Politecnico di Milano
School of Industrial and Information Engineering

Data Mining and Text Mining

July 13, 2019

NAME

CODICE PERSONA/ID

GENERAL INSTRUCTIONS

- **Answers must be clearly written inside the answer box designated for each problem.**
- **All the answers must be adequately motivated.**
- **Pencils are not allowed. The exam consists of 6 sheets of paper. It must be returned with all the 6 sheets. No any other sheet can be added. No sheet can be removed.**
- **This is a closed-book/closed-notes exam.**
- **Only non-programmable calculators are allowed. When asked, logarithms should be computed using base 2**
- **Notes/books/mobile phones are not allowed.**
- **If you are caught using forbidden material, the exam will immediately end and an RP grade will be recorded; then, your Data Mining exam will consist of an oral examination from then on.**

SCORING

- **A problem left unsolved will amount to zero points.**
- **A completely wrong solution will amount to -3 points**

STUDENTS WHO DID THE COURSE PROJECT HAVE 1:40h TO SOLVE PROBLEMS 1, 2, 3, AND 4

ALL THE OTHER STUDENTS HAVE 2:20h TO SOLVE ALL THE SIX PROBLEMS

COURSE PROJECT SCORE

FINAL TIME

GRADES

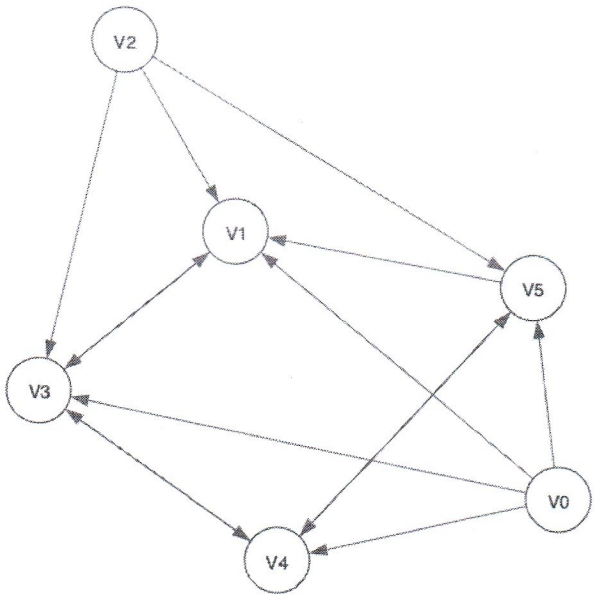
1	2	3
4	5	6

Problem 1 (6 points). Write the pseudocode of Mean Shift Clustering given the data D containing N data points described by d attributes.

Notes: Write one instruction per line. The function definition is specified using Python notation just for your convenience.

	Pseudocode of MeanShiftClustering
1	`def MeanShiftClustering(D):`
2	`  choose a window size (bandwidth)`
3	`  while some points are not assigned to a cluster`
4	`  do`
5	`  Choose the initial location of the search window`
6	`  Center the search window at the location just computed`
7	`  while the process did not converge`
8	`  Assign all the visited points to the same`
9	`  If the final point is nearby to another cluster, join the two`
10	
11	
12	
13	
14	
15	
16	
17	
18	
19	
20	
21	
22	
23	

Problem 2 (6 points). Consider the graph below, identify the communities using a support threshold of 3. Report the final result in the corresponding box.



First we compute the itemsets associated to each node:

{v1, v3, v4, v5}
{v3}
{v1, v3, v5}
{v1, v4}
{v3, v5}
{v1, v4}

Next we apply apriori with a minimum support of 3 and get the following frequent itemsets,

