Introduction to Information Retrieval and Text Mining Assignment 5

Roman Klinger

Institute for Natural Language Processing, University of Stuttgart

2018-01-30

Overview

1. Task 1 (Feature Selection)
2. Task 2 (Perceptron)
3. Task 3 (HAC)
4. Task 4 (Evaluation of Clustering)
5. Task 5 (PageRank)

Outline

1 Task 1 (Feature Selection)

2 Task 2 (Perceptron)

3 Task 3 (HAC)

4 Task 4 (Evaluation of Clustering)

5 Task 5 (PageRank)

Task 1 (Feature Selection) 4 points

Given the following documents assigned with classes c_1 and c_2:

c_1	I drink coffee
c_1	I drink tee
c_2	I take aspirin
c_2	I take paracetamol

The occurrence of which of the words helps best to predict if a document belongs to class c_1 or c_2? Please argue based on *mutual information*.

Task 1 (Feature Selection) 4 points

Given the following documents assigned with classes c_1 and c_2:

c_1	I drink coffee
c_1	I drink tee
c_2	I take aspirin
c_2	I take paracetamol

The occurrence of which of the words helps best to predict if a document belongs to class c_1 or c_2? Please argue based on *mutual information*.

Intuition

- "drink" only occurs with c_1 and "take" only with c_2 $\Rightarrow$ perfect indicator

Task 1 (Feature Selection) 4 points

Given the following documents assigned with classes c_1 and c_2:

c_1	I drink coffee
c_1	I drink tee
c_2	I take aspirin
c_2	I take paracetamol

The occurrence of which of the words helps best to predict if a document belongs to class c_1 or c_2? Please argue based on *mutual information*.

Intuition

- "drink" only occurs with c_1 and "take" only with c_2 $\Rightarrow$ perfect indicator
- "coffee", "tee", "aspirin", "paracetamol" only occur each in one class, helpful as well.

Task 1 (Feature Selection) 4 points

Given the following documents assigned with classes c_1 and c_2:

c_1	I drink coffee
c_1	I drink tee
c_2	I take aspirin
c_2	I take paracetamol

The occurrence of which of the words helps best to predict if a document belongs to class c_1 or c_2? Please argue based on *mutual information*.

Intuition

- "drink" only occurs with c_1 and "take" only with c_2 $\Rightarrow$ perfect indicator
- "coffee", "tee", "aspirin", "paracetamol" only occur each in one class, helpful as well.
- "I" equally distributed between classes$\Rightarrow$ not helpful

Task 1 (Feature Selection)

c_1 I drink coffee

c_1 I drink tee

c_2 I take aspirin

c_2 I take paracetamol

Task 1 (Feature Selection)

c_1	I drink coffee
c_1	I drink tee
c_2	I take aspirin
c_2	I take paracetamol

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right)$$

Task 1 (Feature Selection)

c_1 I drink coffee
c_1 I drink tee
c_2 I take aspirin
c_2 I take paracetamol

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right)$$

MI

$$\begin{aligned} I(drink; Y) &= p(\text{drink}, c_1) \log \left(\frac{p(\text{drink}, c_1)}{p(\text{drink})\,p(c_1)} \right) \\ &+ p(\text{drink}, c_2) \log \left(\frac{p(\text{drink}, c_2)}{p(\text{drink})\,p(c_2)} \right) \\ &+ p(\neg\text{drink}, c_1) \log \left(\frac{p(\neg\text{drink}, c_1)}{p(\neg\text{drink})\,p(c_1)} \right) \\ &+ p(\neg\text{drink}, c_2) \log \left(\frac{p(\neg\text{drink}, c_2)}{p(\neg\text{drink})\,p(c_2)} \right) \\ &= \frac{1}{2} \log \frac{\frac{1}{2}}{\frac{1}{2}\frac{1}{2}} + 0 + 0 + \frac{1}{2} \log \frac{\frac{1}{2}}{\frac{1}{2}\frac{1}{2}} \\ &= 1 \end{aligned}$$

