORITHM kmeans
INPUT (data $\Delta$, distance $d : \Delta^2 \rightarrow \mathbb{R}_{\geq 0}$, centoid number $k$, threshold $\tau$)
OUTPUT (Set of centoids $c_1, c_2, \ldots, c_k$)
  ▷ Assume centroid is structure $c = (v \in DOM(\Delta), B \subseteq \Delta)$, so $c.v$ is the centroid value and $c.B$ is the set of nearest points.
                                              ▷ $\tau$ is a percentage change from previous centroids
                                              ▷ $DOM(\Delta)$ is domain of data
$i = 0$                                       ▷ Initialize iterate where superscript is iteration
for $j = 1, k$ do                             ▷ Initialize Centroids
    $c^i_j.v \leftarrow random(Dom(\Delta))$
    $c^i_j.B \leftarrow \emptyset$
end for
$f_i = 2d((0,0), c^0_1) + 2d((0,0), c^0_2)$
Calculate the difference between current an starting centroids.
repeat
    $i \leftarrow i + 1$
    for $\delta \in \Delta$ do
        $c^i_j.B \leftarrow c.B \cup \{\delta\}$, where $\min_{c^i_j}\{d(\delta, c^i_j.v)\}$
                                              ▷ Associate a data point $\delta$ with the nearest centroid $c^i_j.v$
    end for
    for $j = 1, k$ do
        $c^i_j.v \leftarrow ave(c^i_j.B)$     ▷ Update centroid to average of nearest set of data
        $c^i_j.B \leftarrow \emptyset$
    end for
    $f_i \leftarrow \Sigma^k_{j=1}\Sigma^k_{\ell=1} d(c^i_j.v, c^{i-1}_\ell)$   ▷ Find the difference between previous centroids and current centroids
```

```
25: until $(|f_i - f_{i-1}| < \tau(f_{i-1}))$
26: return $(c_1^i, c_2^i, \ldots, c_3^i)$        ▷ If there's less than $\tau$ change, then we're finished.
```

1.1 Small, partial example

Let's assume the data is:

$$\Delta = \{(2,5),(1,5),(22,55),(42,12),(15,16)\}$$

and the distance function is:

$$d((x_1,y_1),(x_2,y_2)) = [(x_1-x_2)^2 + (y_1-y_2)^2)]^{1/2}$$

and $k = 2$ and $\tau = 0.10$.

