Appello 2021/01/19

1. Let's suppose that we have to solve a classification problem and we would like to generate a classifier model which is INTERPRETABLE. Can you suggest me any such type of classifier?
    a. RULE BASED CLASSIFIER
2. How can we generate a rule based classifier?
3. How can we generate a decision tree?
4. Rule based classifier conflict resolution mechanism.
5. Imbalanced classification problem: how to compare different classifiers? What is the process to find the best classifier? Statistical t-test. Parameter and Non-Parametric test. Assumptions to apply this t-test: normal distribution. Wilcoxon test.
6. Dataset balancing solutions. SMOTE.

1. Can you explain me how the BIRCH algorithm works?
2. How is it possible to locate an instance in a leaf. Parent_CF = Sum of Childs_CF.
3. Branching factor. Second Threshold.
4. Drawbacks? [The result depends on the order by which instances are given to the model.]
5. Clustering tendency. Hopkins test + Formula.
6. Natural Clusters? Why do we need to test is?
7. K-means clustering algorithm. Drawbacks? [Local minimum. Affected by Outliers. Computationally expensive.]
8. How to find the value for k?

1. Types of outliers and how to use proximity based techniques to detect outliers.
2. LOF.
3. Formula of the correctness.

1. Clustering with constraints.

2. Streaming data problem: how to deal with streaming data. Classifier for streaming data.

1. How DBScan works? Definitions. Core object.
2. Density connection concept.
3. What can we guarantee in terms of clusters after applying DBScan?
4. Convex vs Concave clusters. [The problem is that in one case I can not use the silhouette (concave).]
5. How to find y_distance and the minimum number. Heuristic solution does not completely solve the problem. This is why we introduce the OPTICS method.
6. Draw what we can obtain as output of the OPTICS method.

1. Graph mining problem. How to solve it? The traditional distance definition does not work well.
2. SIMRANK. [The main problem of this approach is that we have to store SUV and that it is costly O(n^2).]
3. Describe how the SCAN algorithm works.

1. Preprocessing: how to find redundant attributes?
2. Chi-sqaure. If it is high? If it is low?
3. Degree of freedom.
4. What is the assumption when we compute the expected value?
5. Preprocessing: filter noise in signals. Smoothing. Savitzky–Golay filter.

1. Naïve Bayesian classifier. Why do we need the naïve assumption? Otherwise computing the probability is very costly. Example of the possibility on two classes, and 2 attributes with 3 possible values each. Why is it difficult to compute this probability?
2. K-NN classifier. How to determine the nearest k data points? [Compute distances, sort the list in ascending order and return the first k items in the list.]

3. Complexity of K-NN? [O(n), O(n*log(n)), O(1)]
4. Editing methods. Wilson method. [Wilson, Multiedit, Supervised Clustering].

1. ADABOOST.
2. Classifier Overfitting/Overtraining.

1. Can you explain me how we can generate a decision-tree?
2. Information gain, gain ratio, gini index.

1. Curse of dimensionality.
2. Feature reduction using the heuristic approach with the ground truth.
3. Numerosity reduction. Binning, sampling, clustering. Parametric approach: linear regression model – only store the model paramters.
4. Association rules quality measures. [Support, confidence, all_conf, max_conf, kulc, cosine, IR.]

Appello 2021/02/25

- Outliers:
  1. Talk about outlier detection. Introduce the problem, how many different types of outliers can we have, and present some solutions to deal with this problem.
  2. Focus on the local outlier factor technique.
  3. Reachability density formula. Local outlier factor (LOF).
  4. Explain this formula tells us if a point is an outlier or not from an intuitive point of view.
  5. Local reachability density.
  6. How to detect outliers using the angle.

- How can we learn a decision tree starting from the training set:
  1. Attribute selection methods.
  2. In the hypothesis that information entropy is equal to zero, what is the scenario in the training set?