Task 1 (Feature Selection)

c_1 I drink coffee
c_1 I drink tee
c_2 I take aspirin
c_2 I take paracetamol

$$I(X;Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\left(\frac{p(x,y)}{p(x)\,p(y)}\right)$$

MI

$$\begin{aligned} I(drink;Y) &= p(\text{drink},c_1)\log\left(\frac{p(\text{drink},c_1)}{p(\text{drink})\,p(c_1)}\right)\\ &+ p(\text{drink},c_2)\log\left(\frac{p(\text{drink},c_2)}{p(\text{drink})\,p(c_2)}\right)\\ &+ p(\neg\text{drink},c_1)\log\left(\frac{p(\neg\text{drink},c_1)}{p(\neg\text{drink})\,p(c_1)}\right)\\ &+ p(\neg\text{drink},c_2)\log\left(\frac{p(\neg\text{drink},c_2)}{p(\neg\text{drink})\,p(c_2)}\right)\\ &= \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{1}{2}\frac{1}{2}} + 0 + 0 + \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{1}{2}\frac{1}{2}}\\ &= 1 \end{aligned}$$

$$= I(take;Y)$$

Task 1 (Feature Selection)

c_1 I drink coffee
c_1 I drink tee
c_2 I take aspirin
c_2 I take paracetamol

Task 1 (Feature Selection)

c_1	I drink coffee
c_1	I drink tee
c_2	I take aspirin
c_2	I take paracetamol

$$I(X;Y) =$$

$$\sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right)$$

Task 1 (Feature Selection)

c_1	I drink coffee
c_1	I drink tee
c_2	I take aspirin
c_2	I take paracetamol

$$I(X;Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\left(\frac{p(x,y)}{p(x)\,p(y)}\right)$$

MI

$$\begin{aligned} I(I;Y) &= p(\mathrm{I},c_1)\log\left(\frac{p(\mathrm{I},c_1)}{p(\mathrm{I})\,p(c_1)}\right)\\ &+ p(\mathrm{I},c_2)\log\left(\frac{p(\mathrm{I},c_2)}{p(\mathrm{I})\,p(c_2)}\right)\\ &+ p(\neg\mathrm{I},c_1)\log\left(\frac{p(\neg\mathrm{I},c_1)}{p(\neg\mathrm{I})\,p(c_1)}\right)\\ &+ p(\neg\mathrm{I},c_2)\log\left(\frac{p(\neg\mathrm{I},c_2)}{p(\neg\mathrm{I})\,p(c_2)}\right)\\ &= \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{1}{1}\frac{1}{2}} + \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{1}{1}\frac{1}{2}} + 0 + 0\\ &= 0 \end{aligned}$$

Task 1 (Feature Selection)

c_1 I drink coffee

c_1 I drink tee

c_2 I take aspirin

c_2 I take paracetamol

Task 1 (Feature Selection)

c_1 I drink coffee

c_1 I drink tee

c_2 I take aspirin

c_2 I take paracetamol

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right)$$

Task 1 (Feature Selection)

