# Computer Science
# Data Mining Core
# Homework III Fall 2013
# Indiana University
# Bloomington, IN

Mehmet M. Dalkilic

October 2, 2013

## Preliminaries

The due date is tentatively set Sunday October 13, 11:00 p.m.

**Data**: data $\Delta$, clusters $k$, distance $d : \Delta^2 \to \mathbb{R}_{\geq 0}$
**Result**: centroids $C_1, C_2, \ldots, C_\ell$
**begin**
COMMENT build centroids by random values, pick points, *etc.*
$i \leftarrow 0$
$\mathcal{C}^i = \{C_1^i, C_2^i, \ldots, C_k^i\}$
**repeat**
**for** $\delta \in \Delta$ **do**
find nearest centroid using $d(\delta, C_j^i)$
COMMENT $\delta$ belongs to one and only one centroid
**end**
**for** $C_j^i \in \mathcal{C}^i$ **do**
$C_j^{i+1} \leftarrow$ average$(\{\delta \in C_j^i\})$
**end**
$i \leftarrow i + 1$
**until** threshold on $\mathcal{C}^{i-1}$ and possibly $i$ ;
**end**

**Algorithm 1**: $k$-means

## Problem One: Missing Data

Assume you have a data series $\Delta = a, b, a, b, b, a, ?, a, b, b$ where the ? is missing data. Explain how to replace this using uniform distribution, random distribution imputed from the data, and $\sim \mathcal{B}(p, n)$. What are the sigificant differences between these three approaches?

# Problem Two: 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. Write a reasonably formal statement that uses $\Delta$. What is the size of $\Delta$?
   (a) Is this a large data set?
   (b) Discuss the number of attributes with the data's size
2. Ignoring the `Sample code number` (SCN), how many attributes does $\Delta$ have?
3. How many missing values are there? Give the SCNs for that have missing data. Remove the tuples that have missing data. Let $\Delta^*$ be a cleaned $\Delta$: 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 $d \in \Delta^m$, 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 $(\texttt{SCN}, A_i, data)$ 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.
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 $\Delta^*$ 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 (see Algorithm 1) so that you can cluster $\Delta^*$. Call this program $C_{km}$. For this first iteration, you may assume that $1 \leq k \leq 10$ and that all domain values are numeric. Write $C_{km=i}$ for $i$ blocks (or clusters). Assume there are $k$ blocks $C_1, C_2, \ldots, C_k$. If an element of $C_i$ is correctly clustered in $C_i$, then it is considered a True Positive (TP). If an element that correctly belongs to $C_i$ is clustered in a different $C_j$, then the element is a False Positive (FP). The Positive Predictive Value (PPV) is

$$PPV = \frac{TP}{TP + FP} \tag{1}$$

In this problem we will investigate $C_{km}$'s PPV varying the attributes used. Specifically, create a table that pairs $C_{km=2}$'s PPV using the attributes shown in the table below:

| $C_{km=2}(\Delta^*)$ | PPV |
|---|---|
| $A_1, \ldots, A_9$ | |
| $A_1, \ldots, A_7$ | |
| $A_1, \ldots, A_5$ | |
| $A_1, \ldots, A_3$ | |
| $A_1, A_2$ | |

6. One of the most common techniques in assessing function is using $V$-fold cross validation. The idea is simple. Suppose $|\Delta^*| = N$. Partition $\Delta^*$ 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 |
|---|---|---|
| $C_{km=2}(D^* - \{d_1^*\})$ | $C_{km=2}(d_1^*)$ | |
| $C_{km=2}(D^* - \{d_2^*\})$ | $C_{km=2}(d_2^*)$ | |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $C_{km=2}(D^* - \{d_{10}^*\})$ | $C_{km=2}(d_{10}^*)$ | |

Fill-in the PPV for each fold. Create a weighted PPV, $(1/10)\Sigma_{i=1}^{10} PPV_i$

# Problem Three: Astronomy

There are three files: *main* ($\sim 7K$ galaxies) has the data on the main sample of radio galaxies, *candidates* ($\sim 761K$ galaxies) non-R-AGN galaxies, and *control_2* ($\sim 7K$ galaxies) that are the best matches in *candidates* for each galaxy in *main*.

