Data Science And Database Technology

Homework #2 – Data mining

Objective

Exploit data mining classification algorithms to analyze a real dataset using the RapidMiner machine learning tool.

Dataset

The Breast dataset (Breast.xls, available on the course website) collects medical data about patients who have contracted breast cancer. Each dataset record corresponds to a different patient and consists of a set of patient, treatment, and disease characteristics (e.g., the patient age, the tumor size). Depending upon the tumor is a recurrent or nonrecurrent event in patient life, each record is also labeled with class label "Recurrence events" or "No recurrence events". Such a data attribute, which will be used as class attribute throughout the homework, is reported as the last record attribute.

The complete list of dataset attributes is reported below.

(1) Age
(2) Menopause
(3) Tumor-size
(4) Inv-nodes
(5) Node-caps
(6) Deg-malig
(7) Breast
(8) Breast-quad
(9) Irradiat
(10) **class (class attribute)**

Context

Oncologists want to predict the property of recurrence or not of breast tumors according to patient, tumor, and treatment characteristics. To this purpose, they exploit three different classification algorithms: a decision tree (Decision Tree) and a Bayesian classifier (Naïve Bayes), and a distance-based classifier (K-NN). The Breast dataset is used to train classifiers and to validate their performance.

Questions

Answer to the following questions:

1. Learn a Decision Tree from the whole dataset by setting the minimum gain threshold to 0.01, while keeping the default configuration for all the other parameters. (a) Which attribute is deemed to be the most discriminative one for class prediction? (b) What is the height of the Decision Tree generated? (b) Find a pure partition in the Decision Tree and report a screenshot that shows the example identified.
2. Analyze the impact of the minimal gain (using the gain ratio splitting criterion) and maximal depth parameters on the characteristics on the Decision Tree model learnt from the whole dataset (keep the default configuration for all the other parameters). Report at least 5 different screenshots showing Decision Trees (or portions of them) generated with different configuration settings.
3. Performing a 10-fold Stratified Cross-Validation, what is the impact the maximal gain and maximal depth parameters on the average accuracy achieved by Decision Tree? Report at least 5 screenshots showing the confusion matrices achieved using different parameter settings (consider *at least* all the configurations used to answer Question 2). Keep the default configuration for all the other parameters.
4. Considering the K-Nearest Neighbor (K-NN) classifier and performing a 10-fold Stratified CrossValidation, what is the impact of parameter K on the average classifier accuracy? Report at least 5 screenshots showing the confusion matrices achieved using different K parameter values. Perform a 10-fold Stratified Cross-Validation with classifier Naïve Bayes. Does K-NN perform on average better or worse than the Naïve Bayes classifier on the analyzed data? Report a screenshot showing the confusion matrix achieved by Naïve Bayes on the analyzed dataset.
5. Analyze the Correlation Matrix to discover pairwise correlations between data attributes. Report a screenshot showing the correlation matrix achieved. (a) Does the Naïve independence assumption actually hold for the Breast dataset? (b) Which is the pair of most correlated attributes?

Assignment

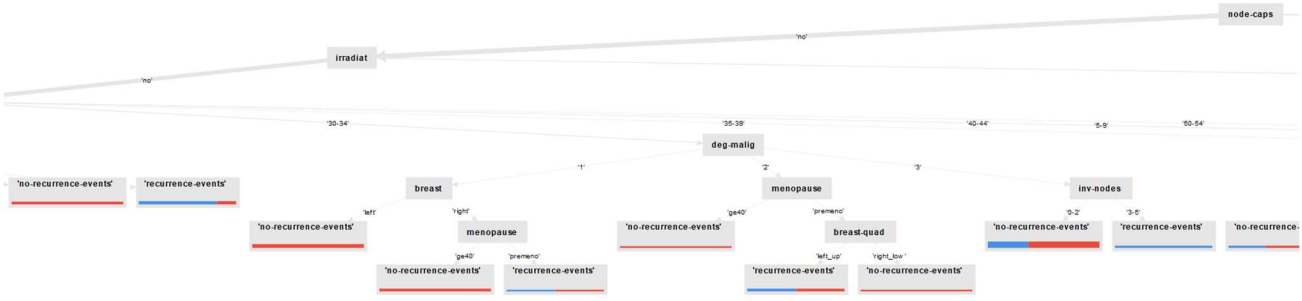
Write a 4-5 page report containing the answers to the above questions.