c_1 I drink coffee

c_1 I drink tee

c_2 I take aspirin

c_2 I take paracetamol

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)\, p(y)} \right)$$

MI

$$\begin{aligned}
I(coffee; Y) &= p(\text{coffee}, c_1) \log \left(\frac{p(\text{coffee}, c_1)}{p(\text{coffee})\, p(c_1)} \right) \\
&\quad + p(\text{coffee}, c_2) \log \left(\frac{p(\text{coffee}, c_2)}{p(\text{coffee})\, p(c_2)} \right) \\
&\quad + p(\neg\text{coffee}, c_1) \log \left(\frac{p(\neg\text{coffee}, c_1)}{p(\neg\text{coffee})\, p(c_1)} \right) \\
&\quad + p(\neg\text{coffee}, c_2) \log \left(\frac{p(\neg\text{coffee}, c_2)}{p(\neg\text{coffee})\, p(c_2)} \right) \\
&= \frac{1}{4} \log \frac{\frac{1}{4}}{\frac{1}{4}\frac{1}{2}} + 0 + \frac{1}{4} \log \frac{\frac{1}{4}}{\frac{3}{4}\frac{1}{2}} + \frac{1}{2} \log \frac{\frac{1}{2}}{\frac{3}{4}\frac{1}{2}} \\
&= 0.25 + 0 - 0.15 + 0.2 \approx 0.3
\end{aligned}$$

Outline

1 Task 1 (Feature Selection)

2 Task 2 (Perceptron)

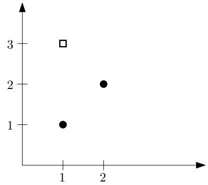
3 Task 3 (HAC)

4 Task 4 (Evaluation of Clustering)

5 Task 5 (PageRank)

Task 2 (Perceptron)

Given are these instances:

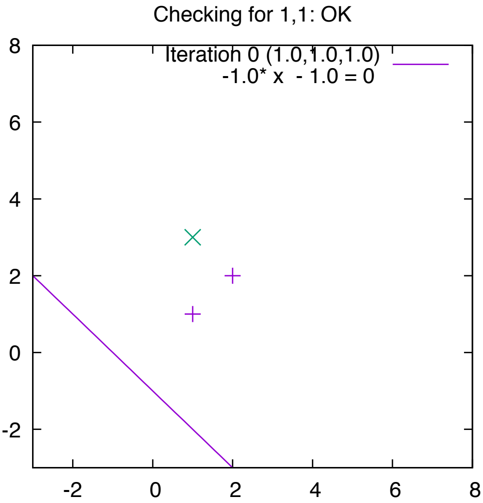


Interpret $-\theta$ as an additional feature weight in the weight vector which always has the feature value 1. Then, the instances are interpreted as vectors $(1,1,1)$, $(1,2,2)$ and $(1,1,3)$.

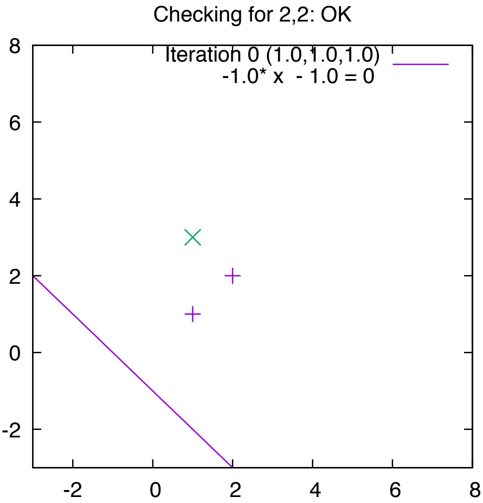
$$\begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix}^T \cdot \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{pmatrix} = \theta \iff \begin{pmatrix} -\theta \\ w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix}^T \cdot \begin{pmatrix} 1 \\ f_1 \\ f_2 \\ \vdots \\ f_n \end{pmatrix} = 0$$

Perform (at least) three iterations of perceptron learning, starting with the vector $(1,1,1)^T = \vec{w}$. In each iteration, all instances are processed, that therefore leads to (at least) 9 weight vector updates.

