Contents

Question & Answers

Exam 07/09/2012

Q. *Shortly describe SVD, why it is used, how it works.*

A. Singular Value Decomposition is a technique for feature selection, which may lead to have less noise in the data, speed up the execution of some algorithms and make them more easily visualizable.
It is based on the fact that any matrix can be decomposed in the following:

$$M = U\Sigma V^T$$

By analyzing those matrix, the most meaningful is Σ, a diagonal matrix that represents "concepts" in the dataset. We can see which are the most important concept in our dataset (the greatest ones) and decide to drop one of them (i.e. drop a column of both U and Σ, and a row in U^T). Then, after the computation of the manipulated matrices, we have an approximation of the original matrix M, with less features.
You stop dropping concept if the energy of the "new" Σ is under 90% of the original Σ, where the energy is defined as the sum of squares of diagonal elements.

Exam 24/05/2013

Q. *According to what seen during the course, what do the terms "completeness" and "optimization" refer to in the context of data mining?*

A. Answer at page 16.

Exam 01/07/2013

Q. *What are the advantages/disadvantages in using supervised and unsupervised discretization?*

A. The discretization is the process to convert continuous features into nominal, by grouping them into intervals.
You can do it in two ways:

- Supervised: you decide the intervals taking into account also the class attribute.
- Unsupervised: you decide the intervals only taking into account the continuous attribute itself.

By using supervised discretization, you try to reach subspace with high purity w.r.t. to class labels, therefore it can be very useful in classification algorithm.
However, this can be useless or even misleading if the attribute you are pre-discretizing is not correlated to the class attribute: in this case it is better to use an unsupervised discretization.

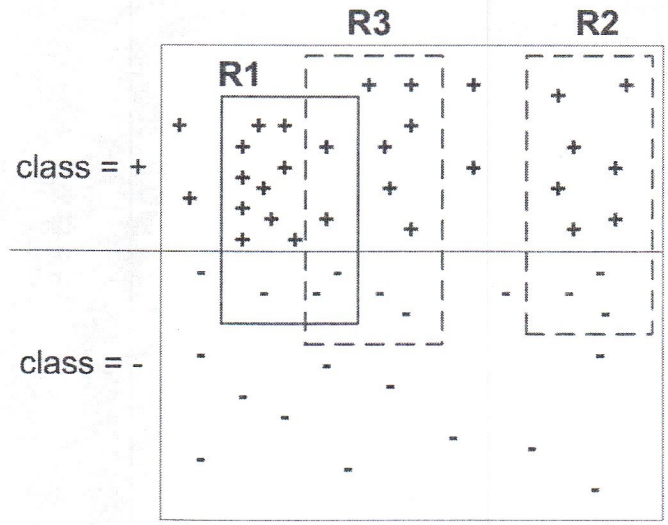
Exam 11/09/2013

Q. *In sequential covering algorithms, the positive examples covered by a generated rule are typically eliminated from the dataset.*
Why this is done? What would happen if all the examples covered by the rule were kept in the dataset? And what if all the covered examples were eliminated?

A. There are three possible approaches when applying sequential covering algorithm. After the generation of the rule you can:

- Eliminate all the covered examples
- Eliminate all the positive covered examples
- Eliminate all the negative covered examples

First of all, me must clarify that if you do not perform the elimination of the examples the algorithm would generate always the same rule.
Take the following example: Let's compute the accuracy of each rule w.r.t. the three possible approaches:



	Eliminate all covered examples	**Eliminate all positive covered examples**	**Eliminate all negative covered examples**
R1	12/15	12/15	12/15
R2	7/10	7/10	7/10
R3	6/8	6/10	8/10

As we can see, the accuracy of R3 changes according to the approach we choose.
The real accuracy of R3, considered all the examples it covers, is 8/12.
From the result showed, we can say that the second approach is actually an underestimate of the accuracy of R3 and might be shown to a customers as a lower-bound of the performance of the rule.
On the other hand, the other two approaches tend to prevent underestimate of the rule.

