University of Southern Denmark
IMADA

DM566: Data Mining and Machine Learning

Spring term 2022

Exercise 5: Clustering: k-means and Silhouette, Evaluation of Classifiers, Probability

Exercise 5-1 Measure for Evaluation of Classifiers (1 point)

Given a data set with known class labels ($f(o)$) of the objects. In order to evaluate the quality of a classifier h, each object is additionally classified using h. The results are given in the table (all three columns) below.

ID	$f(o)$	$h(o)$
O_1	A	A
O_2	B	A
O_3	C	C
O_4	B	C
O_5	A	B

ID	$f(o)$	$h(o)$
O_6	B	B
O_7	C	A
O_8	A	A
O_9	A	A
O_{10}	B	C

ID	$f(o)$	$h(o)$
O_{11}	B	A
O_{12}	C	A
O_{13}	A	B
O_{14}	C	C
O_{15}	B	B

ID	$f(o)$	$h(o)$
O_{16}	C	B
O_{17}	A	A
O_{18}	B	B
O_{19}	C	C
O_{20}	A	B

1. Rewrite the definitions for precision and recall given in the lecture by using TP, TN, FP, and FN.

   Suggested solution:
   $f(o)$ is the actual class of o, $h(o)$ is the class predicted by classifier h.

   - True positives for class i: $TP_i = \{o | f(o) = i \wedge h(o) = i\}$
   - False positives for class i: $FP_i = \{o | f(o) \neq i \wedge h(o) = i\}$
   - False negatives for class i: $FN_i = \{o | f(o) = i \wedge h(o) \neq i\}$
   - (for the sake of completeness, although we do not use this in the computations below:) True negatives for class i: $TN_i = \{o | f(o) \neq i \wedge h(o) \neq i\}$

   Precision:

   $$\text{Precision}(h, i) = \frac{|\{o \in h_i | h(o) = f(o)\}|}{|h_i|}$$

   or

   $$\text{Precision}(h, i) = \frac{|TP_i|}{|TP_i| + |FP_i|}$$

   Recall:

   $$\text{Recall}(h, i) = \frac{|\{o \in f_i | h(o) = f(o)\}|}{|f_i|}$$

   or

   $$\text{Recall}(h, i) = \frac{|TP_i|}{|TP_i| + |FN_i|}$$

2. Using the tables above, compute precision and recall for each class.

Suggested solution:
Confusion matrix:

	A	B	C	$\|f_i\|$	$\|TP\|$	$\|FP\|$	$\|FN\|$
A	4	3	0	7	4	4	3
B	2	3	2	7	3	4	4
C	2	1	3	6	3	2	3
$\|h_i\|$	8	7	5				

Precision:

$$\text{Precision}(h, A) = \frac{4}{4+4} = \frac{1}{2}$$
$$\text{Precision}(h, B) = \frac{3}{3+4} = \frac{3}{7}$$
$$\text{Precision}(h, C) = \frac{3}{3+2} = \frac{3}{5}$$

Recall:

$$\text{Recall}(h, A) = \frac{4}{4+3} = \frac{4}{7}$$
$$\text{Recall}(h, B) = \frac{3}{3+4} = \frac{3}{7}$$
$$\text{Recall}(h, C) = \frac{3}{3+3} = \frac{1}{2}$$

3. To get a complete measure for the quality of the classification with respect to a single class, the F_1-measure (the harmonic mean of precision and recall) is commonly used. It is defined as follows:

$$F_1(h, i) = \frac{2 \cdot \text{Recall}(h, i) \cdot \text{Precision}(h, i)}{\text{Recall}(h, i) + \text{Precision}(h, i)}$$

Compute the F_1-measure for all classes.

Suggested solution:

$$F_1(h, A) = \frac{2 \cdot \frac{4}{7} \cdot \frac{1}{2}}{\frac{4}{7} + \frac{1}{2}} = \frac{8}{15}$$
$$F_1(h, B) = \frac{2 \cdot \frac{3}{7} \cdot \frac{3}{7}}{\frac{3}{7} + \frac{3}{7}} = \frac{3}{7}$$
$$F_1(h, C) = \frac{2 \cdot \frac{1}{2} \cdot \frac{3}{5}}{\frac{1}{2} + \frac{3}{5}} = \frac{6}{11}$$

4. So far, the F_1-measure is only defined for classes and not yet useful to get an overview of the overall performance of the classifiers. To achieve such an overall assessment, one commonly takes the average over all classes using one of the following two approaches:

   · Micro Average F_1-Measure: The values of TP, FP and FN are added up over all classes. Then precision, recall and F_1-measure are computed using these sums.

