Web Mining and Retrieval MidTerm
Examples (a.a. 2018-2019)

April 2019

Docente: R. Basili

Answer to the following questions by marking answers that you assume are correct. Available time: $3 \times n$ *minutes. In the evaluation, every wrong choice lower the score.*

Gruppi: 2/10,1/8,1/2,1/1,1/5

1. (M) Selezionare **tutte** le affermazioni corrette sull'algoritmo *K-Means*
(A) E' un algoritmo di tipo gerarchico. [−0]
(B) E' un algoritmo di tipo agglomerativo. [−0]
(C) Genera un numero k di cluster. [−0]
(D) E' indipenente dalla scelta dei *seed* iniziali. [−0]
(E) Ha una complessita' lineare. [−0]

2. (M) Select **all** the correct statements about the algorithm known as *K-Means*
(A) It is a hierarchical type algorithm. [−0]
(B) It is an agglomerative algorithm. [−0]
(C) It generates a fixed number of k cluster. [−0]
(D) It is independent from the selection of the initial *seeds*. [−0]
(E) It is characterized by a linear complexity. [−0]

3. (M) Selezionare **tutte** le affermazioni corrette sull'algoritmmo *Hierarchical Agglomerative Clustering*
(A) La *Single Link Similarity* e' sempre meglio della *Complete Link Similarity*. [−0]
(B) Piu' esecuzioni riproducono esattamente lo stesso insieme di cluster. [−0]
(C) Genera un numero k di cluster. [−0]
(D) E' un algoritmo di tipo gerarchico [−0]
(E) Possono essere utilizzate differenti metriche per il calcolo della distanza tra due elementi. [−0]

4. (M) Select **all** the correct statements about the algorithm known as *Hierarchical Agglomerative Clustering*
(A) The *Single Link Similarity* metrics is always better than the *Complete Link Similarity*. [−0]
(B) Repeated executions result in exactly the same number of clusters. [−0]
(C) It generates a fixed number of k clusters. [−0]
(D) It is a hierarchical algorithm. [−0]
(E) Different metrics can be applied to compute the distance between two elements. [−0]

5. Data una classe C_i ed il classificatore seguente (Rocchio) ,
$(\sum_{\vec{d}\in C_i} \frac{\beta}{|C_i|}\vec{d} - \sum_{\vec{d}\notin C_i} \frac{\gamma}{|\overline{C_i}|}\vec{d}) \cdot \vec{x} - \tau > 0$, con la soglia $\tau > 0$
segnalare la affermazione corretta?
(A) La sua funzione di separazione è un polinomio di grado $n > 2$. $[-0]$
(B) La sua funzione di separazione è un iperpiano che ha il massimo margine tra gli iperpiani di separazione. $[-0]$
(C) La sua funzione di separazione è un iperpiano il cui gradiente è la differenza tra la media degli esempi positivi e la media degli esempi negativi. $[-0]$
(D) La sua funzione di separazione è un vettore che è la sommatoria di tutti i vettori rappresentanti i documenti positivi. $[-0]$

6. Given a class C_i and the following (Rocchio) classifier,
$(\sum_{\vec{d}\in C_i} \frac{\beta}{|C_i|}\vec{d} - \sum_{\vec{d}\notin C_i} \frac{\gamma}{|\overline{C_i}|}\vec{d}) \cdot \vec{x} - \tau > 0$, with threshold $\tau > 0$
mark the correct statement among the following ones:
(A) The corresponding separating function is a polynomial of degree $n > 2$. $[-0]$
(B) The corresponding separating function is an hyperplane characterized by the maximal margin among all the separating hyperplanes. $[-0]$
(C) The corresponding separating function is an hyperplane whose gradient is the difference between the average of positive examples and the average of the negative examples. $[-0]$
(D) The corresponding separating function is the vector sum of all vectors representing positive documents. $[-0]$

7. Quale delle seguenti affermazioni definisce in termini più generali la *Automatic Text Categorization*?
(A) Dati degli esempi si determina l'iperpiano di separazione che li divide. [−0]
(B) Dati degli esempi positivi e negativi si apprende l'iperpiano per separarli; questo verrà usato per classificare anche i nuovi documenti. [−0]
(C) Dati degli esempi positivi e negativi si apprende la funzione di separazione; questa verrà usata per classificare anche i nuovi documenti. [−0]
(D) Dato un insieme di training e uno di testing si apprende la funzione di separazione sul training e sul testing; in particolare sul testing si stimano i parametri e le prestazioni. [−0]