Exam 07/07/2014

Q. *Explain what are boxplots. Discuss the difference between barplots, histograms and boxplots. When do they should be employed?*

A. They are three ways to visualize data.
Boxplot is a way to show how the a numerical variable is distributed: they show a box with two subsections, delimited by three lines: first quartile, second quartile, and third quartile. The size of each subsections is proportional to the number of samples inside the limits. Outside of the box, we have a lower bound and an upperbound (Max ¡ Q3+1.5 IQR and Min ¿ Q1 - 1.5 IQR): in these zones (even over Max and Min) we typically find outliers.
Histograms is another way to analyze distribution of a numerical variable: they are rectangles whose size is proportional to the number of samples seen in a given range and provide an estimate of the probability of finding an example in that range.
Barplots, instead, do not show the distribution of data, but simply represent the relation between a numerical value and a categorical one. It is often used for comparison of categorical features.

Exam 04/09/2014

Q. *What is feature creation? And why is it used? (provide at least one example and explain the advantage of feature creation in the specific example)*

A. Feature creation is a preprocessing technique that extrapolates some information by the features of the dataset and transforms it into another feature, usually more readable, more visualizable and easy to manipulate. An intuitive example of feature creation is the age w.r.t. to the date of birth: let's assume we have a dataset containing people details, and among them there is a feature called "Date". If we wanted to visualize the number of people for each age, it is useful to extrapolate the "Date" in a new feature "Age" and then plot it.
Furthermore, the attribute "Age", that is numerical, can now be used for a regression.

Exam 24/09/2014

Q. *In the typical rule induction process following a sequential covering approach, are the rules produced overlapping or not?*
If they are, what strategies are used to solve the conflict when two or more rules are advocated for one example?
If they are not, what is the mechanism that guarantees that the rules are not overlapping?

A. Yes, it can happen that rules overlap if you do not eliminate all the examples covered by a new generated rule (you can both eliminate only the positives or only the negatives). What is typically done is to give a precedence to the rules.
A way to prevent overlapping of rules is to eliminate all the covered examples, both positives and negatives.

Q. *Briefly define Information Retrieval (IR), the major IR approaches, and the basic measures for text retrieval.*

A. Information Retrieval aims to find the most relevant documents w.r.t user's input.
There are two main approaches:

- Pull mode: it is typical of search-engine, where the user takes the initiative and asks for something.
- Push mode: it is typical of recommender systems, where the system takes the initiative and provide something that the user is likely to ask for.

As far as text retrieval is concerned, the challenge is to find the most relevant document with respect to a query performed by the user.
The documents, in fact, are ranked according to a function *f(q,d)*, that measure how relevant they are.
There are different approaches:

- Similarity, where $f(q,d) = sim(q,d)$: the function ranks the documents from the most similar to the least similar to the query.
- Probabilistic, where $f(q,d) = P(R = 1|q,d)$: the function provides a probability estimate of the fact that a document is relevant.
- Probabilistic inference model, where $f(q,d) = P(q \rightarrow d)$
- Axiomatic model, where *f(q,d)* must satisfy a set of constraints.

Exam 02/02/2015

Q. *Does Information Gain Ratio always solve the problems introduced by highly branching attributes? If yes, explain why, if no explain what else you would do to solve the problem.*

A. Information Gain Ratio usually penalize highly-branching splits, in order not to result in overfitting. However, even it may help, it can result that still an attribute like "ID" has the highest Information Gain Ratio and thus it is selected for splitting, still leading no a useless Decision Tree.
What you typically do is dropping entirely the attribute "ID" for the building of the tree.

Q. *A company has to select the best clustering algorithm for their data, among a set of ten algorithms. They can provide you with two sets of data.*
One set of data consists in raw data about their customers. The other set contains data that have been labeled by a company expert who labeled customer records as "GOOD" or "BAD" based on their spending level.
How would you organize the comparison? Which data would you use? And how would you use the different data?

A. If I used the labeled data, the algorithm would be "biased" from the label "GOOD" or "BAD" while performing the clustering.
The most correct thing to do is to perform a blind clustering by using the raw data and then compare the results with the labeled data: if the algorithm worked properly, then it would be likely that points from the same clusters (or with a short distance between each other) will have the same label in the labeled dataset.
This way we can see which one(s) from the analyzed algorithm performs better with less information.