```
1: $c_1^0.v \leftarrow (10,7)$
2: $c_1^0.B \leftarrow \emptyset$
3: $c_2^0.v \leftarrow (16,19)$
4: $c_2^0.B \leftarrow \emptyset$
5: Consider the point $\delta_1 = (2,5)$
6: $d(\delta_1, c_1^0.v) = 8$ and $d(\delta_1, c_2^0.v) = 20$, so $c_1^1.B \leftarrow B \cup \{\delta_1\}$
7: Continuing for the rest of $\Delta$ we have $c_1^1.B = \{\delta_1, \delta_2\}$ and $c_2^2.B = \{\delta_3, \delta_4, \delta_5\}$
8: We now find the best representative of each centroid: in this case, the simple average (nearest integer) of each dimension.
9: $c_1^1.v = ((2+1)/2, (5+5)/2) = (2,5)$ and $c_2^1.v = ((22+42+15)/3, (55+12+16)/3)) = (26,28)$
10: We can now find the difference between the previous centroids $(i = 0)$ and current $(i = 1)$
11: We can assume the initial points were all zero; therefore, the difference between the start and $c_0$ is
12: $f_0 = 2d((0,0), c_1^0) + 2d((0,0), c_2^0) = 73$
13: $f_1 = d(c_1^0, c_1^1) + d(c_1^0, c_2^1) + d(c_2^0, c_1^1) + d(c_2^0, c_2^1) = 68$
14: Since $|f_1 - f_0| < \tau(f_0)$, we can stop.
```

1.2 Questions

1. Does the algorithm always converge? Given your answer, what extra formal parameter is needed to the function. How do you decide it's actual value?

2. What is the reason initialization of k-means is problematic?

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

4. In line 12, $f_i = 2d((0,0), c_1^0) + 2d((0,0), c_2^0)$. Why are the distances multiplied by 2?

5. In line 21, we use *ave* to indicate average; however, the centroid is more correctly the best representative of the data that nearest. Give two more ways (functions) that yield a best representative.

6. Modify the algorithm to identify when two centroids are *too* close to one another. There will likely be more than one extra parameter needed to the algorithm. Discuss what your modification does.

7. Let $x = \{a,b,c,d\}, y = \{a,b,e\}, z = \{b,f\}, \mathcal{U} = \{a,b,c,d,e,f\}$. Compute the distances using

$$d(x,y) \quad = \quad \begin{cases} 0, & x = y \\ 1, & x \neq y \end{cases} \tag{1}$$

The signature of the distance function is: $d : \mathsf{Set}^2 \rightarrow \mathbb{R}_{\geq 0}$.

(a) $\neg x = \mathcal{U} - \{a,b,c,d\} = \{e,f\}$

(b) $\neg \mathcal{U} = \emptyset$

(c) $d(x,y) = 1$

(d) $d(x \cap y, \{a, b\}) = 0$

(e) $d(x, x \cup y) =$

(f) $d(\neg(x \cap y), \neg x \cup \neg y) =$

$$\begin{aligned} J(x, y) &= |x \cap y| / |x \cup y| \\ d(x, y) &= 1 - J(x, y) \end{aligned}$$

The signature of the distance function is: $d : \mathsf{Set}^2 \to \mathbb{R}_{\geq 0}$.

(a) $d(x, y) =$

(b) $d(x \cap y, \{a, b\}) =$

(c) $d(x, x \cup y) =$

(d) $d(\neg(x \cap y), \neg x \cup \neg y) =$

8. Assume bit vectors (as strings of 0's and 1's) over the set $\{a, b, c, d, e, f\}$. Then $x = \{a, b, c, d\}$ is $\mathbf{x} = 111100$. Similarly, $\mathbf{y} = 110010, \mathbf{z} = 010001$. Individually, $\mathbf{x}[0] = 1, \mathbf{x}[1] = 1$, *etc.* To remind the student, $\perp$ means the program stopped because of a bad computation. Let the distance function be the Hamming distance between the vector:

$$c(x, y) = \begin{cases} 0, & x = y \\ 1, & otherwise \end{cases} \quad \text{for individual characters} \tag{2}$$

$$d(\mathbf{x}, \mathbf{y}) = \Sigma_{i=0}^{n} c(\mathbf{x}[i], \mathbf{y}[i]) \quad n = ||\mathbf{x}||, \text{the length of the string.} \tag{3}$$

The signature of the distance function is: $d : \mathsf{String}^2 \to \mathbb{R}_{\geq 0}$. Assume we have some string functions and a constant:

$$\begin{aligned} \textvisiblespace &= \text{space} && (4) \\ concat(\texttt{b}, \texttt{at}) &= \texttt{bat} && (5) \\ concat(\texttt{bat}, \epsilon) &= \texttt{bat} && (6) \\ contat(\epsilon, \texttt{bat}) &= \texttt{bat} && (7) \\ upper(\texttt{a}) &= \texttt{A} && (8) \\ space(\texttt{b\textvisiblespace a\textvisiblespace t}) &= \texttt{bat} && (9) \end{aligned}$$

(a) $d(\texttt{x}, \texttt{y}) =$

(b) $d(concat(\texttt{x}, \texttt{x}), concat(\texttt{x}, \texttt{y})) =$

(c) Discuss the two problems above.

(d) $d(\texttt{N\textvisiblespace orth}, \texttt{nort\textvisiblespace h}) =$

(e) $d(upper(space(\texttt{N\textvisiblespace orth})), upper(space(\texttt{nort\textvisiblespace h}))) =$

(f) $d(\texttt{north}, \texttt{south}) =$

(g) Because strings are seldom fixed to any length, finding distance is even more difficult. strings of unequal length. The table below shows *some* strings indicating the direction prefixed or suffixed to addresses:

NORTH
North
N
N.
north
nor
n.
␣n

So the original distance would simply not function, *e.g.,*

$$d(\texttt{N.}, \texttt{North}) \quad = \quad \perp \tag{10}$$

How would you rewrite the distance function to effectively deal with this problem?

9. Assume your data is a set and string pair, $\delta = (\mathsf{Set}\ x, \mathsf{String}\ y)$. Create a metric over this pair. In otherwords,

$$d((\mathsf{Set}\ x_1, \mathsf{String}\ y_1), (\mathsf{Set}\ x_2, \mathsf{String}\ y_2)) \quad =$$

Demonstrate that it is indeed a metric.

2 Application of k-means to medical data

