Give formulas for estimating a text retrieval system's precision, recall, the $F_1$ measure [2 points], and the mean average precision (MAP) [1 point] given the following lists of results for queries $q_1$, and $q_2$, where `R` is a relevant result, and `N` is a non-relevant result:

- Results for $q_1$ : `RNNRR` (10 relevant results for $q_1$ exist in the collection.)
- Results for $q_2$ : `NRRNNN` (3 relevant results for $q_2$ exist in the collection.)

Three judges $j_1, j_2$, and $j_3$ were appointed to assess the relevance of ten documents. Estimate the Cohen's kappa $\kappa$ given the following lists of the judges' relevance judgements: [2 points]

- Relevance judgements of $j_1$ : `NNNRNRRNNR`
- Relevance judgements of $j_2$ : `NRNRRRRNNN`
- Relevance judgements of $j_3$ : `RNRNRRNRNR`

Are the judges in agreement according to your estimate of $\kappa$?

$P_{q_1} = \frac{3}{5}$, $P_{q_2} = \frac{2}{6}$, $R_{q_1} = \frac{3}{10}$, $R_{q_2} = \frac{2}{3}$

Macro-averaging P, R

$$P = (P_{q_1} + P_{q_2})/2 = \frac{14}{30} = 0{,}4\overline{6}$$

$$R = (R_{q_1} + R_{q_2})/2 = \frac{29}{60} = 0{,}48\overline{3}$$

Macro-averaging $F_1$

$$F = \frac{2 \cdot R \cdot P}{R + P} = \frac{251}{1005} \doteq 0{,}5482$$

Micro-averaging $F_1$

$$F_{q_1} = \frac{2 \cdot R_{q_1} \cdot P_{q_1}}{R_{q_1} + P_{q_1}} = \frac{2}{5} = 0{,}4$$

$$F_{q_2} = \frac{2 \cdot R_{q_2} \cdot P_{q_2}}{R_{q_2} + P_{q_2}} = \frac{4}{9} = 0{,}\overline{4}$$

$$F = (F_{q_1} + F_{q_2})/2 = \frac{190}{450} = 0{,}4\overline{2}$$

Micro-averaging P, R

$$P = \frac{3+2}{6+6} = \frac{5}{11} = 0{,}\overline{45}$$

$$R = \frac{3+2}{10+3} = \frac{5}{13} \doteq 0{,}3846$$

Macro-averaging $F_1$

$$F = \frac{2 \cdot R \cdot P}{R + P} = \frac{5}{12} = 0{,}41\overline{6}$$

Micro-averaging $F_1$

$$F_{q_1} = \frac{2 \cdot R_{q_1} \cdot P_{q_1}}{R_{q_1} + P_{q_1}} = \frac{2}{5} = 0{,}4$$

$$F_{q_2} = \frac{2 \cdot R_{q_2} \cdot P_{q_2}}{R_{q_2} + P_{q_2}} = \frac{4}{9} = 0{,}\overline{4}$$

$$F = (F_{q_1} + F_{q_2})/2 = \frac{190}{450} = 0{,}4\overline{2}$$

$$MAP = \frac{1}{2}\left(\frac{1}{10}\left(\frac{1}{1} + \frac{2}{4} + \frac{3}{5}\right) + \frac{1}{3}\left(\frac{1}{2} + \frac{2}{3}\right)\right)$$

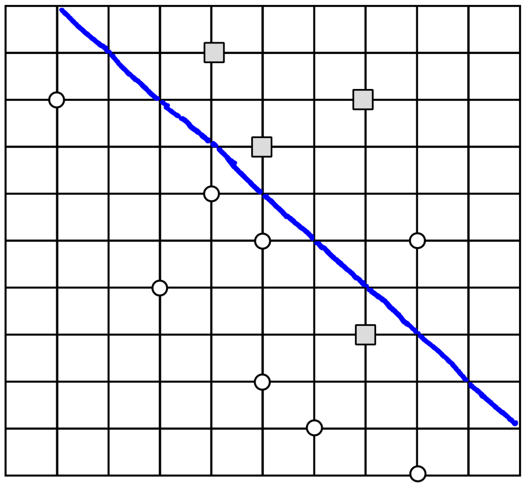
$P(R) = \frac{15}{30} = \frac{1}{2}$ $\quad P(N) = \frac{15}{30} = \frac{1}{2}$ $\quad P(E) = P(R)^2 + P(N)^2 = \frac{1}{2}$

$P(A) = \frac{2}{10} = \frac{1}{5}$ $\quad \kappa = \frac{P(A) - P(E)}{1 - P(E)} = \frac{-\frac{3}{10}}{\frac{5}{10}} = -\frac{3}{5} = -0{,}6$

No, the judges are not in agreement.