Exam 29/06/2015

Q. *What is the Bag of Words model and how does it relate to the Vector Space Model?*

A. Bag of Words model is a way to represent documents. They are represented as vector in a high dimensional space, where each axis corresponds to a keywords and the order of words does not count. Each word is then represented as a position in a vector: a position is set to "1" if the word is contained the document, "0" otherwise.
Vector Space Model exploits this high-dimensional space to measure similarity between a query and the documents in the space. The similarity function for Vector Space Model is:

$$sim(q, d) = \sum_{i=1}^{n} x_i y_i$$

That is a multiplication between each component of two documents.
There are different types of VSM:

- Basic VSM: words represented as 1/0 bits
- TF representation: word position contains the occurrences of the words in the document
- TF-IDF: word position is calculated as TF*IDF, where IDF = log((M+1)/k)

Once a similarity function between the query and all the documents is calculated, you rank them sorting according to their similarity score, in order to retrieve the most relevant ones.

Q. *You work at "I Know Data", a leader in Big Data. You manage two teams (A and B), which report their findings every week.*
Team A is working on a data set containing 200'000 records described by 120 attributes and one class attribute described by four labels (bad, average, good, fantastic); team B is working on a dataset containing 300'000 records described by 240 attributes and one class attribute described by 2 labels (ugly and pretty).
Both teams are applying various supervised classification methods.
During the weekly meeting, team A reports a classification accuracy of 0.53 (that is 53%), team B reports an area under the ROC curve of 0.53; what is your opinion of the two results?
Are they satisfactory? Is there a team that has a better result? In case, there is, which one?

A. Having an area below the curve of 0.53 is a very poor result in this case, since it is almost equal to random guessing between two target labels, therefore the work proposed by team B is not so significant: random guessing would give almost the same results, but in much less time.
On the other hand, team A resulted in an accuracy of 0.53, that is: the predicted label has the 53% of possibility to be correct. This is a quite better result than random guessing, since team A worked on a dataset with 4 target labels, and random guessing would have an accuracy of 0.25.
Eventually, team A showed a better result.

Exam 08/07/2015

Q. *Your company hired three consultants to develop a classification model from your clients' database in which every client was previously labeled as high or low depending on its spending level.*

- *The first consultant proposes to apply clustering to identify the structure of the data using all the data.*
- *The second consultant will apply Naïve Bayes (using the spending as the class) and will use tenfold crossvalidation to evaluate the model produced.*
- *The third consultant will apply Random Forest (also using the spending as the class) and also apply tenfold crossvalidation.*

You team leader calls you in her office and asks you an opinion about the three solutions. Elaborate.

A. As this is a classification problem, the solution proposed by the first consultant is not correct: once you have the clusters, indeed, you cannot classify any further example without running again the clustering algorithm.
The two others solution are more appropriate to this specific case and I would choose one way or another according to the domain: Naive Bayes, for example, might give bad results if the numeric variables are not normally distributed (this is probably the case). On the other hand, Random Forests are in general a good solution since they are very precise and they hardly overfit if the number of tree is high enough. After the application of Naive Bayes and Random Forest and the related CV, we can perform a t-Test in order to see if there is a significant difference between the two algorithms.

Exam 01/02/2016

Q. *Briefly describe SVD and for which purpose it is used.*

A. Answer at page 3.

Q. *Mr. Fonevoda gave you a database that contains the data of 1100 customers who claimed had their cell-phone cloned. An investigation showed that only 1000 of the filed claims were true. Accordingly, each claim in the dataset has been labeled as "cloned" or as "fraud" depending on whether the claim was true or not.*
Mr Fonevoda asks you to provide an automatic tool that can help the company identifying the possible frauds. For this purpose, you built a classification model from the dataset and evaluated the model performance applying a 10 fold cross-validation. According to your evaluation, your model has an accuracy of the 90% with a 0.1% variance.
However, when Mr. Fonevoda tests your model on an unseen dataset of 100 claims (containing 50 "fraud" cases and 50 "cloned" cases), your model results in a 50% accuracy.
How can you explain such a poor performance on the unseen cases?
How would you improve your model?

A. In a specific case where your model should find the "outliers", using the accuracy as a performance evaluation metric is a mistake:
In this case, if an algorithm classified the total 1100 customers' phones as "true" would result in a 90% accuracy , but will lose any "false" phone.
In fact, when applied on a dataset more balanced, the algorithm have terrible performances.

Then, the first thing to do is changing the evaluation metric, tending to evaluate the model w.r.t. Precision, Recall and F1.
In order to improve your model, you should work on a more balanced dataset, so what you can do is training a model this same dataset after a bootstrap.
Furthermore, you should set a Cost Matrix, giving an higher cost of misprediction (FN in particular), and use it both to train your algorithm and to evaluate it.