   · Macro Average F_1-Measure: Precision and recall are computed for each class individually, afterwards the average precision and average recall are used to compute the F_1-measure.

   Compute the Micro- and Macro-Average F_1-measures for the example above. What do you observe?

   Suggested solution:
   Micro Average F_1:

$$\begin{aligned}
|TP| &= 4+3+3 = 10 \\
|FP| &= 4+4+2 = 10 \\
|FN| &= 3+4+3 = 10 \\
\text{Precision} &= \frac{|TP|}{|TP|+|FP|} \\
&= \frac{10}{10+10} = \frac{1}{2} \\
\text{Recall} &= \frac{|TP|}{|TP|+|FN|} \\
&= \frac{10}{10+10} = \frac{1}{2} \\
F_1 &= \frac{2 \cdot \frac{1}{2} \cdot \frac{1}{2}}{\frac{1}{2}+\frac{1}{2}} \\
&= \frac{1}{2}
\end{aligned}$$

   Macro Average F_1:

$$\begin{aligned}
\text{Average precision} &= \frac{\frac{1}{2}+\frac{3}{7}+\frac{3}{5}}{3} \approx 0.51 \\
\text{Average recall} &= \frac{\frac{4}{7}+\frac{3}{7}+\frac{1}{2}}{3} = 0.5 \\
F_1 &\approx \frac{2 \cdot 0.5 \cdot 0.51}{0.5+0.51} \approx 0.505
\end{aligned}$$

Exercise 5-2 Events and Sample Spaces (1 point)

1. We have a system of several fuses. We can examine each single fuse to see whether it is defective. The sample space for this experiment can be abbreviated as $\Omega = \{N, D\}$, where N represents not defective, D represents defective.
   Now we want to examine three fuses in sequence and note the result of each examination. What is the sample space Ω?

   Define two events on this sample space as verbal descriptions (similar to a die roll coming out even) and write the list of the sample space elements that are members of each of these events.

   Suggested solution:
   The sample space of an experiment, denoted by Ω, is the set of all possible outcomes of that experiment.

   An outcome for the entire experiment is any sequence of Ns and Ds of length 3, so

   $$\Omega = \{NNN, NND, NDN, NDD, DNN, DND, DDN, DDD\}$$

2. As an experiment, we observe the number of pumps in use at a seven-pump gas-station, so simple events are the numbers $0 - 6$ (pumps in use). Given the events $A = \{2, 4, 5, 6\}$, $B = \{0, 1, 4, 6\}$, and $C = \{1, 2, 3, 5\}$, which simple events are contained in

   (a) $A \cup B$?

   Suggested solution:
   $A \cup B = \{0, 1, 2, 4, 5, 6\}$

   (b) $A \cup C$?

   Suggested solution:
   $A \cup C = \{1, 2, 3, 4, 5, 6\}$

   (c) $A \cap B$?

   Suggested solution:
   $A \cap B = \{4, 6\}$

   (d) $A \cap C$?

   Suggested solution:
   $A \cap C = \{2, 5\}$

   (e) $\overline{A}$?

   Suggested solution:
   $\overline{A} = \{0, 1, 3\}$

   (f) $\overline{A \cup C}$?

   Suggested solution:
   $\overline{A \cup C} = \{0\}$

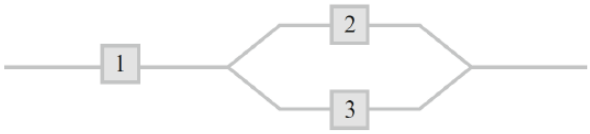
   (g) $\overline{B \cup C}$?

   Suggested solution:
   $B \cup C = \{0, 1, 2, 3, 4, 5, 6\} = \Omega$
   $\overline{B \cup C} = \{\}$

(h) $\overline{A \cap B}$?