Answers

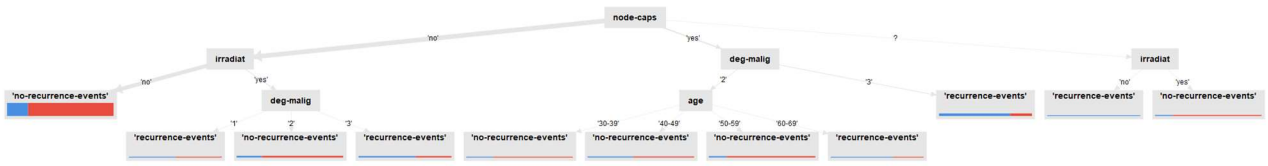
1) Decision tree
   a) The most discriminative attribute is the one used as a root node. In this case it is ***"node-caps".***
   b) The height of a Tree is the maximum depth, so the number of nodes needed to reach the furthest leaf node, starting from the root. In this case the height is **6** (not counting the root node).
   c) The screenshots below represent a pure partition, that is a node that is no longer splittable because it already perfectly describes a class label.

*Left image - Leaf node represents only records with class label **"recurrence-events"***
*Right image - Leaf node represents only records with class label **"no-recurrence-events"***

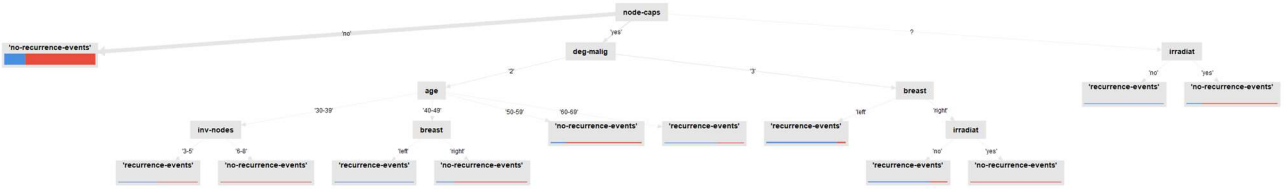
2) Decision Tree for five different configurations



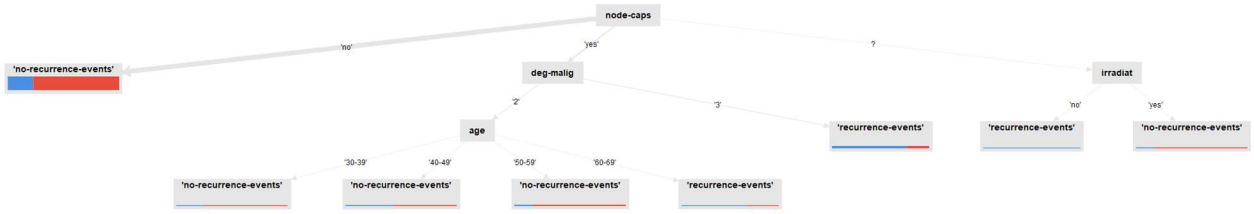
1) Minimal Gain: 0.01 - Max Depth: 10 (screenshot of a part of the tree)



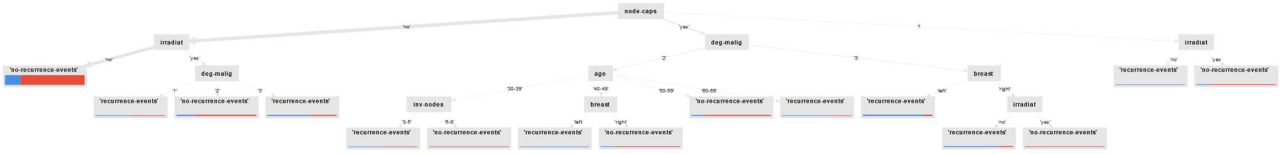
2) Minimal Gain: 0.01 - Max Depth: 4 (full tree)



3) Minimal Gain: 0.05 - Max Depth: 10 (full tree)



4) Minimal Gain: 0.05 - Max Depth: 4 (full tree)



5) Minimal Gain: 0.025 - Max Depth: 5 (full tree)

3) Confusion matrices for five different configurations (with Decision Tree):
Decreasing the value of the max depth will stop the splitting process of the tree at a certain depth. If we set this too low, we lose accuracy.
Decreasing the value of the minimal gain, instead, produce a higher accuracy because the tree will continue the splitting process for a longer period. This means we develop a bigger tree and gain in accuracy.

accuracy: 67.48% +/- 6.59% (micro average: 67.48%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	37	45	45.12%
pred. 'no-recurrence-events'	48	156	76.47%
class recall	43.53%	77.61%	