Q. *You are the CTO of "We Do Stuff" and you need to hire a data mining consultant to extract knowledge from your customer database that consists of one million records described by 10 attributes. Your duty is to examine all the proposals arrived to your company.*

- *Consultant A suggests to use Bagging with k-nearest neighbor and 1000 models.*
- *Consultant B suggests to apply bagging with neural networks and 10000 models.*
- *Consultant C suggests to apply random forests with 5000 models.*
- *Consultant D suggests to modify random forests to use them with the basic OneRule algorithm to generate an ensemble of 10000 models.*

Elaborate.

A. The proposal of consultant A is not that valid, because Bagging is an ensemble which works better with unstable classifiers, whereas kNN is a stable one, therefore probably the difference between directly a single kNN model (or Bagging with few models) should not worth the complexity of the algorithm.
Consultant B is proposing a good solution, since NNs are an unstable classifier, except for the fact that neural networks are usually slow to train, so, 10000 models might be too computationally expensive.
Consultant C and D suggest valid alternatives, but the application of basic random forests with 5000 models leads to a very large time complexity, whereas using a simple algorithm like OneRule with a larger number of models could lead to the same result in less time.
To conclude, the best solution proposed is the one from Consultant D.

Exam 04/07/2016

Q. *The company StereotypeThis is working on a system to profile people living in a certain geographical area. They collect several variables about each person like for example, their age, their salary, what they own, where do they live, etc. Given a population of hundreds of thousands individuals they want to segment the population and extract a small number of profiles that describe stereotypical citizens. For example, one profile might describe the "typical" senior citizen of the "typical" student from another region.*
StereotypeThis received four proposals.

- *Proposal 1. Company A, suggests first to apply hierarchical clustering; then since the clustering would just add the cluster label attribute, company A wants to apply a classification algorithm, namely Rote Learning, to derive a compact model of the target cluster attribute. This way, the model produced by rote learning would provide the needed profiles.*
- *Proposal 2. Company B, wants to use a representative based clustering, namely CURE which also allows for arbitrary shapes, and provide the final clusters representatives as the needed profiles.*
- *Proposal 3. Company C, wants to use another representative based clustering, namely BFR, and then since they state that the representation of clusters would be still too complex to be presented, they will apply decision trees using cluster labels as the target class.*
- *Proposal 4. Company D wants to use the CURE algorithm and then they want to apply the CBA* ***(not treated in 2017/18, therefore not considered in the answer)*** *algorithm to extract the profiles that the company requires.*

StereotypeThis asks you to help them to the rank the proposals from the best one to the worst one. Elaborate!

A.

- Proposal 1: It is a very efficient proposal, because Hierarchical Clustering is a quite precise algorithm that allows you to find all the possible cluster. We also know that it's an algorithm with a very high time-complexity $O(N^2 \log(N))$, therefore the choice of a classification algorithm would speed up the classification of further samples and might give a hint on the quality of the clustering.
- Proposal 2: Representative-based clustering are faster than other techniques, but also very unstable. Furthermore, they do not perform good in case of different sizes/densities between the real clusters, which is very likely the case in this context.
- Proposal 3: As a representative-based algorithm, BFR suffers from the same drawbacks of CURE, so I would discard this as an initial step. I still appreciate the choice of a Decision Tree as a classification algorithm, since it is a very intuitive algorithm also to out-of-field people.

My final rank then is (in order of preference): 1,3,2.

Exam 14/07/2016

Q. *The company StereotypeThis was not happy with your last consulting but still decided to give you another chance. The company is trying to develop a pipeline for customer segmentation and thus it is shopping for clustering methods. They are currently evaluating several products that provide density-based algorithms, partition-based algorithms (like k-means), etc.*
In particular, the company NoProblemWithClustering is proposing a set of partitional clustering algorithms that automatically determine the number of clusters. The company claims that this is an advantage.
List non-trivial situations in which this is not the case.

A. As we know, k-Means suffers from the problem of the choice of the initial centroids: having k real clusters, it is quite unlikely that the stochastic algorithm picks k centroids each belonging to a different real cluster. Therefore, knowing a-priori the number of clusters, might not be an advantage: they might be chosen "randomly wrong". What it is sometimes done in these cases, in fact, is to choose an higher number of centroids and then merge some clusters with post-processing.
Again, we know that k-Means does not perform fine when the real clusters have different sizes/densities or non-globular shapes, because they might not be correctly recognized.
Again, here you can set an higher number of centroids and then reduce them with postprocessing.

