SD201: Final Exam

Mauro Sozio
Oana Balalau (TA), Maximilien Danisch (TA)
J.B. Griesner (TA), Raphael Charbey (TA)

Question 1: Decision tree, 3/20 points

Animal	Give Birth	Live in Water	**Class**
whale	Y	Y	M
dolphin	Y	Y	M
orca	Y	Y	M
human	Y	N	M
boa snake	Y	N	NM
Surinam toad	Y	N	NM
owl	N	N	NM
platypus	N	N	M
salmon	N	Y	NM
eel	N	Y	NM

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

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 the majority rule (breaking ties arbitrarily).

After having built the decision tree, determine whether the tree can be pruned so as to improve the generalization error. Recall that the training errors are given by the number of records that are misclassified by the decision tree, while the generalization error is equal to the number of training errors plus 0.5 times the number of leaves. 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 the majority rule (breaking ties arbitrarily). Show all the steps. Predict the class value of the record: ('seal', GB='Y',LW='Y') using the decision tree which minimizes the generalization error.

Question 2: Naive Bayes, 2/20 points

Animal	Give Birth	Live in Water	**Class**
whale	Y	Y	M
echidna	N	N	M
human	Y	N	M
boa snake	Y	N	NM
owl	N	N	NM
salmon	N	Y	NM
boa snake	Y	N	NM

Table 2: Data specifying for each animal whether it gives birth, lives in water and is mammal or non-mammal (class).

Consider the data given in Table 2. Using the Naive Bayes classifier, predict the class value of the record: ('salmon', GB='N',LW='Y'). Show all the steps.

Question 3: K-means, 3/20 points

Exercise 3 (Clustering) We are given the following eight points in the 2-dimensional euclidean space. $P_1 = (0,1), P2 = (1,2), P3 = (2,1), P4 = (1,0), P5 = (3,1), P6 = (4,2), P7 = (5,1), P8 = (4,0)$. Suppose that $P3 = (2,1)$ and $P6 = (4,2)$ are chosen as initial centroids for the K-means algorithm (K=2). Show step by step the clustering you would obtain by running K-Means on the previous set of points, while specifying for each clustering the current set of centroids. Recall that the algorithm terminates when the current set of centroids does not change. Draw the result for each step.

	a	b	c	**class**
1	0	1	1	0
2	0	1	1	1
3	0	0	0	0
4	0	0	0	0
5	1	1	0	1
6	1	1	0	1
7	1	1	0	1
8	1	1	1	0
9	1	1	1	0

Table 3: Binary vectors for a binary classification problem.

Question 4: PageRank, 3/20 points

Compute the PageRank vector (without introducing random jumps) of the directed graph G depicted in Figure 1. You can use any of the methods we considered in our course. Argue by using any argument from Markov Chain theory that the stationary distribution of the Markov chain corresponding to G exists and it is unique and can be computed by the power iteration method regardless of the initial distribution.

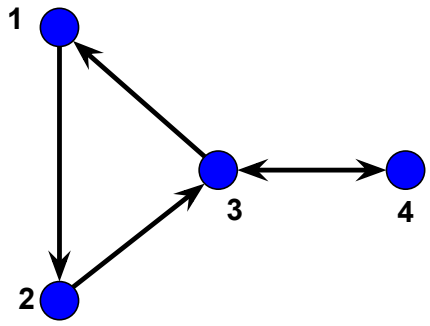

Figure 1: A directed graph G. Observe that the edge $(3, 4)$ is bidirectional.

Question 5 (K-means++, 3/20 points) Consider the deterministic variant of the $k-$means++ algorithm where the set of initial centroids are selected in the following way. Let $\mathcal{X}$ be a set of points in $\mathbb{R}^d$ given in input. The first centroid is chosen arbitrarily in $\mathcal{X}$. Then at step t, given the set of centroids $C_t = c_1, \ldots, c_t$, $(1 \leq t \leq k-1)$ selected up to step t, we select the point at maximum distance from the centroids in C_t. Formally, we choose the point $p \in \mathcal{X}$ such that $\min_{c \in C_t} d(c, p)$ is maximum (if there are multiple choices we pick one of them arbitrarily). Give an example where such a variant of k-means++ does not perform well, i.e. it computes a set of centroids with SSE much larger than the optimum solution. Compare the performance of such a variant of k-means++ with the original k-means++ algorithm.

Question 6 (densest subgraphs, 3/20 points) Let $G = (V, E)$ be an undirected graph. Let $H_1 = (V_1, E_1)$ and $H_2 = (V_2, E_2)$ be two densest subgraphs in G, i.e., for any subgraph $H = (V_H, E_H)$ of G it holds that $\frac{|E(H)|}{|V(H)|} \leq \frac{|E_i|}{|V_i|}, i = 1, 2$. Let $\widehat{H} = (V_1 \cap V_2, E_1 \cap E_2)$ be the graph obtained by the intersection of H_1 and H_2. What can we say about the density of $\widehat{H}$?

Question 7: MDL, 3/20 points

Given the data shown in Table 3, build a tree with minimum description length (MDL). For simplicity, you can assume that the cost of the tree is given by 1/2 the number of nodes in the tree (including nodes that are not leaves) plus the number of records that are not classified correctly by the decision tree. The class of a leaf node is determined by the majority rule (breaking ties arbitrarily). Prove that the tree you built has minimum description length among all possible decision trees for the input data.