This problem examines Wolberg's breast cancer data[1]. The data 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`

1. Suppose you're working to help a clinic serve a community that has limited resources to identify then treat breast cancer. The cost of a biopsy is from \$1000 to \$5000. The cost of a masectomy is \$15,000 to \$55,000 (these are representative costs in 2015).

   (a) What was the total cost of the biopsies?

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

2. Ignoring the `Sample code number` (SCN), how many attributes does Δ have?

3. How many missing values exist (total)? How many patients have missing values? Give the SCNs for that have missing data. Of these data, would you have recommended re-examination for the women? What would be the cost be? What is the error rate (in otherwords, given x as the number of patients, what is $f(x) = y$ where y is the number of mistakes. Remove the tuples that have missing data. Let Δ^* be a cleaned Δ: the tuples with the missing values are removed. `R` offers several ways to remove unknown data, though you are free to write your own code. Let $\Delta^m = \Delta - \Delta^*$. For each $\delta \in \Delta^m$, replace the unknown data using one of the techniques we discussed in class; alternatively, you may employ your won approach. No matter how you decide to replace the unknowns, explain fully. The final data should be presented as $(\texttt{SCN}, A_i, \mathit{data})$ where `SCN` is the tuple key, A_i is the attribute, and *data* is the new data.

   (a) Is the amount of missing data significant?

   (b) Assess the significance of either keeping or removing the tuples with unknown data. You should consider both the morbidity and cost.

4. Assume the attribute `Clump Thickness` is A_1, `Uniformity of Cell Size` is A_2 and so on. Attribute A_{10} has only two domain values and is the classifier. For Δ^* and the attributes $A_i, 1 \leq i \leq 9$

   (a) which A_i has the greatest variance? You will write an `R` function that takes a list of numbers and returns the variance.

   (b) which A_i has the lowest entropy? You may use the `R` package `entropy` by Hausser and Strimmer.

   (c) Fill-in the table below with the KL distance for attribute pairs. For this we construct a mass function P_i over A_i by simple counting. For a cell whose row, column entries are A_i, A_j, find $d_{KL}(P_i||P_j)$. You may use an existing `R` function for this, but you need to provide sufficient package details for someone who would consider using that package.

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9
A_1	0								
A_2		0							
A_3			0						
A_4				0					
A_5					0				
A_6						0			
A_7							0		
A_8								0	
A_9									0

The KL distance between attributes of the cancer set.

5. Implement k-means so that you can cluster Δ^*. Since we have labels for the data, we can measure the quality of the clustering. If an element of δ is correctly clustered in c_i (nearest to the correct centroid), then it is considered a True Positive (TP). If an element that correctly belongs to c_i is clustered in a different c_j (in other words, nearer and incorrect centroid), then the element is a False Positive (FP). The Positive Predictive Value (PPV) is

$$PPV = \frac{TP}{TP + FP} \tag{11}$$

   Investigate varying the number of blocks as well as the attributes used. There are a modest number of attributes, so should use the powerset. Discuss the result with simply finding pairs of correlations.

6. One of the most common techniques in assessing function is using V-fold cross validation. The idea is simple. Suppose $|\Delta^*| = N$. Partition Δ^* into $V = 10$ sets $D^* = \{D_1^*, D_2^*, \ldots, D_{10}^*\}$ such that each $|D_i^*| = \frac{N}{10}$ tuples and all D_i, D_j are pairwise disjoint. The task is to use $V - 1$ sets to train and the remaining d to test.

Train	Test	PPV Result
$\texttt{kmeans}(D^* - \{D_1^*\})$	D_1^*	α_1
$\texttt{kmeans}(D^* - \{D_2^*\})$	D_2^*	α_2
$\vdots$	$\vdots$	$\vdots$
$\texttt{kmeans}(D^* - \{D_{10}^*\})$	D_{10}^*	α_{10}

   The total PPV is then

$$PPV(\Delta) = (1/10)\Sigma_{i=1}^{10}\alpha_i \tag{12}$$

Calculate the PPV using V-fold cross validation. Discuss your results.

3 Checking Results

```
INPUT TOTInterBlockDis(Set $X = \{B_1, B_2, \ldots, B_n\}$, distance $d$)
                                        ▷ $X$ is a partition and $B_i$ are the blocks.
OUTPUT $R_{\geq 0}$ $v$
$v \leftarrow 0$
for $i = 1, n - 1$ do
    for $j = i + 1, n$ do
        $v \leftarrow v + \texttt{InterBlockDis}(B_i, B_j, d)$
    end for
end for
```

```
return $v$

INPUT InterBlockDis(Set $X = \{x_1, \ldots, x_n\}$, Set $Y = \{y_1, \ldots, y_m\}$, distance $d$)
OUTPUT $R_{\geq 0}$ $v$
$v \leftarrow 0$
for $i = 1, n$ do
    for $j = 1, m$ do
        $v \leftarrow v + d(i, j)$
    end for
end for
return $v$
```

References

[1] William H. Wolberg and O.L. Mangasarian. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. 87:9193–9196, 1990.