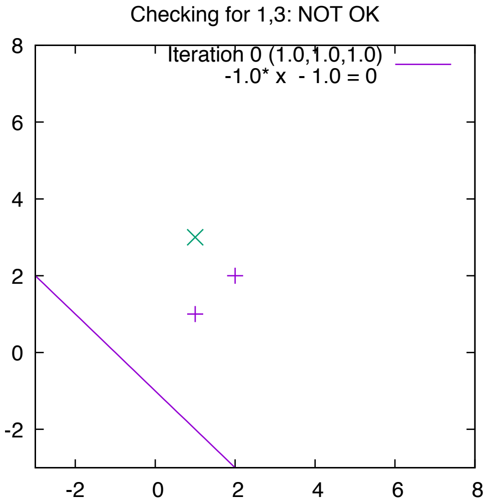
Task 2 (Perceptron)



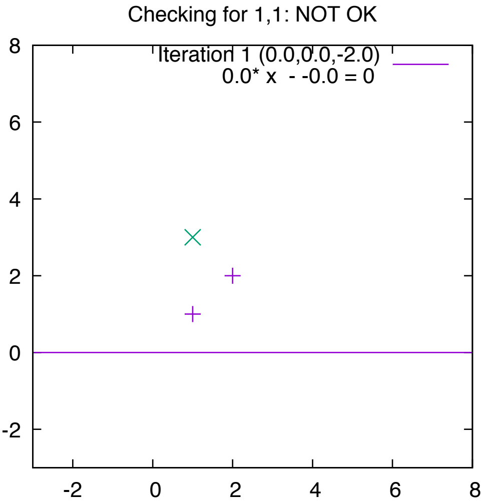
Task 2 (Perceptron)



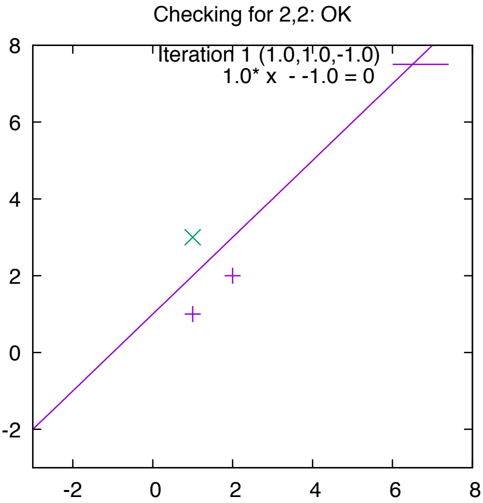
Task 2 (Perceptron)



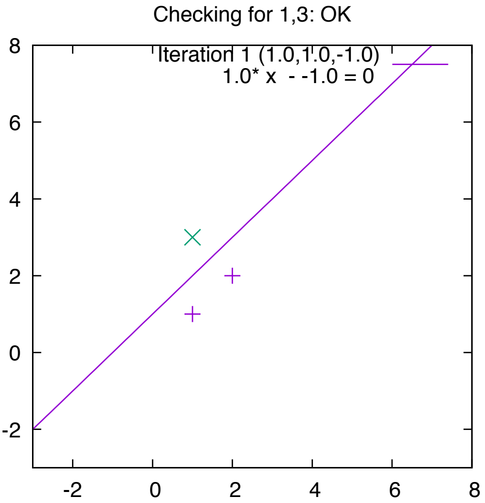
Task 2 (Perceptron)



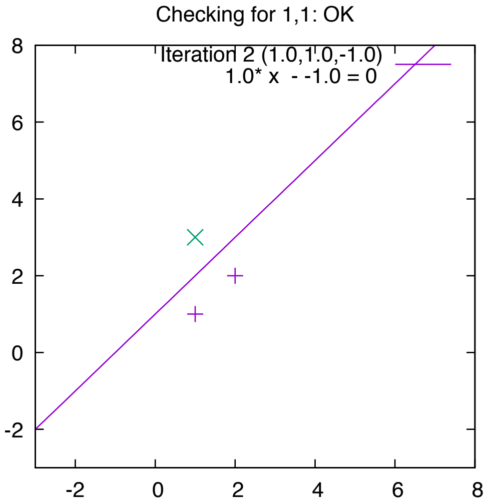
Task 2 (Perceptron)



