2023-02-22 3rd appello 2023

1st candidate (via Teams, probably exchange student 6 CFU):

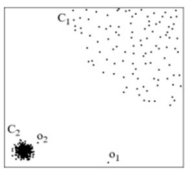
1. How it works chi-square for the redundancy of features in the pre-processing phase?
    a. Write the formula of chi-square and comment it
    b. How is computed the *expected* value in the formula of previous question?
    c. If you just compute the totals of (?) this is just an estimation of what?
2. Can you write the formula of the Savitzky – Golay filter?
    a. Explain it
    b. What is the difference between Savitzky Golay filter and the rectangular filter?
3. Can you write the formula of precision and recall?
    a. What is the formula of accuracy?
4. Explain how the k-means clustering algorithm works
    a. Write the cost function that we want to minimize in the iterative approach that we use in k-means (What do you want to minimize with the algorithm? What is the error function that you want to minimize?)

2nd candidate (in presence, project presentation too):

1. Can you talk about sequential patterns? (definition of sequence, subsequence, itemset, … which algorithms for seq patterns: Apriori All, Apriori Some, FPGrowth for seq pattern mining)
    a. What are in general the steps of an algorithm of sequential pattern mining?
    b. How can we generate the large itemsets from the single item frequent itemset?
    c. Name one algorithm for seq pattern mining that is not based on Apriori? (FPGrowth for seq pattern). Explain on what it is based and how it works.
        i. What is the problem with parallel projection approach? (amount of memory needed)
        ii. Parallel and partition projection
2. Can you talk about precision and recall?

3rd candidate (in presence):

1. Outlier detection in general: explain what is the problem and describe some solutions. (global, collective, contextual outliers; supervised, unsupervised, semi-supervised approaches; statistical, proximity approaches)
    a. Can you explain one of the approaches that you mentioned (statistical, proximity (distance, density))? (talks about distance approach, write the formula for determining if an object is an outlier) $\frac{\|\{o' | dist(o,o') \leq r\}\|}{\|D\|} \leq \pi$
        i. What is pi in the formula?
        ii. What is the problem of distance based approach? (draws a plot in which some points are very dense and clustered together, … (problem of different densities))

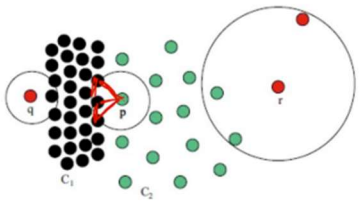


    b. Talk about density based approach, write the formula for determining ***outlierness*** in terms of density approach
        i. Why the outlierness formula solves the problem shown in previous plot?

$$O_k(x_i) = \frac{D_k(x_i)}{\frac{1}{k}\sum_{j \in Nk(xi)} D_k(x_j)}$$

ii. What is the problem that persists with this approach? (If clusters are close, outlierness gives unintuitive results)

c. What is another approach based on density? (LOF: Local Outlier Factor; writes reachability distance definition; local reachability *density* formula)
   i. Write the definition of LOF



In this example, p has higher outlierness than q and r:
- The green points are not part of the KNN list of p for small k

---

4th candidate (in presence):

1. Clustering with Constraints: what is the problem and what is the solution? (Must link, cannot link constraints, δ-constraint (Minimum separation) formula, ε-constraint formula, talks about penalty factor)
   a. What is the algorithm that we have to use (COP K-Means)
      i. What is the cost function of k-means?
      ii. What is the aim of minimizing this cost function? (to obtain a cluster compact and well separated clusters)
2. How do you evaluate the quality of a clustering? (distinguishes between the case in which you have ground truth and when you have not)
   a. Talk about methods to adopt when ground truth is available. (Precision and recall bcubed formulas)
      i. What do you think about these metrics? (the problem here is that you expect that clusters are compact, this is not always true)

$$\text{Precision BCubed} = \frac{\sum_{i=1}^{n} \frac{\sum_{o_j: i \neq j, C(o_i)=C(o_j)} \text{Correctness}(o_i, o_j)}{\|\{o_j | i \neq j, C(o_i) = C(o_j)\}\|}}{n}$$

$$\text{Recall BCubed} = \frac{\sum_{i=1}^{n} \frac{\sum_{o_j: i \neq j, L(o_i)=L(o_j)} \text{Correctness}(o_i, o_j)}{\|\{o_j | i \neq j, L(o_i) = L(o_j)\}\|}}{n}$$

5th candidate (in presence):

