TD: Mining Lab - Solutions

Mauro Sozio, Oana Balalau (TA)
firstname.lastname@telecom-paristech.fr

Exercise 1 (PageRank)

Compute the PageRank scores of the nodes in the graph G shown in Figure **??**, without introducing random jumps. First, write down the system of linear equations and solve it. Then, compute the PageRank scores using the PageRank algorithm and report the results for the first three iterations.

Solution:

1) We introduce a variable for each node: x_1, x_2, x_3, x_4.
The *in* and *out* links represent the relations between the variables:

$$\begin{cases} x_1 = x_4 \\ x_2 = x_1/2 \\ x_3 = x_1/2 + x_2 \\ x_4 = x_3 \\ x_1 + x_2 + x_3 + x_4 = 1 \end{cases} \iff \begin{cases} x_1 = x_4 \\ x_2 = x_1/2 \\ x_3 = x_1 \\ x_4 = x_3 \\ x_1 + x_2 + x_3 + x_4 = 1 \end{cases} \iff \begin{cases} x_1 = 2/7 \\ x_2 = 1/7 \\ x_3 = 2/7 \\ x_4 = 2/7 \end{cases}$$

2)
The first 3 iterations of PageRank:

$$\begin{cases} r^{k+1} = Mr^k \\ j \to i, \text{j has k succ} \implies M_{ij} = 1/k \end{cases}$$

We build the M matrix:

$$M = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1/2 & 0 & 0 & 0 \\ 1/2 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \tag{1}$$

Then we compute the vectors:

$$r^1 = \begin{pmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{pmatrix} \quad r^2 = \begin{pmatrix} 1/4 \\ 1/8 \\ 3/8 \\ 1/4 \end{pmatrix} \quad r^3 = \begin{pmatrix} 1/4 \\ 1/8 \\ 1/4 \\ 3/8 \end{pmatrix} \quad r^4 = \begin{pmatrix} 3/8 \\ 1/8 \\ 1/4 \\ 1/4 \end{pmatrix} \tag{2}$$

Animal	Give Birth	Live in Water	**Class**
a_1 = Dolphin	Y	Y	M
a_2 = Platypus	N	Y	M
a_3 = Sardine	N	Y	NM
a_4 = Human	Y	N	M
a_5 = Pigeon	N	N	NM

Table 1: Data specifying for each animal whether it gives birth, lives in water and is mammal or non-mammal (class).

Exercise 2 (Classification)

Question 1 (Decision tree)

Consider the data given in Table 1. Construct a decision tree using the following rules: (i) at each step compute the gini index for every possible split considering all attributes and select the split with the best gini value; (ii) Stopping rule: when the gini value of a node is zero or no further split is possible; and (iii) the class of a leaf node is determined by a majority rule.

After having built the decision tree, determine whether the tree can be pruned so as to improve the number of generalization errors. Recall that the training errors are given by the number of records that are misclassified by the decision tree, while the generalization errors are equal to the number of training errors plus 0.5 times number of leaves in the whole current tree. Each subtree should be pruned (i.e. replaced by a leaf node) until no further improvement is possible. Each leaf node is then labeled by a majority rule. Show all the steps. Predict the class value of the record: (a_6, GB='N',LW='N') with the decision tree minimizing the generalization errors.

Solution:

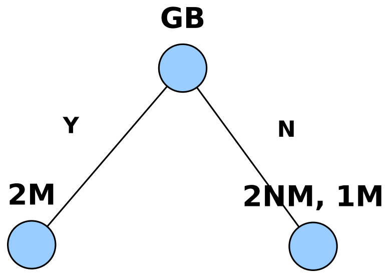

Figure 1: $G_Y = 1 - 1 = 0$
$G_N = 1 - (2/3)^2 - (1/3)^2 = 4/9$
$G_{split} = 2/5{*}0{+}3/5{*}4/9 = 4/15$

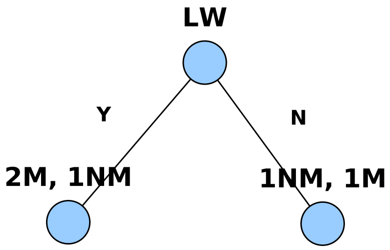

Figure 2: $G_Y = 1 - (2/3)^2 - (1/3)^2 = 4/9$
$G_N = 1 - (1/2)^2 - (1/2)^2 = 1/2$
$G_{split} = 3/5 * 4/9 + 2/5 * 1/2 = 7/15$

We will choose for the first split the attribute Give Birth. We write the tree and then we try to improve the **generalization error** by pruning it.

The tree that has the minimum generalization error is in Figure 4.

The class of (a_6, GB='N',LW='N') is NM.