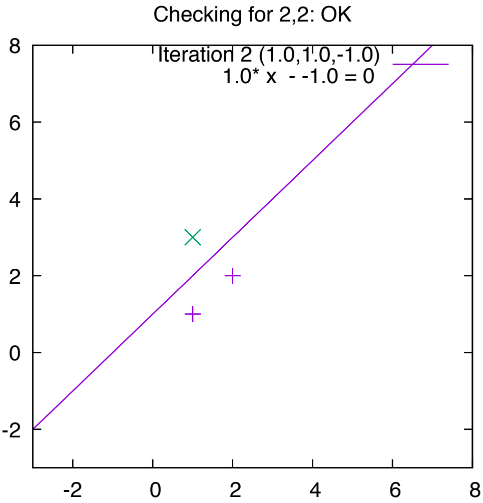
Task 2 (Perceptron)



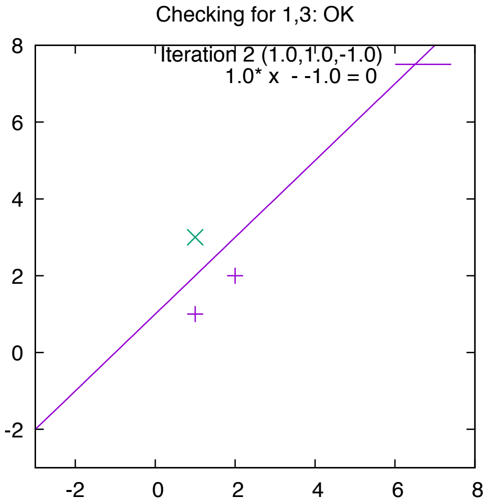
Task 2 (Perceptron)



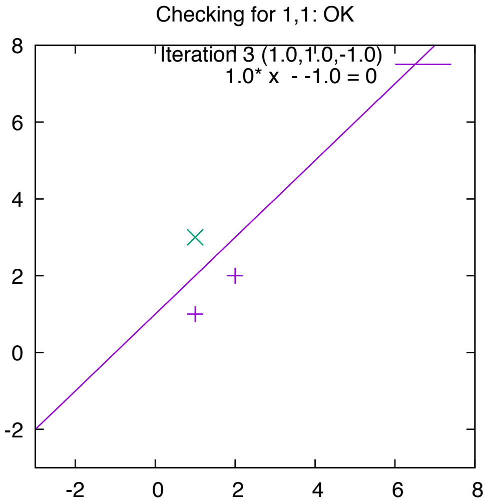
Task 2 (Perceptron)



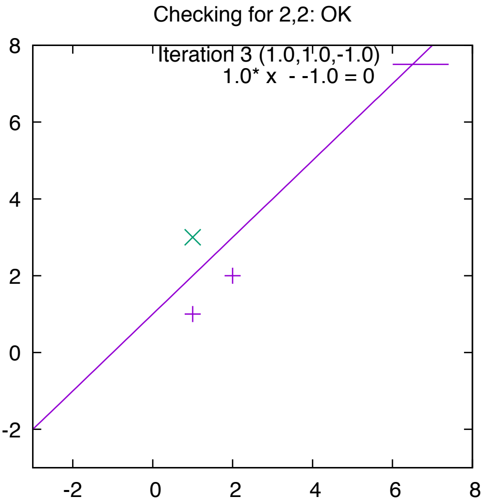
Task 2 (Perceptron)



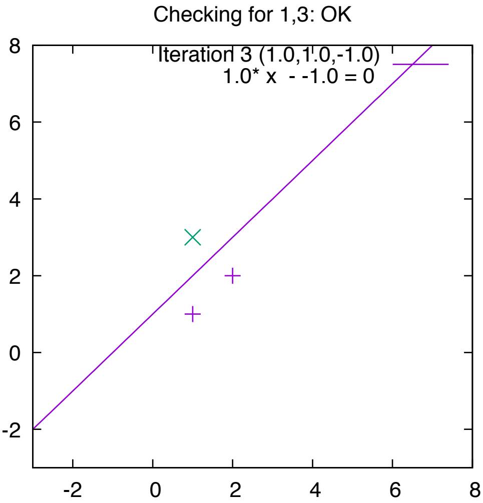
Task 2 (Perceptron)