1. Talk about feature selection in the preprocessing phase, focus on heuristic approach. (define Mutual Information formula between two features, and the normalized mutual information formula; iteratively put in the set of selected features the ones that maximizes the G score (gives definition of G score and its formula))
   a. Make an example of Mutual Information on age and height.
   (Note that Mutual Information works only on features that can assume a finite defined set of values, so you have to define a categorical set of values for age (e.g. young, adult, old) and for height (low, medium, high)) (writes the formula of Normalized Mutual Information for all the combinations of the two attributes, writes the entropy formula)
   b. What is the value of Normalized Mutual Information in various cases? When I(X, Y) is low what does it means? (That the attributes X and Y are independent). Why the Mutual Information is zero when two features incorrelated? (the probability P(x, y) is equal to P(x) * P(y), then the logarithm of MI formula gives 0).
   c. Let suppose in the example of age and height we have a class label with three values (C1, C2, C3) and that those values are defined by the attribute age (when age is young, class is C1, adult is C2, old is C3). What happens in this case in terms of Mutual Information? How we can see in the formula that we have the maximum value of I (between age and class)?

d. Describe one iteration of feature selection by starting from an empty set of selected features S = {empty set}. (we iterate the selection of features until the cardinality of S is equal to k)
    i. How we can determine the best value for k? (you can test many values of k and understand which gives the best result)
e. Do you think that we can use Normalized Mutual Information Gain approach to select features only for classification problems? (since ground truth is needed, we cannot use it for clustering).
    i. What is another approach that is good for clustering or in general for when you do not have ground truth available? Chi-square test
        1. How do you calculate the expected value for an attribute?

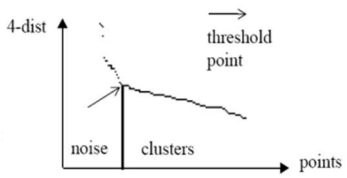
6th candidate (in presence):

1. We talked about Association Rules, talk about FP Growth. (how many scans are needed, why it improves other approaches (Apriori – like algorithms))
    a. Expose an execution of FP Growth in its phases. How the tree is built?
        i. After that you obtained the ordered list of items, what do we do to the dataset of transactions? (you have to sort the items in the transactions by the decreasing frequency order previously obtained)
        ii. After you built the tree, how do you extract the frequent patterns? (start creating frequent itemsets by increasing frequency considering a minimum threshold)
    b. What is the advantage of FP Growth compared to apriori? (we do not have to generate all intermediate candidate set, and that usually the FP tree can fit in memory, only two scans of dataset)
    c. How we can limit the number of frequent pattern to return to the user? (we can return just the closed pattern)
        i. Give the definition of closed pattern
        ii. If I just return the closed pattern to the user, will the user have all the information that he needs? (with closed patterns we do not loose information
        iii. Let assume that we return the closed items only, and that the returned patterns are {i1, i2, i3} with support 10, and {i1, i2} with support 15. Can we determine the support of i3? (since the given ones are all the closed itemsets, we can deduce that there is no itemset that includes i3 with a support higher than 10, so the support of i3 is 10)
            1. In the previous example what is the support of i1? The one of i2? The one of {i1, i3}?
    d. How do you transform a pattern in an association rule?
        i. Can you write the formula of the confidence of an association rule? (conf(A -> B) = P(B|A) = (sup(A U B) / sup (A)))
            1. Do we search for high or low values of confidence? (high)
            2. What happens if the support of B is low?
            3. What we cannot see from this formula?

7th candidate (in presence):

1. Can you explain me how it works DBSCAN? (notion of core object, parameters, …)

    a. When we terminate to add points to a cluster in DBSCAN?
    b. What is the characteristic of a cluster obtained with DBSCAN? (all the points inside the cluster are… Density connected)
        i. What is the definition of density connected?
    c. What is the problem we have with DBSCAN? (to establish the EPS and min points parameters)
        i. How we can determine those parameters effectively?
2. What is another method of clustering based on density? (OPTICS)
3. Can you write the formula of precision?
    a. And the formula of recall?



8th candidate (in presence):

1. Can you explain me how it works decision tree learning? (example, entropy, divide & conquer approach)
    a. Define entropy formula
    b. How do you generate the nodes in the tree to create the splits? (e.g. by using info gain)
        i. What is Info gain? (formula and definition)
2. Do you remember the Naïve Bayes Classifier?
    a. Given the formula of conditioned probability, how we can compute the various probabilities that appear in the formula? What is the Naïve assumption?
    b. Why it is so costly to determine the probability of P(X | H) without the independency assumption? (/ Why we have to make the assumption of independence?) (Practical difficulty: requires initial knowledge of many probabilities, significant computational cost)

9th candidate (in presence):

1. Explain the Naïve Bayes classifier.
    a. How to compute the P(X | Ci) given the assumption of independence?
        i. Let assume we have two attributes for X (x1 and x2) and assume that each of them can assume 3 values. Write the calculation to determine P(X | Ci) with the naïve assumption of independence.
2. Explain what is the ROC curve.
    a. Usually in TPR we have just one point in the ROC curve and not a curve. How do we compute the AUC (area under curve)? You have to connect the point of the model in the graph to (0,0) and to (1,1).
    b. What is the area under the curve in that case? (The one between the curve (/ polygon) and the line between (0,0) and (1,1) (diagonal line))