Exam 12/09/2016

Q. *Briefly define the concept of completeness and optimization as presented during the course and discuss them in the context of decision trees and association rules.*

A. The concept of completeness is related to the capacity of an algorithm to produce all the interesting patterns contained in a dataset, where as "interesting" we mean something that it's worth it to analyze and that might give a better and compact description of the space.
The concept of optimization is related to the capacity of an algorithm to produce only the interesting patterns, without showing some useless ones in the results.
As far as decision trees, we might have some troubles both with completeness and optimization: indeed, when you have an highly-branching attribute (such as an ID), you could build a 100% performing decision tree (on the train-set), but it does not give you any interesting pattern in the data.
Again, even without an attribute like that, a Decision Tree might require some pruning because it tends to split every possible subspace in the data, resulting is some interesting patterns and some not really worth to analyze.
As far as instance-based learning, we have seen kNN: in this algorithm we have the issue to choose k, if k is too large you might end up in finding too generic patterns, that includes points not really related to each other, whereas if k is too small, the classification is highly sensible to noise and not all the interesting patterns are found.

Q. *The company ZoolanderData asked three companies DataOne, ShouldClassify, and ProbData to help them compare four algorithms: a basic decision tree returning class labels, a naïve bayes, Logistic regression, and a version of k-nn modified to return probabilities values.*

- *DataOne compared the four algorithms using ROC curves. The result shows that the decision tree is the best performing algorithm with the largest area below the curve.*
- *ShouldClassify applied crossvalidation and compared the four algorithms. Their results show confirm the results presented by DataOne.*
- *ProbData also used ROC curves to compare the four algorithms but using the area below the curve their results show that logistic regression is the best performing algorithm.*

ZoolanderData asks you to comment on the results presented by the three companies.

A. Both DataOne and ProbData made a mistake, since ROC curve can only be computed for probabilistic classifiers: therefore there is no reason to compare probabilistic and non-probabilistic classifiers with this method. The most opportune method to compare this algorithm is crossvalidation, used by ShouldClassify.

Exam 23/11/2016

Q. *Briefly define the concept of completeness and optimization as presented during the course and discuss them in the context of decision trees and association rules.*

A. Answer at page 16.

Q. *The company ZoolanderData asked three companies DataOne, ShouldClassify, and ProbData to help them compare four algorithms: a basic decision tree returning class labels, a naïve bayes, Logistic regression, and a version of k-nn modified to return probabilities values.*

- *DataOne compared the four algorithms using ROC curves. The result shows that the decision tree is the best performing algorithm with the largest area below the curve.*
- *ShouldClassify applied crossvalidation and compared the four algorithms. Their results show confirm the results presented by DataOne.*
- *ProbData also used ROC curves to compare the four algorithms but using the area below the curve their results show that logistic regression is the best performing algorithm.*

ZoolanderData asks you to comment on the results presented by the three companies.

A. Answer at page 16.

Exam 30/01/2017

Q. *Your company is trying to solve the following problem. You have some data for which you have no additional information apart from the data themselves. You want to extract as much information as possible.*
You interview three data mining consultants for an opening in your firm. The three consultants propose the following approaches.

- *Consultant A: First we apply a decision tree. Then we prune it as much as possible so that we can identify large portions of the problem space. Then, we apply hierarchical clustering on the subspaces identified by the decision tree to find good description of the subproblems.*
- *Consultant B: I disagree with A. We should first apply clustering and then apply decision tree using the results of the clustering process. In this way we can extract a description of the clusters.*
- *Consultant C: They are both wrong. First apply a hierarchical clustering so that you can find some structure in the data. Then, on each cluster we just found we apply k-mean so that we can actually have a compact description of the clusters.*

Which solution is the best? Why the other ones are worse?

A. The solution proposed by Consultant A is completely wrong, because this one is not a classification problem: Decision Tree is a supervised algorithm, meaning that the building of the tree is leaded by the knowledge of the target variable. In this case, we do not have a target variable, therefore applying Decision Tree is not possible.

The solution proposed by C is quite heavy, since it applies two clustering algorithm: it can give a compact vision of the space, but not as clear as the one proposed by Consultant B: running a clustering algorithm and then a Decision Tree in sequence can give a hint on how the clustering procedure was carried on: Decision Tree is an easily-understandable algorithm, since it clearly shows the attributes which the splits are made on.