The features for each galaxy are:

- A unique galaxy ID
- galaxy type
- right ascension
- declination
- $u$-band magnitude
- $r$-band magnitude
- the spectroscopic redshift

The galaxy type is one of three integers: 0, 1, 2. 0 means it is a brightest cluster galaxy, 1 is a member of a galaxy cluster, and 2 is a galaxy that is 'in the field.' The right ascension and declination are the coordinates of the galaxies. The $u$ and $r$ magnitudes are measures of the brightness of the galaxies, and $u - r$ gives a number that we call the color. The larger $u - r$ value is, the redder the galaxy is. Finally, the spectroscopic redshift is related to the distance to the galaxy.' The four criteria we used to choose the galaxies in the control sample are galaxy type, $r$-band magnitude, $u$-band minus $r$-band, which we call '$u - r$ color', and spectroscopic redshift.

The task is to build a data set, called *control*, from *candidates*. For each galaxy $g \in main$, there is a set of galaxies $G' \subset candidate$ such that $\min_{R(g,g')}\{g' \in candidate\} = G'$

The definition of $R$ is given in the subsection below.

**Data**: data $main + candidate$, distance $R : \Delta^2 \to \mathbb{R}_{\geq 0}$
**Result**: *control*
**begin**
  $control \leftarrow \emptyset$
  **for** $g \in main$ **do**
    $control \leftarrow control \cup \{g' \in candidate | \min\{R(g, g')\}\}$
  **end**
**end**
**return** *control*

**Algorithm 2**: Build control

## Populating *control*

The distance function assume the galaxy type is identical (from *main* and *candidate*). Each control galaxy was selected to closely match the $r$ magnitude, $u-r$ color, and the redshift of its corresponding R-AGN. These properties were used for selection because the redshift and $r$ magnitude give the optical luminosity, which together with the $u-r$ are a proxy for the stellar mass of the galaxy while the $u - r$ color is a proxy for the star formation history. The selection was done by minimizing the function:

$$R(g, g') = \left( (\frac{z_g - z_{g'}}{0.2596})^2 + (\frac{r_g - r_{g'}}{8.6})^2 + (\frac{(u_g - r_g) - (u_{g'} - r_{g'})}{11.84})^2 \right)^{1/2} \quad (2)$$

where $z_g$ means, for example, the value $z$ in $g$.

1. Prove or disprove that $R$ is a metric.
2. Run the *control* algorithm and compare the output to *control_2*.
3. Run $k - means$ on *candidate* using *main* as the centroids. Because the galaxy type must be identical, you should first partition *candidate* into three blocks that represents galaxy types 0,1,2 and similarly for *main*. For each block in *candidate*, use the corresponding block in the partitioned *main* as the centroids.
4. Perform data analysis on the features of *main* and *candidate*.

# Entropy

**Entropy** Let $X$ be a finite *r.v.* on some a probability space. The mean information content of $X$ is called entropy, written $H(X)$, and is found by

$$H(X) \quad = \quad \Sigma_{x \in X} \; P(X = x) \log \frac{1}{P(X = x)} \tag{3}$$

We will abbreviate Eq.3 by writing $p_x$ to mean $P(X = x)$. Further, we'll use $x$ in the integration assuming it's from from $X$.

**Joint Entropy** Let $X, Y$ be finite *r.v.*'s on some probability space. The mean information content of both $X, Y$ is call joint entropy, written $H(X, Y)$, and is found by

$$H(X, Y) \quad = \quad -\Sigma_y \Sigma_x \; p_{xy} \log p_{xy} \tag{4}$$

**Definition** Let $X, Y$ be finite *r.v.*'s on some probability space. The mean conditional information content of $Y$ given $X$ is called conditional entropy, written $H(Y|X)$, and is found by

$$H(Y|X) \quad = \quad \Sigma_x \; p_x H(Y|X = x) = \Sigma_x \Sigma_y \; p_{xy} \log(\frac{p_x}{p_{xy}}) \tag{5}$$

**Definition** Let $X$ be a finite *r.v.* with two probability masses $P, Q$. The KL distance, written $d_{KL}(P||Q)$, is found by

$$d_{KL}(P||Q) \quad = \quad \Sigma_x \; p_x \log \frac{p_x}{q_x} \tag{6}$$

# 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.