8. Which one of the following statements corresponds to the most general definition for *Automatic Text Categorization*?
(A) Given some examples the hyperplane that separates them is determined by training. [−0]
(B) Given some positive and negative examples the hyperplane that separates them is first acquired; then it is used to classify also the new incoming documents. [−0]
(C) Given some positive and negative examples the separating function is first acquired; it will be used to classify the new incoming documents. [−0]
(D) Given a training and a test set, the separating function is first acquired on the training and test set; in particular, testing is used to estimate the parameters and the performances. [−0]

9. Cosa s'intende per n-fold cross validation?
(A) Dati degli esempi di training e di testing si apprendono i modelli sul training e si testano sul testing. [−0]
(B) Dati degli esempi di training e di testing si apprendono i modelli sul testing e si testano sul training. [−0]
(C) Si divide il corpus di documenti in n parti; a rotazione una viene usata per il testing e $n-1$ sono usate per il training. [−0]
(D) Si divide il training in n parti e si addestra il classificatore n volte; ogni volta si misura la performance sul test-set. [−0]

10. What does *n-fold cross validation* mean?
(A) Given the training and testing examples, models are acquired from training and testing, and then are tested on the test set. [−0]
(B) Given the training and testing examples, models are acquired from the testing data, and then are tested on the training set. [−0]
(C) The corpus is partitionaed in equal n parts; at each step, one partition is used a testing set and the remaining $n-1$ are used for the training. [−0]
(D) The training set is divided in n parts e the classifier is trained n times; at every time, the performance is measured over the test-set. [−0]

11. Data una classe C_i ed il classificatore seguente (Rocchio) ,
$(\sum_{\vec{d} \in C_i} \frac{\beta}{|C_i|} \vec{d} - \sum_{\vec{d} \notin C_i} \frac{\gamma}{|\overline{C_i}|} \vec{d}) \cdot \vec{x} - \tau > 0$, con la soglia $\tau > 0$
qual'è una differenza tra Rocchio e Naive Bayes?
(A) Naive Bayes usa l'assunzione di indipendenza delle parole mentre in Rocchio tale assunzione non viene fatta. [−0]
(B) Naive Bayes è un classificatore lineare mentre Rocchio è probabilistico. [−0]
(C) Per ogni documento ed una classe, Rocchio fornisce in output un valore empirico (*score*) mentre Naive Bayes una probabilità. [−0]
(D) Nessuna differenza perchè entrambi sono definiti sul Vector Space Model [−0]

12. Given a class C_i ed the following (Rocchio) classifier,
$(\sum_{\vec{d} \in C_i} \frac{\beta}{|C_i|} \vec{d} - \sum_{\vec{d} \notin C_i} \frac{\gamma}{|\overline{C_i}|} \vec{d}) \cdot \vec{x} - \tau > 0$, with threshold $\tau > 0$
what is difference between Rocchio and the Naive Bayes methods?
(A) Naive Bayes imposes the independence property among words while in Rocchio such assumption is not made. [−0]
(B) Naive Bayes is a linear classifier while Rocchio is a probabilistic one. [−0]
(C) For every incoming document and a class, Rocchio outputs an empirical *score* while Naive Bayes produces a probability. [−0]
(D) There is no difference among them (except the formula), as they are both defined in the (same) underlying Vector Space. [−0]

13. (M) Segnalare **tutte** le affermazioni errate tra le seguenti:
(A) Gli algoritmi di learning generano esempi da classificare sulla base dei processi di addestraemnto. [−0]
(B) Un algoritmo di learning basato su esempi apprende una funzione efficiente di classificazione. [−0]
(C) Un processo markoviano non costituisce un modello di apprendimento per alcun problema di classificazione binario. [−0]
(D) Nessuna delle altre affermazioni e' errata. [−0]
(E) E' generativo un algoritmo di learning che genera gli esempi positivi sulla base della probabilitá delle loro proprieta' salienti. [−0]

14. (M) Mark **all** he wrong statements among the following ones:
(A) Learning algorithms generate example to be classified on the basis of learning processes. [−0]
(B) An example-based learning algorithm induces an efficient classification function. [−0]
(C) A Markov process does not non provide any learning model for a binary classification problem. [−0]
(D) None of the other statements is wrong. [−0]
(E) It is generative a learning algorithm that generates positive examples on the basis of the probability of their salient features. [−0]