Write your solution only on this side of the sheet!

You are given the following training dataset containing observations of two classes (○ and □):



Is your dataset linearly separable? [2 points] Construct a linear classifier using *the two closest points from different classes*, and draw the separating hyperplane. [2 points] What is the training error (the ratio of incorrectly classified observations and all observations) of your linear classifier? [1 point]

No, there is no hyperplane that separates the two classes.

The training error is $\frac{2}{12} = 0{,}08\overline{3}$.

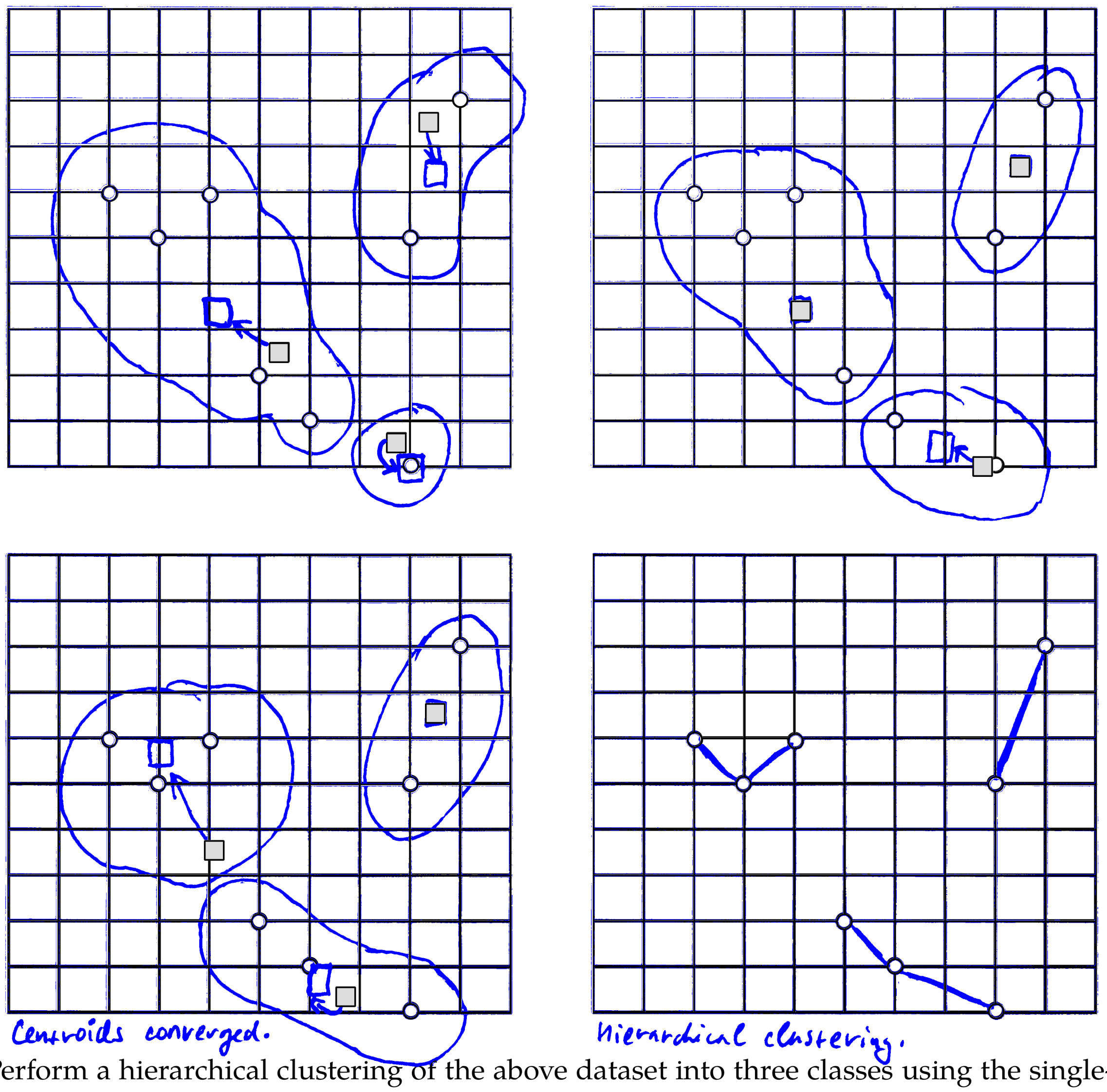
0007 sheet 3 učo points

Explain the following aspects of the $K$-means flat clustering algorithm [2 points]:

1. What do we need to decide before using the algorithm?
2. What is the input and the output of the algorithm?
3. What are the two steps that take place in every epoch?
4. How do we decide when to stop the algorithm?