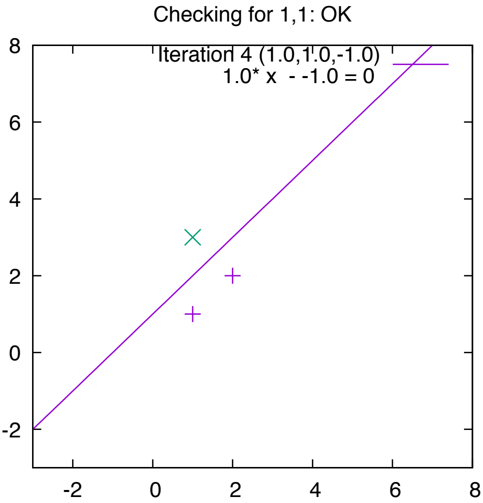
Task 2 (Perceptron)



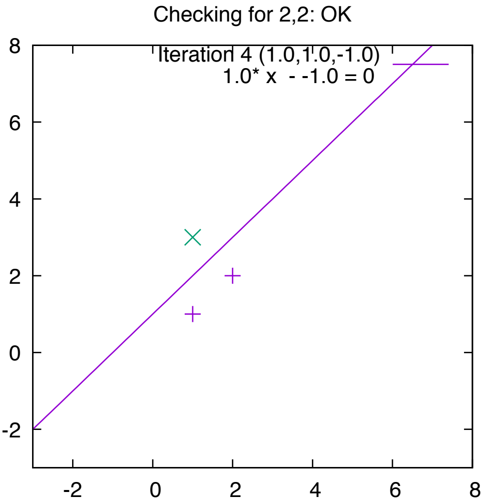
Task 2 (Perceptron)



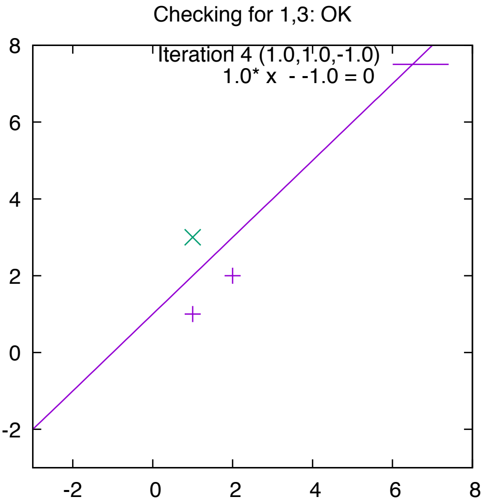
Task 2 (Perceptron)



Task 2 (Perceptron)



Task 2 (Perceptron)

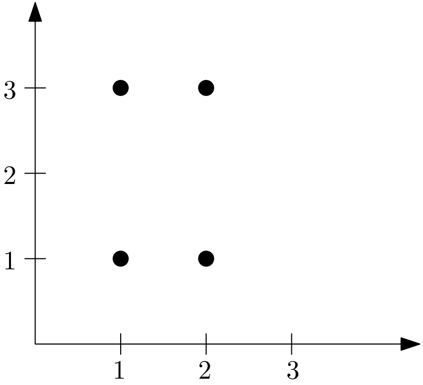


Outline

1. Task 1 (Feature Selection)
2. Task 2 (Perceptron)
3. Task 3 (HAC)
4. Task 4 (Evaluation of Clustering)
5. Task 5 (PageRank)