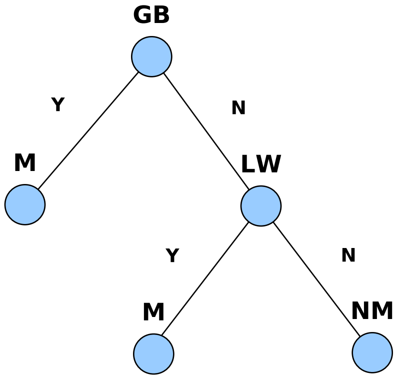

Figure 3: T.E. = 1, G.E. = 1 + 1.5 = 2.5

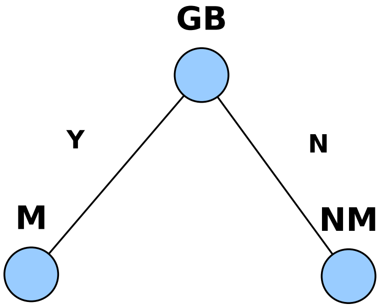

Figure 4: T.E. = 1, G.E. = 1 + 1 = 2

Question 2 (Naive Bayes)

Consider the data given in Table 1. Using the Naive Bayes classifier we learned during our class, predict the class value of the record: (a_6, GB='N',LW='N'). Show all the steps.

Solution:

$$\begin{aligned} P(M|GB=N,LW=N) &= \frac{P(GB=N,LW=N|M)P(M)}{P(GB=N,LW=N)} \\ &= \frac{P(GB=N|M)P(LW=N|M)P(M)}{P(GB=N,LW=N)} \\ &= \frac{1/3 * 1/3 * 3/5}{1/5} = 1/3 \end{aligned} \tag{3}$$

$$\begin{aligned} P(NM|GB=N,LW=N) &= \frac{P(GB=N,LW=N|NM)P(NM)}{P(GB=N,LW=N)} \\ &= \frac{P(GB=N|NM)P(LW=N|NM)P(NM)}{P(GB=N,LW=N)} \\ &= \frac{1 * 1/2 * 2/5}{1/5} = 1 \end{aligned} \tag{4}$$

The class of (a_6, GB='N',LW='N') will be NM.

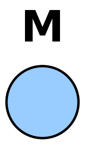

Figure 5: T.E. = 2, G.E. = 2 + 0.5 = 2.5

Question 3 (k-Nearest Neighbors)

Consider the data given in Table 1. Using the kNN classifier with $k = 3$, assuming features are binary variables $\{0, 1\}$ and using the L_2-norm, predict the class value of the record: (a_6, GB='Y',LW='Y').

Solution

Suppose 'N' is 0 and 'Y' is 1. Which are the closest 3 points to (0,0)?

The closest point is a_1 and at equal distance we have the points a_2, a_3, a_4. We notice that no matter which 3 points we select, the class of a_6 will always be M.

Exercise 3 (Clustering)

Question 1 (k-means)

Use the k-means algorithm and Euclidean distance to cluster the following 8 points into 3 clusters: A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9).

Suppose that the initial seeds (centers of each cluster) are A1, A4 and A7. Run the k-means algorithm for 1 step only. Draw two 10 by 10 space with all the 8 points and show the clusters and the centroids before and after the first step.

How many more iterations are needed to converge? Draw the result for each step.

Solution:

We will not show the drawings as they are straightforward.

Let d be the Euclidean distance and s_1, s_2, s_3 the seeds. We will compute for each node the distance from it to each of the seeds and we will assign the node to the closest cluster.

We take A1 as example.
$d(A1, s_1) = 0$
$d(A1, s_2) = \sqrt{13}$
$d(A1, s_3) = \sqrt{65}$
A1 will belong to cluster 1.

After computing for each node, we will have:
Cluster 1: $\{A1\}$, Cluster 2: $\{A3, A4, A5, A6, A8\}$, Cluster 3: $\{A2, A7\}$

The centers of the new clusters:
$s_1 = (2, 10)$
$s_2 = ((8 + 5 + 7 + 6 + 4)/5, (4 + 8 + 5 + 4 + 9)/5) = (6, 6)$
$s_3 = ((2 + 1)/2, (5 + 2)/2) = (1.5, 3.5)$

After the 3rd iteration the clusters will not change.

Question 2 (Hierarchical clustering)

Construct the dendrogram of the following hierarchical clustering algorithm applied on the previous 8 points: (i) start with each point in its own cluster; (ii) iterate: merge the clusters of the two closest points that are in different clusters; and (iii) Stopping rule: stop when the distance between the two closest points that are in different clusters is larger than ϵ. We set $\epsilon = \sqrt{10}$.

Solution:
We show the clusters at each step, the dendrogram is easy to construct after:
$\{A1\}, \{A2\}, \{A3\}, \{A4\}, \{A5\}, \{A6\}, \{A7\}, \{A8\}$
$\{A1\}, \{A2\}, \{A3, A5\}, \{A4\}, \{A6\}, \{A7\}, \{A8\}$
$\{A1\}, \{A2\}, \{A3, A5, A6\}, \{A4\}, \{A7\}, \{A8\}$
$\{A1\}, \{A2\}, \{A3, A5, A6\}, \{A4, A8\}, \{A7\}$
$\{A1, A4, A8\}, \{A2\}, \{A3, A5, A6\}, \{A7\}$
$\{A1, A4, A8\}, \{A2, A7\}, \{A3, A5, A6\}$