1) Minimal Gain: 0.01 - Max Depth: 10

accuracy: 71.00% +/- 6.95% (micro average: 70.98%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	31	29	51.67%
pred. 'no-recurrence-events'	54	172	76.11%
class recall	36.47%	85.57%	

2) Minimal Gain: 0.01 - Max Depth: 4

accuracy: 70.64% +/- 6.20% (micro average: 70.63%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	24	23	51.06%
pred. 'no-recurrence-events'	61	178	74.48%
class recall	28.24%	88.56%	

3) Minimal Gain: 0.05 - Max Depth: 10

accuracy: 71.33% +/- 6.58% (micro average: 71.33%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	26	23	53.06%
pred. 'no-recurrence-events'	59	178	75.11%
class recall	30.59%	88.56%	

4) Minimal Gain: 0.05 - Max Depth: 4

accuracy: 71.72% +/- 5.81% (micro average: 71.68%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	31	27	53.45%
pred. 'no-recurrence-events'	54	174	76.32%
class recall	36.47%	86.57%	

5) Minimal Gain: 0.025 - Max Depth: 5

4) K-NN accuracy study: if the value of K is low, the algorithm is often highly affected from noise and outliers. In the study below we can also see that the accuracy is lower when classifying test data. When we increase the value of K we can see that the accuracy is better, but we must pay attention to not set K way too high. That could lead to prioritize bigger categories, loosing information of smaller ones that contains few samples (because they would be always “outvoted”).

accuracy: 73.77% +/- 5.98% (micro average: 73.78%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	26	16	61.90%
pred. 'no-recurrence-events'	59	185	75.82%
class recall	30.59%	92.04%	

1) K value = 5

accuracy: 66.44% +/- 7.28% (micro average: 66.43%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	30	41	42.25%
pred. 'no-recurrence-events'	55	160	74.42%
class recall	35.29%	79.60%	

2) K value = 1

accuracy: 65.73% +/- 8.62% (micro average: 65.73%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	39	52	42.86%
pred. 'no-recurrence-events'	46	149	76.41%
class recall	45.88%	74.13%	

3) K value = 2

accuracy: 74.84% +/- 6.23% (micro average: 74.83%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	25	12	67.57%
pred. 'no-recurrence-events'	60	189	75.90%
class recall	29.41%	94.03%	

4) K value = 7

accuracy: 75.20% +/- 5.18% (micro average: 75.17%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	23	9	71.88%
pred. 'no-recurrence-events'	62	192	75.59%
class recall	27.06%	95.52%	

5) K value = 9

a) The screenshot below represents the confusion matrix of the Naïve Bayes classification algorithm. If we compare it to the K-NN study, we can conclude that the K-NN algorithm yielded a better accuracy (respect to Naïve Bayes) 3 out of 5 times.
So, **K-NN performs better on average then Naïve Bayes.**

accuracy: 72.45% +/- 7.70% (micro average: 72.38%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	41	35	53.95%
pred. 'no-recurrence-events'	44	166	79.05%
class recall	48.24%	82.59%	

5) The correlation matrix is the following:

Attribut...	age	menopa...	tumor-s...	inv-nodes	node-ca...	deg-mal...	breast	breast-...	irradiat
age	1	0.241	-0.045	-0.001	0.052	-0.043	0.067	-0.024	-0.011
menopa...	0.241	1	0.019	-0.011	0.130	-0.161	0.077	-0.096	-0.075
tumor-size	-0.045	0.019	1	-0.131	0.058	0.133	-0.022	-0.056	-0.022
inv-nodes	-0.001	-0.011	-0.131	1	-0.465	-0.213	0.040	0.063	0.399
node-caps	0.052	0.130	0.058	-0.465	1	0.098	0.024	-0.036	-0.197
deg-malig	-0.043	-0.161	0.133	-0.213	0.098	1	-0.073	0.018	-0.074
breast	0.067	0.077	-0.022	0.040	0.024	-0.073	1	0.175	-0.019
breast-q...	-0.024	-0.096	-0.056	0.063	-0.036	0.018	0.175	1	-0.005
irradiat	-0.011	-0.075	-0.022	0.399	-0.197	-0.074	-0.019	-0.005	1

The Naïve hypothesis says that attributes are statistically independent. If this assumption were true, we would only see values very close to 0 in the correlation matrix. As we observe, though, we have values different from 0, so **the Naïve assumption is not verified for this dataset.**
Looking at the Pairwise Table and ordering by correlation (from high to low) we can see that the attributes ***"inv-nodes"*** and ***"irradiant"*** have the highest correlation (0.399), while ***"inv-nodes"*** and ***"node-caps"*** have the lower correlation (-0.465).