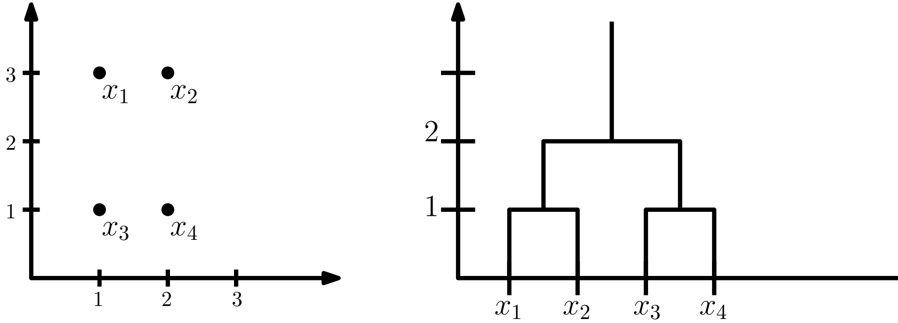
Task 3 (HAC)

Perform hierarchical agglomerative clustering of these instances with single link and with complete link clustering. Is the result the same? Is there a difference?



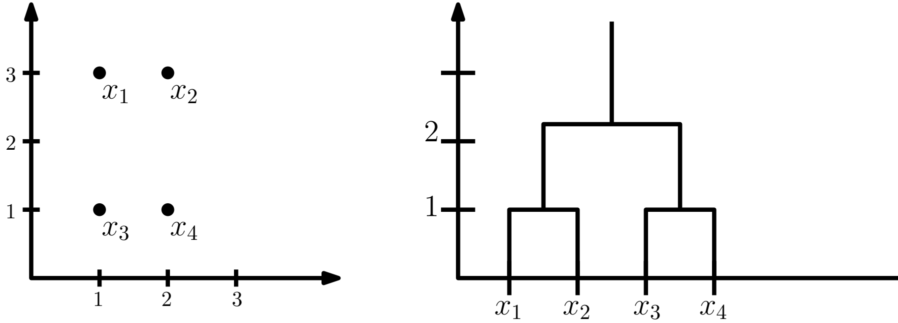
Task 3 (HAC)

Single Link



Task 3 (HAC)

Complete Link

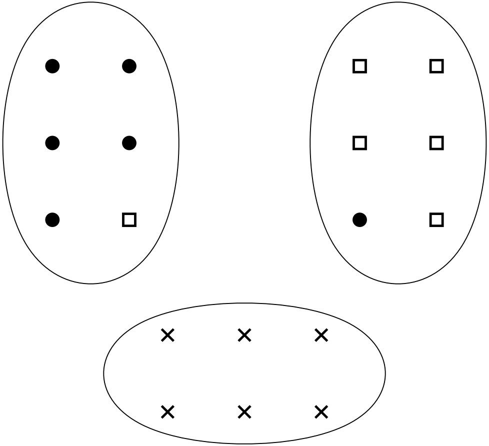


Outline

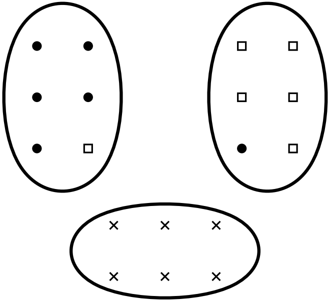
1. Task 1 (Feature Selection)
2. Task 2 (Perceptron)
3. Task 3 (HAC)
4. Task 4 (Evaluation of Clustering)
5. Task 5 (PageRank)

Task 4 (Evaluation of Clustering)

What is the rand index of the following clustering, assuming that cross, circle and box are gold annotations of classes?