Exam 26/06/2017

Q. *Discuss the computational complexity of classification using k-nearest neighbors with a plain table-based representation and how such complexity might be reduced.*

A. KNN has a very high complexity when used in its plain table representation. For every element (n elements), we must scan the whole table (n^2). The total complexity is in the order of $O(n^3)$. A possible optimization in terms of complexity is using KD-Trees: the search space is split into hierarchical regions and the search is performed locally one region at a time.
When classifying an example, the tree is navigated until a leaf is reached and, during backtracking, only the neighboring subspaces are checked.
The criterion for choosing direction and point of split must be chosen a priori.
An example is:

- direction: along the direction with the greatest variance
- point: the median (or mean if data is skewed) along this direction

The worst case scenario still has a complexity of $O(n^3)$ but the average computation time decreases.

Exam 10/07/2017

Q. *The company Amazona is contacting you since they have a database in which several attributes are organized in hierarchies. Typically, a hierarchy is a sequence of attributes that represent the same information with different level of details going from the most general one down to the most specific one. For instance, the address of a client is represented by the hierarchy of attributes Nation, Province, City, ZipCode, and Street.*
The attribute Nation is the most general one, Province provides a more detailed information than Nation, City is more specific than the Province, ZipCode provides a more specific location within the City, and Street provides and even more specific location. Another example of hierarchy of attributes regards the products a customer bought like ProductClass (for example, "Electronics"), ProductType (for example, TV), Brand (e.g.,Samsung), and ID (for example, UE40J5100AW).

The company asks you to extend a classification method to exploit such hierarchies of attributes so that a more specific attribute is not used until more general ones are employed. For instance, the classification model generated by your method should not use ZipCode unless Nation and Province are also used.

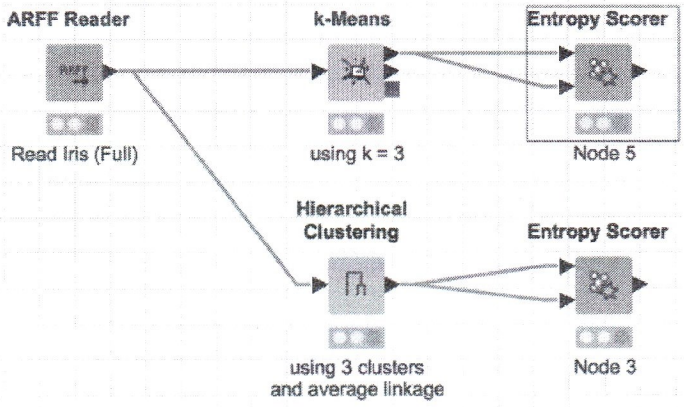
Describe how you would modify one of the classification algorithms discussed during the lectures to fit the company requirements highlighting the advantages and limitation of your proposed solution.

A. A valid algorithm applicable to this specific example is Decision Tree, which is a classification algorithms which assigns a label to a tuple by mean of a tree branched according to the attributes of the dataset.
Decision Trees are built by splitting and branching according to the values of the attributes, examples are therefore classified by following a path down to a leaf. A common criterion to decide the attribute to split on is the Information Gain, which gives a precedence to an attribute w.r.t. another according to the purity of a sub-dataset generated by the split.

You can modify this algorithm by stressing the fact that you can only branch on an attribute if the tree is not already split according to the relative father-attribute: for instance, you can only split on Province, if in the upper levels there is a split on Nation.
Decision Trees are quite efficient and this modification could match the needs of the company, but they tend to overfit, especially with highly-branched attribute. Furthermore, forcing the split according to this hierarchy may penalize splits with an higher information gain.

Exam 04/09/2017

Q. *You are attending a presentation of a consulting company. The presenter is showing some results of the analyses she performed. At some point, you get distracted and don't follow the presentation. Suddenly, you check the screen and you see the following schema. Your boss asks you to explain what the schema indicates and if you have any opinion about the pros and cons of the proposed approach. What do you reply to your boss?*



A. The schema proposed is a KNIME workflow, showing two parallel clustering algorithms.
On one hand, we have k-Means, that is a clustering algorithm which selects k (in this case k=3) random points as initial centroids of the clusters and then, for all the remaining points, they are assigned to the cluster of the nearest centroids. Then the centroids are updated and the process is repeated for every point until all the clusters stabilize.