Suggested solution:
$\overline{A \cap B} = \{0, 1, 2, 3, 5\}$

3. Three components are connected to form a system as shown in this diagram:



Because the components in the 2-3 subsystem are connected in parallel, that subsystem will function if at least one of the two individual components functions. For the entire system to function, component 1 must function and so must the 2-3 subsystem. The experiment consists of determining the condition of each component (S (success) for a functioning component and F (failure) for a non-functioning component).

(a) What outcomes are contained in the event D that exactly two out of the three components function?

Suggested solution:
$D = \{SSF, SFS, FSS\}$

(b) What outcomes are contained in the event E that at least two of the components function?

Suggested solution:
$E = \{SSF, SFS, FSS, SSS\}$

(c) What outcomes are contained in the event G that the system functions?

Suggested solution:
$G = \{SSF, SFS, SSS\}$

(d) List the outcomes in $\overline{G}$, $D \cap G$, $D \cup G$, $E \cup G$, and $E \cap G$.

Suggested solution:
It is helpful to consider $\Omega = \{SSS, SSF, SFS, SFF, FSS, FSF, FFS, FFF\}$.

- $\overline{G} = \Omega \backslash G = \{SFF, FSS, FSF, FFS, FFF\}$
- $D \cap G = \{SSF, SFS\}$
- $D \cup G = \{SSF, SFS, FSS, SSS\}$
- $E \cup G = \{SSF, SFS, FSS, SSS\}$
- $E \cap G = \{SSF, SFS, SSS\}$

Exercise 5-3 Conditional Probability (1 point)

Suppose that of all individuals buying a fastfood burger, 70% include fries in their purchase, 30% include a soda, and 15% include both fries and a soda. Consider randomly selecting a buyer and let $A = \{\text{fries purchased}\}$ and $B = \{\text{soda purchased}\}$.

Then $\Pr(A) = 0.7$, $\Pr(B) = 0.3$, and $\Pr(\text{both purchased}) = \Pr(A \cap B) = 0.15$.

1. Given that the selected individual included fries in their purchase, what is the probability that a soda was also purchased?

   Suggested solution:

   Recall that for any two events A and B with $\Pr(B > 0)$, the conditional probability of A given B is defined by:

   $$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

   Thus, the solution is

   $$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)} = \frac{0.15}{0.7} = 0.21$$

2. Given that the selected individual included a soda in their purchase, what is the probability that fries were also purchased?

   Suggested solution:

   $$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{0.15}{0.3} = 0.5$$

Exercise 5-4 Silhouette and k-means implementations in scikit-learn (1 point)
This exercise will be done as a lab-session in class. Explore the code on
https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html.
To do this as lab session in class, bring your laptop with Python, NumPy, SciPy, and SciKit-Learn installed.

1. What is the termination criterion in k-means in the scikit-learn implementation?

   Suggested solution:
   Understanding some of the input parameters to the k-means function:

   - n_clusters
     The number of clusters to form as well as the number of centroids to generate.
     Default is 8.
   - n_init
     Number of time the k-means algorithm will be run with different centroid seeds. The final results will be the best output of the runs in terms of inertia.
     Default is 10.
   - max_iter
     Maximum number of iterations of the k-means algorithm for a single run.
     Default is 300.
   - tol
     A parameter that controls the tolerance with regard to the changes in the within-cluster sum-squared-error to declare convergence.

   So the termination criterion is n_init and max_iter, as it will run k-means n_init times with at most max_iter iterations per run.

   The k-means implementation in scikit-learn stops early if it converges before the maximum number of iterations is reached.

   Termination also depends on the tol parameter, as this determines when convergence is reached.

2. Why can we get negative Silhouettes in this example?

   Suggested solution:
   Recall how silhouette coefficient is calculated.
   Let $a(o)$ be the distance between o and its "own" cluster representative.
   Let $b(o)$ be the distance between o and the closest "foreign" cluster representative.

   $$s(o) = \frac{b(o) - a(o)}{\max(b(o), a(o))}$$

   Note that the silhouette coefficient can only be negative if o has a shorter distance to the "foreign" cluster representative, than to its "own".