- How BIRCH works:
  1. What is the advantage? (Streaming data: incremental mining)
  2. What are the disadvantages? (Only numerical attributes in CF)
  3. BIRCH can manage outliers using the size of the Diamater (D). [The final result you obtain depends on the order by which instances are added in sequence: THIS IS THE BIGGEST DISADVANTAGES.]
  4. Cluster quality evaluation methods. Intrinsic and extrinsic.
  5. Silhouette index.
  6. Bicubd, Precision, recall, correctness.
  7. Why did they introduce this indexes in the literature if we can not really use them in real world application where we do not know the class label? [This is done because we use this metrics for benchmarking when new clustering algorithms are proposed.]

- When we have to cope with classification/clustering, we have to compare different algorithms. Statistical confidence related to the best model.
    1. T-test, t-test assumption: normal distribution
    2. Confidence interval
    3. 10-fold cross validation
    4. Wilcoxon test
    5. Difference between t-test and Wilcoxon

- Make some considerations about the preprocessing stage: how we prepare the data for data mining algorithms? Data dimensionality reduction.
    1. Attribute selection using Mutual Information, normalized mutual information and entropy. The attribute which maximizes the mutual information with the class label and minimizes the mutual information with all the already selected attributes.
    2. The disadvantage: we have to select at the beginning the number of attributes we want to select. This problem is not so dramatic in PCA, why?
    3. Eigen values and eigen vectors in PCA, how are they used?

- Clustering:
    1. Can you please explain me how K-means work?
    2. Write the cost function.
    3. Make the plot of the cost function based on the value of k.
    4. How can we determine the optimal value of k?
    5. The Hopkins test for clustering WITH FORMULAS.
    6. Possible values for the Hopkins test.

- Graph mining:
    1. What is the problem? Possible solutions?
    2. Tell me some techniques to perform graph clustering.
    3. Simrank.

4. Assumption on the terminating conditions of the iterations.
5. Similarity/dissimilarity measures for clustering.
6. Single, complete and average linkage.

- Clustering:
  1. DENCLUE.

Appello 2021/07/29

- DENCLUE (he asked all the details)
  - how to evaluate the clusters?
  - is silhoutte useful for evaluating clusters found by DENCLUE? [NO]
- Naive Bayesian
  - write formulas
  - we want to compare different probabilities threshold, how to use ROC?
  - how to compare two classifiers with the help of ROC?
  - Explain precision and recall
  - suppose we have dependencies, how to classify?
- Explain SimRank
  - why do we use simRank?
  - how can we exploit it? (in hierarchical clustering)
  - how to cluster graphs directly? (SCAN) -> B-cubed?
- FPGrowth?
  - what is closed itemset? max itemset?
  - which one to offer to users? (none, offer association rules)
  - what are the metrics to measure the quality of rules?
  - null-variant and imbalanced problems?

Appello 2021-02-26

- Correlation coefficient:
    1. What is reasoning behind the formula of the correlation coefficient intuitively?
    2. Scatterplot
    3. Correlation coefficient = 1, what type of scatter plot does it correspond to?
    4. Correlation coefficient = -1.
    5. Text Vector.

- Outlier detection:
    1. Techniques we can use: in particular techniques based on proximity? Can you analyze some of them?
    2. Local outlier factor (LOF).

- Hierarchical clustering:
    1. Introduce the concept and then describe at least on algorithm that we described during the lectures.
    2. Linkage methods
    3. Distance measures: formulas.
    4. Which type of distance do we use with divisive hierarchical clustering techniques?
    5. How to compare the performance of two clustering methods?
    6. Write the formulas for:
        - Silhouette
        - bcubed precision and recall

Appello

1. LOF: what is the specific problem for which this technique was introduced? [Deal with different density]
2. Distance based outlier detection.
3. Outlier detection in high dimensionality spaces. Angle technique.
4. How to compare different clustering algorithm.
    - Silhouette coefficient. Possible values and meaning.
    - Intrinsic vs extrinsic methods.
    - Precision and Recall in Bcubed approach. Definition of correctness.
5. Heuristic attribute selection method based on mutual information.