On the other hand, we have Hierarchical Clustering using 3 clusters and average linkage to compute the distance between clusters. This algorithm starts with N clusters, where N is the number of points in the dataset, and iteratively creates a cluster merging the two nearest points.
Hierarchical Clustering is, in general, more efficient than k-Means, which suffers from the problem of the choice of the initial centroids. Unfortunately, Hierarchical Clustering is very slow (complexity of $O(N^2 \log N)$, because you have to compute the distance between all the points, whereas k-Means is faster, since you only have to compute the distance from the centroids.

The two proposed approaches may lead to think that they are equal, but this is not true, because k-Means is stochastic. Furthermore, we do not have information about the distribution of the data: if the real clusters have different sizes/densities or are not-globular k-Means could perform very badly.

Exam 27/11/2017

Q. *True or False: Two different decision trees that both correctly classify a set of training examples, will also classify any other testing example in the same way (i.e., both trees will output the same class for any other example). Explain.*

A. The answer is false. Take the case of a split on a numerical variable: when building a tree, you base the splits on the values present in the training set, so if you have values between 50 and 60 a decision tree could split on 52 and another on 56. Then, when a test with value 55 will be classified, the two decision trees may give different results.

Q. *Briefly discuss the differences between sequential covering algorithms, decision trees and nearest neighbor methods.*

A. The three algorithms proposed are classification algorithms.
In particular, sequential covering and decision trees are similar because the first one produces rules (IF-THEN-ELSE), while you can extract indirect rules from the second one. By the way, there are differences in which the rules of both the algorithms are built:

- Sequential Covering Algorithms generate rules to cover the whole dataset, but when a rule is generated it eliminates samples from the dataset (choosing between eliminating all the samples covered, only the positive ones or only the negatives one).
- Decision Trees, instead, build the classification according to an evaluation of purity on the sub-dataset of an attribute-split. Another important difference lays in the fact that Sequential Covering cannot handle with missing values.
- Nearest Neighbor methods, instead, does not produce rules and does not train over a dataset (the other two do): it simply takes a point and calculates the distance from the others, then it picks the nearest ones and assigns a target label according to a criterion (e.g. majority voting). This last algorithm is instead influenced by:
  - The way to compute distance;
  - How many neighbors to pick (if too small it may be sensible to noise, if too large may be inaccurate).
  - Nearest Neighbors is not suitable for too large dataset and performs very bad with a large amount of attributes, because with many dimensions all the points are equally distant.

Exam 16/01/2018

Q. *Your colleagues do not believe that clustering and classification techniques can be combined in several ways. What arguments/examples would you use to convince them?*

A. An example of usage in combination of these two techniques is when you do not have a target label for classification (therefore, classification algorithms would be useless). What you can do is running a clustering algorithm (hierarchical, for instance) in order to have different groups and use those as target labels. Then, for further examples, instead of running again the clustering algorithm (which is quite costly), you can run directly the classification one built over the labels provided by the groups.
Instead, if you do have the target labels, you can compare the results of a classification algorithm w.r.t. a clustering one: if two samples have the same labels it is likely they will end up in the same cluster because they are similar, and vice versa. If a clustering reflects exactly this property, it can be informative on how the label were assigned.

Q. *You have a dataset containing ten thousand examples; you applied two classification algorithms A and B and measured their performance using accuracy and standard 2/3-1/3 holdout; the accuracy over the test set was 74.2% for algorithm A and 75.5% for algorithm B. Your supervisor is not satisfied and tells you that he is not convinced that B is actually better than A.*

1. *Do you agree with your supervisor? If yes, why, if not, why not?*
2. *If you agree, propose and detail an alternative procedure to compare the two algorithms and come out with a more reliable answer.*

A. Yes, I agree with the supervisor, because the performance difference might be due to chance. What you can do is performing a paired t-Test in the following way:
For both the algorithms, you perform a k-fold crossvalidation over the whole dataset, resulting in k performance results for both the algorithms.

$$\theta_1^A ... \theta_k^A \qquad \theta_1^B ... \theta_k^B$$

Then you define:

$$\delta_i = \theta_i^A - \theta_i^B \qquad \mu = \frac{1}{k}\sum_i^k \delta_i \qquad \sigma = \sqrt{\frac{1}{k}\sum_i^k (\delta_i - \mu)^2}$$

And the hypothesis for the test:

$$H_0 : \mu = 0 \qquad H_1 : \mu \neq 0$$

If you reject H_0 with a certain confidence, then the showed results are due to chance and they have significantly different performance.
There are many alternative procedures to compare classification algorithms, in case of probabilistic algorithm one of them is ROC curve: you typically choose the one with the higher area below the curve.