Task 4 (Evaluation of Clustering)



- $\mathrm{TP} = \binom{5}{2} + \binom{5}{2} + \binom{6}{2} = 10 + 10 + 15 = 35$
- $\mathrm{FP} = 5 + 5 + 0 = 10$
- $\mathrm{FN} = 5 + 5 = 10$
- $\mathrm{TN} = \binom{18}{2} - \mathrm{TP} - \mathrm{FP} - \mathrm{FN} = 153 - 35 - 10 - 10 = 98$
- Rand index $= \frac{35+98}{153} \approx 0.87$

Outline

1 Task 1 (Feature Selection)

2 Task 2 (Perceptron)

3 Task 3 (HAC)

4 Task 4 (Evaluation of Clustering)

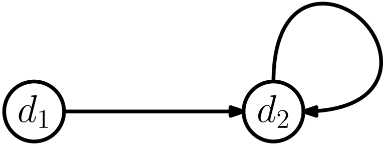
5 Task 5 (PageRank)

Task 5 (PageRank)

What is the page rank value for two documents d_1 and d_2 in which exactly one link in d_1 points to d_2 and one link points from d_2 to itself?

Task 5 (PageRank)

What is the page rank value for two documents d_1 and d_2 in which exactly one link in d_1 points to d_2 and one link points from d_2 to itself?

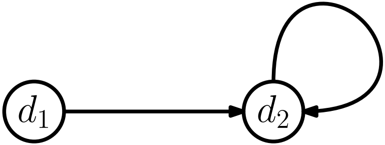


- Link matrix:

$$\begin{matrix} 0 & 1 \\ 0 & 1 \end{matrix}$$

Task 5 (PageRank)

What is the page rank value for two documents d_1 and d_2 in which exactly one link in d_1 points to d_2 and one link points from d_2 to itself?



- Link matrix:

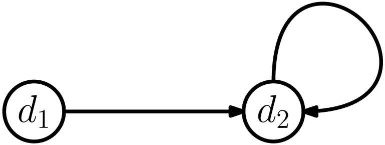
$$\begin{matrix} 0 & 1 \\ 0 & 1 \end{matrix}$$

- Probability transition matrix:

$$\begin{matrix} 0 & 1 \\ 0 & 1 \end{matrix}$$

Task 5 (PageRank)

What is the page rank value for two documents d_1 and d_2 in which exactly one link in d_1 points to d_2 and one link points from d_2 to itself?



- Link matrix:

$$\begin{matrix} 0 & 1 \\ 0 & 1 \end{matrix}$$

- Probability transition matrix:

$$\begin{matrix} 0 & 1 \\ 0 & 1 \end{matrix}$$

- Dead ends/teleportation (with teleportation rate of 0.1):

$$\begin{matrix} 0.05 & 0.95 \\ 0.05 & 0.95 \end{matrix}$$

Task 5 Solution

		$p_{11} = 0.05$	$p_{12} = 0.95$
d_1	d_2	$p_{21} = 0.05$	$p_{22} = 0.95$
0.5	0.5		

Task 5 Solution

		$p_{11} = 0.05$	$p_{12} = 0.95$
d_1	d_2	$p_{21} = 0.05$	$p_{22} = 0.95$
0.5	0.5	0.05	0.95

Task 5 Solution

		$p_{11} = 0.05$	$p_{12} = 0.95$
d_1	d_2	$p_{21} = 0.05$	$p_{22} = 0.95$
0.5	0.5	0.05	0.95
0.05	0.95		

Task 5 Solution

		$p_{11} = 0.05$	$p_{12} = 0.95$
d_1	d_2	$p_{21} = 0.05$	$p_{22} = 0.95$
0.5	0.5	0.05	0.95
0.05	0.95	0.05	0.95

Task 5 Solution

		$p_{11} = 0.05$	$p_{12} = 0.95$
d_1	d_2	$p_{21} = 0.05$	$p_{22} = 0.95$
0.5	0.5	0.05	0.95
0.05	0.95	0.05	0.95

Introduction to Information Retrieval and Text Mining Assignment 5

Roman Klinger

Institute for Natural Language Processing, University of Stuttgart

2018-01-30