COMP 474/6741 Intelligent Systems (Winter 2024)

Worksheet #5: Introduction to Machine Learning

Task 1. A quick refresher: Based on the output below, compute precision@k $= \frac{1}{k} \cdot \sum_{c=1}^{k} \text{rel}(c)$ for the three recommender systems (for $k = 1, 2, 3$):

	System@k			precision@k		
	1	2	3	1	2	3
system 1	✗	✗	✓	0	0	0.33
system 2	✗	✓	✓	0	0.5	0.66
system 3	✓	✓	✓	1	1	1

1/1 [rel(1)] = 0

1/2 [rel(1) + rel(2)] = 0

1/3 [rel(1) + rel(2) + rel(3)] = 0.33

Task 2. Here is a dataset of documents with two attributes, to be grouped into two clusters. Apply k-Means clustering, by computing the *Euclidian distance* $d(\vec{p}, \vec{q}) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$ of each data point to the two initial centroïds and assigning each document to its closest cluster:

Centroids are located @ Cluster1 (1.0, 1.0)

	Centroïd	
	a1	a2
Cluster 1	1.0	1.0
Cluster 2	5.0	7.0

p vector of Doc q vector of cluster

	a1	**a2**	**Distance to C1**	**Distance to C2**	**Cluster**	**Tag**
Doc1	1.5	2.0	1.118	6.103	1	#Travel
Doc2	3.0	4.0	3.606	3.606	1	#Food
Doc3	4.5	5.0	5.315	2.062	2	#Travel
Doc4	3.5	4.5	4.301	2.915	2	#Food

(Ignore the "Tag" column for now, we'll use it in the next question!)

Task 3. Now apply the kNN classification algorithm on the new document below to determine its tag. Use $k = 3$ and the Euclidian distance $d(\vec{p}, \vec{q}) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$ (just like for k-Means-clustering):

	a1	**a2**	**d-Doc1**	**d-Doc2**	**d-Doc3**	**d-Doc4**	**Tag?**
Doc5	2.5	3.5	1.8028	0.7071	2.5	1.414	# Food

You can now auto-assign a *tag* to the new document based on a majority vote of the k nearest neighbors.

here p vector is Doc5
q vector is d-Doc1 or d-Doc2 or ... or etc.

k =3

3 closest = Doc1, 2, 4

Tags = Travel, Food, Food

Food = 2

Travel = 1

Task 4. Should we invest \$100m in producing this new movie? We'll use machine learning to predict the rating (1–5 stars) of a movie, by applying the *regression* version of the kNN algorithm. Here's our training data:

$\sqrt{(109-135)^2 + (5-0)^2 + (3-5)^2} = 26.55$

#	Movie	Length	#Zombies	#Explosions	Rating
1	Movie 1	135	0	5	★★★
2	Movie 2	90	123	2	★★★★★
3	Movie 3	159	2	1	★
4	Movie 4	109	5	3	

To find the predicted rating for Movie #4, first find the two nearest neighbors (i.e., $k = 2$), using the same calculation as before:

$d(\vec{m}_4, \vec{m}_1) =$ 26.552 $d(\vec{m}_4, \vec{m}_2) =$ 19.026 $d(\vec{m}_4, \vec{m}_3) =$ 50.13 $\Longrightarrow$ Closest = m1 , m2

Now, compute the *average* of the ratings of the k nearest movies for $k = 2$ (convert the ⋆ rating into a value in [1...5]): This is your predicted rating for Movie 4! (3 + 5) / 2 = 4 star rating for movie 4

Task 5. Here are three different systems that classified 500 data items:

	Target	*system 1*	*system 2*	*system 3*
	X1 √	X1 ×	X1 √	X1 √
	X2 √	X2 ×	X2 ×	X2 √
	X3 √	X3 ×	X3 √	X3 √
	X4 √	X4 ×	X4 √	X4 √
	X5 √	X5 ×	X5 ×	X5 √
	X6 ×	X6 ×	X6 ×	X6 √
	X7 ×	X7 ×	X7 ×	X7 √
	… ×	… ×	… ×	… ×
	… ×	… ×	… ×	… ×
	X500 ×	X500 ×	X500 ×	X500 ×

Last time, we already calculated *Precision* and *Recall* for these systems (you can verify that the alternative formulas here give you the same results). Now, compute the *Accuracy* and F_1-*Measure*:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

	system 1	system 2	system 3
Precision	n/a	1.0	0.71
Recall	0	0.6	1.0
Accuracy	0	0.6	1
F_1-Measure	0	0.75	0.83

Sys 1
Accuracy = $\frac{0+0}{5}$
$F_1\text{-m} = \frac{2 \cdot 0}{0+0} = 0$
P = $\frac{0}{0+0}$ n/a
R = $\frac{0}{0+5}$

Sys 2
Accuracy = $\frac{3+0}{5}$
$F_1\text{-M} = \frac{2 \cdot (1)(0.6)}{(1)+0.6} = 0.75$
P = $\frac{3}{3+0} = 1$
R = $\frac{3}{3+2} = 0.6$

Sys 3
A = $\frac{5+0}{5} = 1$
$F_{1\text{-m}} = \frac{2 \cdot (0.71) \cdot (1)}{0.71 + 1} = 0.83$
P = $\frac{5}{5+2}$
R = $\frac{5}{5+0}$