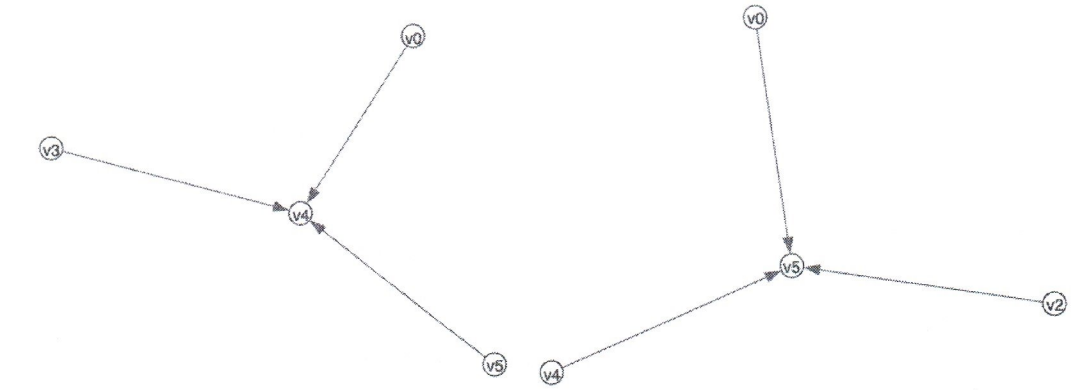
{v4} 0.5000000 3
{v5} 0.5000000 3
{v3} 0.6666667 4
{v1} 0.6666667 4
{v1,v4} 0.5000000 3
{v3,v5} 0.5000000 3

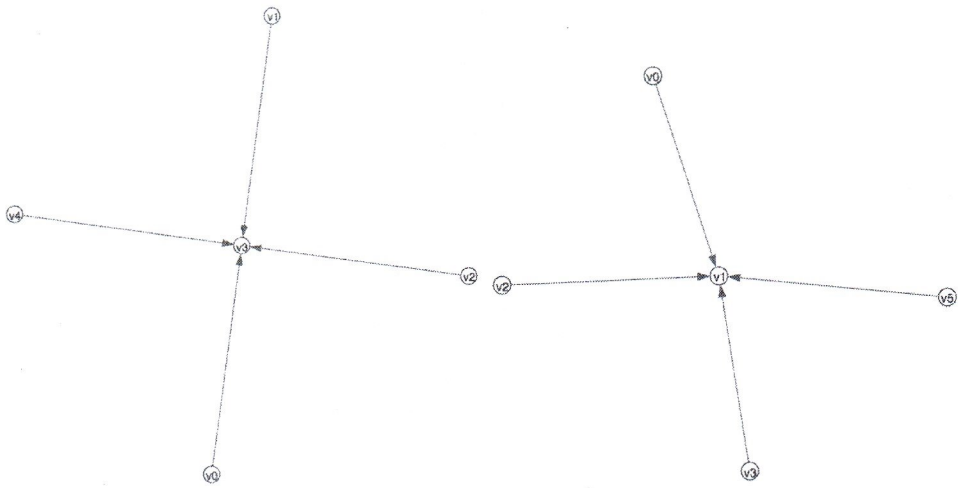
We identify the communities comprising,

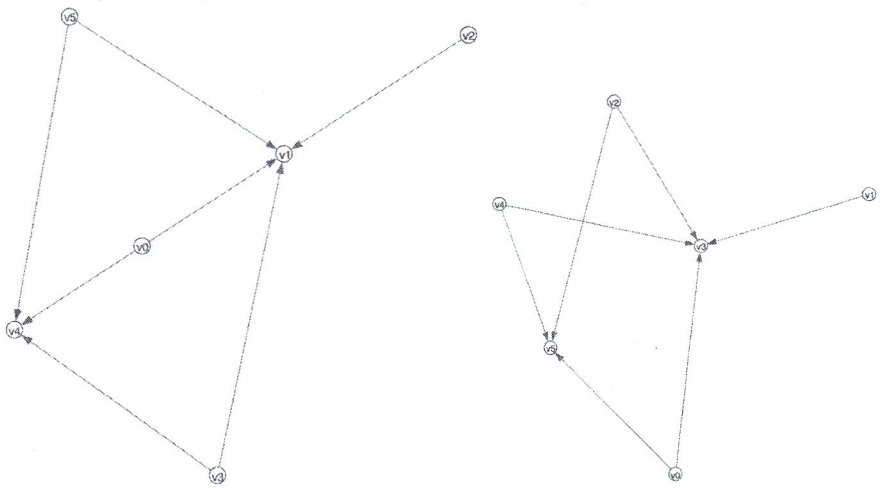
- All the nodes with an edge going into v4
- All the nodes with an edge going into v5
- All the nodes with an edge going into v3
- All the nodes with an edge going into v1
- All the nodes with an edge going either into v1 or v4
- All the nodes with an edge going either into v3 or v5

Problem 2 (Continued).