1. We need to know the number of classes K and initial mean estimates (seeds). 2. The input are unclassified points and K seeds. 3. Reassigning points, recomputing centroids. 4. Centroids converged.

Given the points ○, and the seeds □, run the $K$-means algorithm for three epochs. Draw the state of the algorithm at the beginning and after every epoch; no computation should be necessary. What is the output of the algorithm? [2 points]



Centroids converged.

hierarchical clustering.

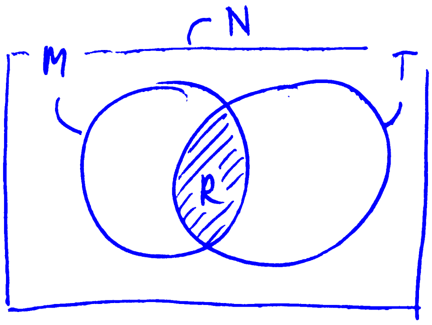
Perform a hierarchical clustering of the above dataset into three classes using the single-link hierarchical agglomerative clustering algorithm. Is the output the same as the output of the $K$-means flat hierarchical clustering algorithm above? [1 point]

Yes, it is.

0007 sheet 4 učo points

You maintain a text retrieval system. Let $E_1$ denote the complete set of documents in the index of your system and let $E_2$ denote the complete set of documents in the index of a competing system. Suppose the indices of both systems are independent uniform random samples without replacement from the World Wide Web $N$. The size of $E_1$ is $|E_1| = 130$ trillion $(130 \cdot 10^{12})$ documents. You take a uniform random subsample of documents without replacement from $E_1$ and you submit each document to the competing system. This gives you an estimate $x = 0.2$ of the conditional probability $P(d \in E_2 \mid d \in E_1), d \in N$. You repeat the same procedure with $E_2$, obtaining an estimate $y = 0.4$ of the conditional probability $P(d \in E_1 \mid d \in E_2), d \in N$. Assume the estimates $x, y$ are the true probabilities. What is the size $|E_2|$ of the competing system's index? [3 points]

The grey parrot, native to equatorial Africa, is categorized as an endangered species by the International Union for Conservation of Nature (IUCN). Suppose you take a uniform random sample $M$ without replacement of size $|M| = 8\,000$ from the grey parrot population $N$ and mark the sampled animals. After returning the marked animals back into the population, you take a second independent uniform random sample $T$ without replacement of the same size $|T| = 8\,000$ from the population. The number of marked animals $R = M \cap T$ in the second sample is $|R| = 10$. What is the most likely size $|N|$ of the grey parrot population? [2 points]

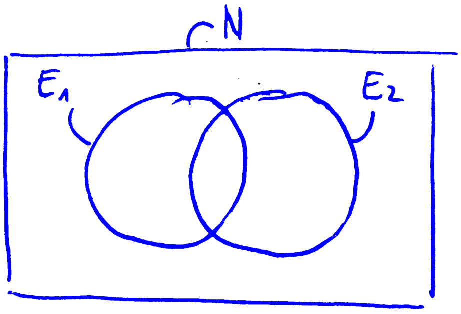



$$\forall x \in N: P(x \in M) = \frac{|M|}{|N|}, \quad P(x \in R) = \frac{|R|}{|N|}$$

$$P(x \in T) = \frac{|T|}{|N|}.$$

$$\forall x, y \in N: P(x \in R) = P(x \in M, y \in T) = P(x \in M) \cdot P(y \in T) = \frac{|M||T|}{|N|^2}.$$

$$\frac{|R|}{|N|} = \frac{|M||T|}{|N|^2} \iff |N| = \frac{|M||T|}{|R|} = \frac{8000^2}{10} = 6{,}4 \cdot 10^6.$$




$$\forall d \in N: P(d \in E_2 \mid d \in E_1) = 0{,}2,$$

$$P(d \in E_1 \mid d \in E_2) = 0{,}4$$

$$P(d \in E_1) = \frac{130 \cdot 10^{12}}{|N|}, \quad P(d \in E_2) = \frac{|E_2|}{|N|}.$$

$$0{,}2 = \frac{0{,}4 \cdot \frac{|E_2|}{\cancel{|N|}}}{\frac{130 \cdot 10^{12}}{\cancel{|N|}}} \iff |E_2| = \frac{0{,}2 \cdot 130 \cdot 10^{12}}{0{,}4} = 65 \cdot 10^{12}.$$