15. (M) Segnalare **tutte** le affermazioni corrette tra le seguenti:
(A) Un modello generativo a' basato su un insieme di variabili stocastiche e su un insieme di dipendenze che le legano alla probabilita' totale del fenomeno da apprendere. [−0]
(B) I modelli generativi costituiscono esempi di algoritmi di Semi-Supervised Learning e sono necessari quando non esistono abbastanza dati di addestramento. [−0]
(C) Le catene di Markov nascoste (Hidden Markov Models) costituiscono modelli generativi per task di sequence labeling. [−0]
(D) Nessuna delle altre affermazioni e' corretta. [−0]
(E) E' generativo un algoritmo che classifica gli esempi positivi sulla base della probabilita' delle sue proprieta' salienti. [−0]

16. (M) Mark **all** the correct statements among the following ones:
(A) A generative model is based on a set of stochastic variables and a set of dependencies that contrain them to the total probability of the targeted phenomenon. [−0]
(B) Generative models are examples of Semi-Supervised Learning algorithms and become necessary when the size of the training data set is not enough [−0]
(C) Hidden Markov models correspond to generative models for sequence labeling tasks. [−0]
(D) None of the other statements is correct. [−0]
(E) It is generative a learning algorithm that generates positive examples on the basis of the probability of their salient features. [−0]

17. Dire se i vettori $\underline{2}$ e $\underline{3}$ compaiono nello stesso cluster finale prodotto da un algoritmo di tipo k-mean (con $k = 2$) basato su una metrica di tipo euclideo, applicato al seguente insieme:

Vector	x-dim	y-dim
0	0	0
1	1	4
2	2	3
3	3	6
4	5	0
5	6	2
6	0	1
7	8	5
8	6	9

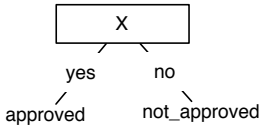
Si assuma che i due seed siano costituiti dai vettori $\underline{1}$ e $\underline{5}$.
(A) Si' [−0]
(B) No. [−0]
(C) Si ma solo se non considero una metrica di tipo *single-link* [−0]
(D) Dipende dai seed [−0]

18. Tell if vectors $\underline{2}$ and $\underline{3}$ appear in the same final cluster obtained by applying a k-mean algorithm (with $k = 2$) based on the euclidean metrics when applied to the above data set.
Assume that the seed correspond to vectors $\underline{1}$ and $\underline{5}$.
(A) Yes [−0]
(B) No. [−0]
(C) Yes, but only by avoiding the application of a metrics such as *Single-link* [−0]
(D) It depends on the seeds [−0]

19. Given the following training set

Age	**Has_job**	**Own_house**	**Class**
young	no	no	not_approved
young	$Y1$	yes	approved
middle	yes	no	approved
middle	no	$Y2$	not_approved
old	yes	yes	approved

in which cases the following decision tree is sufficient for the problem according to the C4.5 algorithm (by Quillan):



(A) X=Has_job if $Y1$=yes and $Y2$=no. [−0]
(B) X=Own_house only if $Y1$=yes and $Y2$=no. [−0]
(C) X=Own_house only if $Y1$=yes and for each value of $Y2$. [−0]
(D) X=Age only if $Y2$=yes and for every value of $Y1$. [−0]

20. A Hidden Markov Model is described by a symbol $O = \{Cof, Tea, Cap\}$ and state $S = \{CP, TP\}$ vocabulary, whereas emission matrix is

$$\mathbf{E} = \begin{pmatrix} p(Cof|CP) & p(Tea|CP) & p(Cap|CP) \\ p(Cof|TP) & p(Tea|TP) & p(Cap|TP) \end{pmatrix} = \begin{pmatrix} .65 & .15 & .2 \\ .2 & .8 & 0 \end{pmatrix}$$

and transitions are defined by:

$$\mathbf{T} = \begin{pmatrix} p(CP|CP) & p(TP|CP) \\ p(CP|TP) & p(TP|TP) \end{pmatrix} = \begin{pmatrix} .5 & .5 \\ .3 & .7 \end{pmatrix}$$

If the initial state is always CP what is the most likely state sequence associated to the sequence (Cof, Tea)?
(A) It is not possible to estimate the most likely sequence. $[-0]$
(B) (CP, TP) $[+0]$
(C) (TP, TP) $[-0]$
(D) (CP, CP) $[-0]$
(E) None of the other statements is correct. $[-0]$