Domande scritto 1° appello 2020

- adaboost;
- boxplot;
- denclue;
- simrank;
- esercizio con una tabella e dava una association rule e chiedeva se era strong in base a un minsup e una minconf. Sempre nell'esercizio poi chiedeva di vedere se i due item erano correlati e dava da calcolare poi allconf, maxconf, kulcinsky e cosine

Domande scritto 3° appello 2020

- Descrivere apriori algorithm e suoi limiti.
- Chi square, cosa è come si trova e quando si usa.
- Cosa è la roc curve.
- Denclue.
- Scan.
- Descrivere i 4 tipi di bicluster

Domande di orale post-scritto (mix):

- Come si fanno a confrontare due classificatori e come si fa a dire se un clustering è fatto bene
- Outliers
- K-Means
- Random Forest
- Info gain (con formule)
- Difetti decision tree
- Imbalanced dataset

1st appello and 2nd appello 2021

- **12 CFU**
  - Rule based classifier, formulas of foil_gain and foil_prune, conflict resolution strategies
  - Imbalanced datasets, use CV with F-Measure, why rebalancing, SMOTE?
  - which is the null hypothesis in a statistical test for classifiers? T-test + Wilcoxon, assumptions, explanation
  - How BIRCH works, advantages and disadvantages
  - Hopkins statistic
  - k-means, how it works and what are the drawbacks, Elbow metric
  - Types of outliers and proximity based approaches for outlier detection
  - Outlier detection in high dimensions
  - Silhouette metric, Bcubed precision and recall
  - Clustering with constraints and what are constraints
  - Datastreams and how to implement a classifier for datastreams, Hoeffding Tree
  - Graph clustering in general and Simrank
  - SCAN
  - Kmeans, methods of assessing clustering tendency, elbow, silouhette
  - precision, recall, accuracy e le roc curve con le relative formule di questi argomenti
- **6 CFU**
  - DBSCAN, explanation, advantages and disadvantages
  - How to tune epsilon and MinPts in DBSCAN
  - OPTICS
  - DENCLUE
  - Chi-square
  - Smoothing

Raccolte dal Gruppo Telegram AIDE

- Kmeans, methods of assessing clustering tendency, elbow, silouhette
- A me ha chiesto precision, recall, accuracy e le roc curve con le relative formule di questi argomenti
- Normally, he doesn't ask about neither Spark nor Hadoop.
- Diciamo che vuoi analizzare un segnale
  - how do we preprocess it? (Smoothing)
  - can you explain me the philosophy behind the method? (Muovi la finestra)
  - how many parameters? (Ampiezza, frequenze)
  - how do you establish the size of the window? (Smoothing ratio)
- Clustering
  - k-means a parole tue
  - is there any heuristics to determine k? (Elbow)
  - how do we measure clustering tendency? (Hopkins)
  - DBSCAN a parole tue
  - compare DBSCAN e k- means
  - OPTICS
  - Che cosa è un test non parametrico
- Classification
  - kNN
  - Info Gain, formula
  - ROC Area
- Outlier
  - che tipi ci sono
  - come li tratti localmente
- SCAN per un grafo con la formulazione precisa
- LOF
- hierarchical clustering
- Clustering con constraints, come si procede in generale

1st appello 2021

- **12 CFU**
  - Rule based classifier, formulas of foil_gain and foil_prune, conflict resolution strategies
  - Imbalanced datasets, use CV with F-Measure, why rebalancing, SMOTE?
  - which is the null hypothesis in a statistical test for classifiers? T-test + Wilcoxon, assumptions, explanation
  - How BIRCH works, advantages and disadvantages
  - Hopkins statistic
  - k-means, how it works and what are the drawbacks, Elbow metric
  - Types of outliers and proximity based approaches for outlier detection
  - Outlier detection in high dimensions
  - Silhouette metric, Bcubed precision and recall
  - Clustering with constraints and what are constraints
  - Datastreams and how to implement a classifier for datastreams, Hoeffding Tree
  - Graph clustering in general and Simrank
  - SCAN
  - Apriori
- **6 CFU**
  - DBSCAN, explanation, advantages and disadvantages
  - How to tune epsilon and MinPts in DBSCAN
  - OPTICS
  - DENCLUE
  - Chi-square
  - Smoothing