Exam 29/01/2018

Q. *Your colleagues do not believe that clustering and classification techniques can be combined in several ways. What arguments/examples would you use to convince them?*

A. Answer at page 23.

Q. *You have a dataset containing ten thousand examples; you applied two classification algorithms A and B and measured their performance using accuracy and standard 2/3-1/3 holdout; the accuracy over the test set was 74.2% for algorithm A and 75.5% for algorithm B. Your supervisor is not satisfied and tells you that he is not convinced that B is actually better than A.*

1. *Do you agree with your supervisor? If yes, why, if not, why not?*
2. *If you agree, propose and detail an alternative procedure to compare the two algorithms and come out with a more reliable answer.*

A. Answer at page 23.

Exam 16/05/2018

Q. *Consider a data set that has been imported into a pandas dataframe called data and the output of the shape and describe commands below: You are asked to preprocess and explore the data using the six*

```
In [23]: data.shape
Out[23]: (400, 17)
```

```
In [21]: data.describe()
```

Out[21]:

	age	bp	sg	al	wbcc	rbcc
count	391.000000	388.000000	353.000000	354.000000	294.000000	269.000000
mean	51.483376	76.469072	1.017408	1.016949	8406.122449	4.707435
std	17.169714	13.683637	0.005717	1.352679	2944.474190	1.025323
min	2.000000	50.000000	1.005000	0.000000	2200.000000	2.100000
25%	42.000000	70.000000	1.010000	0.000000	6500.000000	3.900000
50%	55.000000	80.000000	1.020000	0.000000	8000.000000	4.800000
75%	64.500000	80.000000	1.020000	2.000000	9800.000000	5.400000
max	90.000000	180.000000	1.025000	5.000000	26400.000000	8.000000

```
In [22]: data.describe(exclude=[np.number])
```

Out[22]:

	rbc	pc	pcc	ba	htn	dm	cad	appet	pe	ane	class
count	248	335	396	396	398	398	398	399	399	399	400
unique	2	2	2	2	2	2	2	2	2	2	2
top	normal	normal	notpresent	notpresent	no	no	no	good	no	no	ckd
freq	201	259	354	374	251	261	364	317	323	339	250

most adequate preprocessing and visualization techniques among the ones considered during the course. For each technique specify (1) the name of the technique as presented in the course, (2) the attributes to which it is applied and (3) your goal for applying it (what are you trying to understand or what is expected outcome of the preprocessing).
Note that (1) the order is important, the techniques should be listed in the order you plan to use them; (2) the same technique cannot be specified twice.

Technique	Imputation of missing values
List of attributes	All attributes except rbc, pc, wbcc, rbcc
Goal	This technique is used to deal with the fact that some tuples in the dataset are missing some values. You can use the mean for numerical attributes (or a regression) and the mode for categorical ones.

Technique	Histograms and boxplots
List of attributes	Numerical variables
Goal	These plots shows me the distribution of numerical variables. In particular, boxplots show also the percentiles and highlights possible outliers.

Technique	Normalization (log(1+x))
List of attributes	al (and eventually other skewed variables pointed out by histograms)
Goal	Normalization is in general a good practice when you have a numerical variable that is quite skewed. In this case, "al" is quite skewed.

Technique	Winsorization
List of attributes	bp, wbcc if not dropped
Goal	This technique aims to eliminate outliers, by setting them to the 75th percentile (can be another, but in general it is done with this). In this case we can see that bp and wbcc have a max which is quite higher than the 75th percentile, therefore it is reasonable to apply winsorization on this numerical variables.

Technique	Clustermap analysis
List of attributes	Numerical variables
Goal	The Clustermap analyzes the relationship between the numerical variables: this may indicate may which are are the most correlated and, eventually, inform me about which variables can be dropped.

Technique	PCA
List of attributes	Numerical variables
Goal	In order to easily visualize the data, you can apply PCA with 2 or 3 principal components, plot them in a 2D or 3D space and see if some patterns can be found.

+ outliers
+ new attributes (features) if there are, e.g., timestamps