List of Exercises: Data Mining 1
November 28th, 2018

1. We trained a model on a two-class balanced dataset using five-fold cross validation. One person calculated the performance of the classifier by measuring the accuracy in each fold and then averaging the results. Another person summed up all TP, FP, TN, FN, and then calculated the accuracy from these numbers. Will the results be different? If they are, which result is more trustable?

2. Draw the full decision tree for the parity function of four boolean attributes, A, B, C and D. Is it possible to simplify the tree?

3. The data instances of Table 1 are sorted by decreasing probability value for the positive class (P), as returned by a classifier. For each instance, compute the values for the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), for a threshold of 0.5. Compute the true positive rate (TPR) and false positive rate (FPR). Plot the ROC curve for the data.

Table 1: Table for question (6)

Instance	Class	Predicted Probability
1	P	0.95
2	N	0.85
3	P	0.78
4	P	0.66
5	N	0.60
6	P	0.55
7	N	0.53
8	N	0.52
9	N	0.51
10	P	0.40

4. Suppose that we want to select between two prediction models, M_1 and M_2. We have performed 10 rounds of 10-fold cross-validation on each model, where the same data partitioning in round i is used for both M_1 and M_2. The error rates obtained for M_1 are 30.5, 32.2, 20.7, 20.6, 31.0, 41.0, 27.7, 26.0, 21.5, 26.0. The error rates for M_2 are 22.4, 14.5, 22.4, 19.6, 20.7, 20.4, 22.1, 19.4, 16.2, 35.0. Comment on whether one model is significantly better than the other considering a significance level of 1%.

5. According to Han and Kamber (cf. reference book) what is the difference between classification and prediction?

6. Give three advantages and three disadvantages of Decision Tree models and of Support Vector Machines.

7. Knowing that 30% of the portuguese population has hypercholesterolaemia, that 60% of the same population is overweight, and that 80% of the patients with hypercholesterolaemia are overweight, what is the probability of an overweighted person having hypercholesterolaemia? Explicitly present your reasoning.

8. Considering the qualitative Bayesian model shown in Figure 1, and class variable LungCancer, what are the relevant nodes necessary to compute the probability of the class given the other variables?

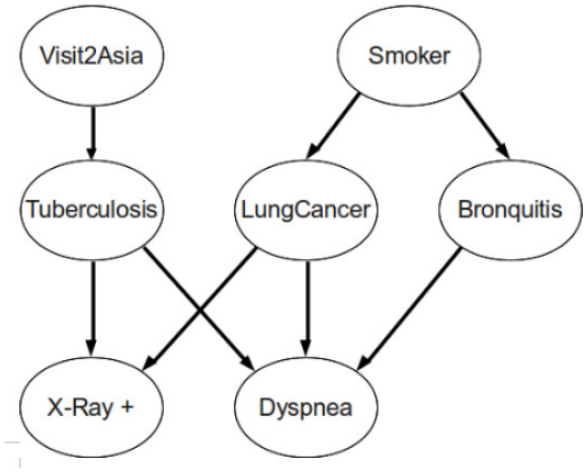

Figure 1: Bayes Network for question (8)

9. Given the dataset of Table 2 and class variable C, what would be the resulting probability table after Laplace correction?

Table 2: Table for question (9)

A	B	C
T	T	T
T	F	T
F	F	T
F	T	F
T	F	F
F	F	F

10. For dataset shown in Table 3, show the Naive Bayes network topology (assume the class variable is Flu), and calculate $P(Flu = Yes \mid Chills \wedge Fever)$. (This Table was created by André Rodrigues).

Table 3: Synthetic dataset from clinic patients.

Runny nose	Headache	Chills	Fever	Flu
no	strong	true	false	no
no	strong	true	true	no
little	**strong**	**true**	**false**	**yes**
high	**mild**	**true**	**false**	**yes**
high	**no**	**false**	**false**	**yes**
high	no	false	true	no
little	**no**	**false**	**true**	**yes**
no	mild	true	false	no
no	**no**	**false**	**false**	**yes**
high	**mild**	**false**	**false**	**yes**
no	**mild**	**false**	**true**	**yes**
little	**mild**	**true**	**true**	**yes**
little	**strong**	**false**	**false**	**yes**
high	mild	true	true	no

11. Consider that a model is trained and tested on the same dataset. What can you say about the

performance of this classifier on new unseen data?

12. Consider that you are given a train-test split of a dataset. Assume that you create successively better models with respect to the testset when training over and over again with the same training set. What can you say about the performance of the final model?

13. What are the main differences between Support Vector Machines and Neural Networks?