Note that the topology has been automatically generated so it might appear different from what you draw.







Problem 3 (6 points). Answer the following questions using the corresponding boxes:

Define Permutation Importance and why it is used.
It is a method to evaluate how important a single feature is in a trained model.

	Pseudocode to evaluate Permutation Importance
1	Train a model on a dataset.
2	Shuffle the values in a single column (e.g., the values of a single features).
3	Apply the model both to the original data and to shuffled data.
4	Compute the feature importance as the loss of performance when the model is applied to the shuffled data.
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	
15	

Problem 3 (continued).

Define Partial Dependence Plots and explain why they are used

Partial dependence plots show how a feature affects predictions.

	Pseudocode to compute Partial Dependence Plots
1	Train a model on data.
2	Run the model on each sample by changing only the value of target feature(s) (from a set of values in the feature range).
3	Compute and plot the model output for each value of the feature.
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	
15	

Problem 4 (6 points). Given the dataset below, where "value" is the class attribute,

- Compute the Naïve Bayes classifier using the Laplace Estimator and report the results in the corresponding tables
- Compute the predicted class for (good, small, no)

location	size	pets	value
good	small	yes	high
good	big	no	high
good	big	no	high
bad	medium	no	medium
good	medium	only cats	medium
good	small	only cats	medium
bad	medium	yes	medium
bad	small	yes	low
bad	medium	yes	low
bad	small	no	low

Counts (Integer Values)

Attribute Location	Class High	Class Medium	Class Low	Attribute Size	Class High	Class Medium	Class Low
Good	3+1	2+1	0+1	Small	1+1	1+1	2+1
Bad	0+1	2+1	3+1	Medium	0+1	3+1	1+1
				Big	2+1	0+1	0+1
Attribute Pets	Class High	Class Medium	Class Low				
Yes	1+1	1+1	2+1				
No	2+1	1+1	1+1				
Only Cats	0+1	2+1	0+1				

Problem 4 (Continued.)

Frequency Values

Attribute Location	Class High	Class Medium	Class Low	Attribute Size	Class High	Class Medium	Class Low
Good	4/5	3/6	1/5	Small	2/6	2/7	3/6
Bad	1/5	3/6	4/5	Medium	1/6	4/7	2/6
				Big	3/6	1/7	1/6
Attribute Pets	Class High	Class Medium	Class Low				
Yes	2/6	2/7	3/6				
No	3/6	2/7	2/6				
Only Cats	1/6	3/7	1/6				

Computation to Predict the class of (good, small, no)

For each class we have to compute,

P(class|example) = P(good|class) P(small|class) P(no|class) P(class)

The predicted class will be the one with the highest probability.

Problem 5 (6 points). You and your boss attended a presentation of a data scientist who applied several classification methods to a data set containing 20000 data points. The data scientist first applied hold out splitting the data in train and test set. Then applied the 7 classification methods listed below and computed performance as the Mean Absolute Error (MAE) on the test set. The presenter concluded that "Bayesian Ridge" was the best method of the seven considered since it reached the lowest MAE value. At the end your boss wants to know whether you liked the presentation or would apply a different procedure to compare the seven approaches.

	Regressor	MAE
0	linear	7.248779
1	lasso	7.264249
2	ridge	7.248766
3	elastic_net	7.262760
4	AdaBoost	15.213487
5	bayesian ridge	7.247135
6	Lasso LARS	30.632625
7	xgb	12.283350

It would be better to apply k-fold crossvalidation instead of just comparing algorithms based on one single value, especially considering that the dataset contains a rather limited number of data points.

We can compare crossvalidation results using t-test and since we are making several comparisons we would be better off using Bonferroni adjustment to make the comparison more robust.

Problem 6 (3 points). You are participating to a round table where several experts discuss the techniques based on decision trees. Prof. JohnTree states that the ID3/C4.5 decision tree algorithms (the ones seen during the course) are guaranteed to find an optimal tree (that is, a tree that best classifies the training tuples over all possible trees). MartinBoost begs to differ and states that the only approach that guarantees optimality are boosting algorithms like xgboost. RudyForest joins the discussion and declares that Random Forests are the only possible mean to guarantee optimality. Comment the three statements below and explain which one in your opinion is the most correct one.

JohnTree

MartinBoost

RudyForest

Additional Comments