StuDocu.com

Klausur 29 November Ws 2016, Fragen und Antworten

Business Analytics (IN2028) (Technische Universität München)

StuDocu is not sponsored or endorsed by any college or university
Downloaded by mayk use (mehmetemre5252@hotmail.com)

Area 1: Naive Bayes [15 P]

Use a Naive Bayes model to classify the unknown instance. Round your results to 4 decimal places. For the numeric attribute the following distributions were estimated:

$P(age = x\ |low) = 0.5 * e^{-0.5x}$

$P(age = x\ |high) = 0.25 * e^{-0.25x}$

The following training set is given.

Quality	Working	Age	Class
Bad	Yes	2	High
Bad	Yes	2	High
Good	Yes	4	High
Good	No	1	Low
Bad	Yes	3	High
Bad	Yes	1	Low
Good	Yes	2	Low
Good	Yes	2	Low
Good	Yes	1	Low
Good	Yes	3	Low

Unknown Instance

Quality	Working	Age	Class
Bad	No	3	?

Hint: Give attention on the zero-frequency problem.

Downloaded by mayk use (mehmetemre5252@hotmail.com)

Solution Area 1

		Class	
		High	Low
Quality	Bad	**¾**	**1/6**
	Good	¼	5/6
	∑	1	1
Working	Yes	4/4	5/6
	No	**0/4**	**1/6**
	∑	1	1
Age	1	0.195	0.3
	2	0.152	0.184
	3	**0.118**	**0.112**
	4	0.092	0.068

Zero-frequency →

		Class	
		High	Low
Quality	Bad	4/6	2/8
	Good	2/6	6/8
	∑	1	1
Working	Yes	5/6	6/8
	No	1/6	2/8
	∑	1	1
Age	1	0.195	0.3
	2	0.152	0.184
	3	**0.118**	**0.112**
	4	0.092	0.068

Class	High	4/10
	Low	6/10
	∑	1

$$P(\dots|High) = \frac{4}{10} * \frac{4}{6} * \frac{1}{6} * 0.118 = 0.00524$$

$$P(\dots|Low) = \frac{6}{10} * \frac{2}{8} * \frac{2}{8} * 0.112 = 0.0042$$

$$P_N(\dots|High) = \frac{0.00524}{0.00524 + 0.0042} = 0.5551$$

$$P_N(\dots|Low) = \frac{0.0042}{0.00524 + 0.0042} = 0.4449$$

Downloaded by mayk use (mehmetemre5252@hotmail.com)

Area 2: Decision Trees [30 P]

a) A laundry manager wants to optimize his energy and detergent usage. Therefore he tested the washing results for various settings. Use the info Gain criterion to build the first level of a decision tree. The attribute *Temperature* has to be splitted with binary splits.

Temperature	Detergent	Dirt	Class: Result
30	Low	Low	Good
30	Low	High	Bad
30	Low	Medium	Bad
30	High	High	Bad
40	Low	Low	Good
40	Low	Medium	Bad
40	High	High	Bad
60	Low	Medium	Good
60	High	Medium	Good
60	High	High	Good

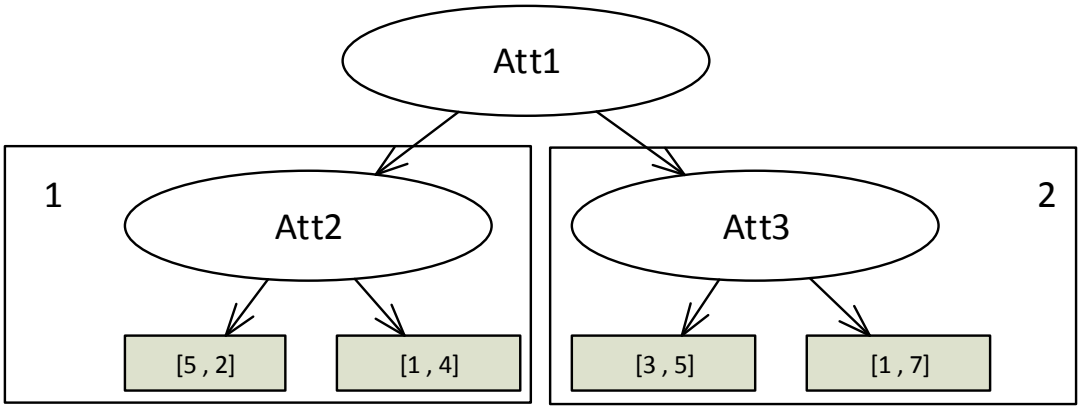
b) What is the advantage of Gain Ration when compared to info Gain? Provide also a brief example. (2-3 sentences)

c) What is the reason for pruning and how does the pruning method subtree replacement work? (3-4 sentences)

d) The following tree was constructed without any pruning. Apply the pruning method subtree replacement to the subtrees 1 and 2. (Notation: [#class1, #class2])

Hint: You do not have to calculate the value of the formula, you can use the values from the table in the appendix.

$e = (appendix),\ f = \frac{E}{N}, for\ c = 0.25, z = 0.69$



Downloaded by mayk use (mehmetemre5252@hotmail.com)

Solution Area 2

a)

Notation: [good, bad]
Class: [5, 5]
Info(Class) = Info([5, 5]) = 1

Gain(Temperature, Split 35) = Info(Class) – Info([1, 3], [4, 2]) = 1.0 – 0.875 = 0.125
Gain(Temperature, Split 50) = Info(Class) – Info([2, 5], [3, 0]) = 1.0 – 0.604 = 0.396
Gain(Detergent) = Info(Class) – Info([3, 3], [2, 2]) = 1.0 – 1.0 = 0.0
Gain(Dirt) = Info(Class) – Info([2, 0], [2, 2], [1, 3]) = 1.0 – 0.725 = 0.275

We choose Temperature with binary Split at 50.

b)

Gain Ratio takes size and number of branches into account and reduces bias to highly branching attributes.

c)

Pruning is a method to reduce overfitting.
Subtree Replacement: estimate error rates for node and leaves. If estimated error rate of leaves greater than error rate of node → prune otherwise keep.

d)

1. Err leaves: 0.386, err node: 0.598 → no pruning
2. Err leaves: 0.362, err node: 0.331 → prune

Downloaded by mayk use (mehmetemre5252@hotmail.com)

Area 3: Evaluation [25 P]

A classifier has been evaluated by a 5-fold Cross Validation.

a) Explain in two to three sentences, how a stratified 5-Fold Cross Validation works.

b) Visualize the evaluation result of the classifier with a Gain Curve. The results are given in the following table. Show also the Gain Curves of a random and an ideal classifier and name the axes. Use 10% steps.

No.	True Class	Score
1	+	0.99
2	+	0.97
3	+	0.95
4	-	0.93
5	+	0.92
6	-	0.90
7	+	0.87
8	-	0.85
9	-	0.84
10	-	0.81

c) A colleague of yours suggests to use a 10-fold Cross Validation instead of the 5-fold one. Which changes in the performance can you expect and why? (2-3 sentences)

d) Does the order of the instances in the dataset affect the outcome of a cross validation? Why? (1-2 sentences)

Downloaded by mayk use (mehmetemre5252@hotmail.com)

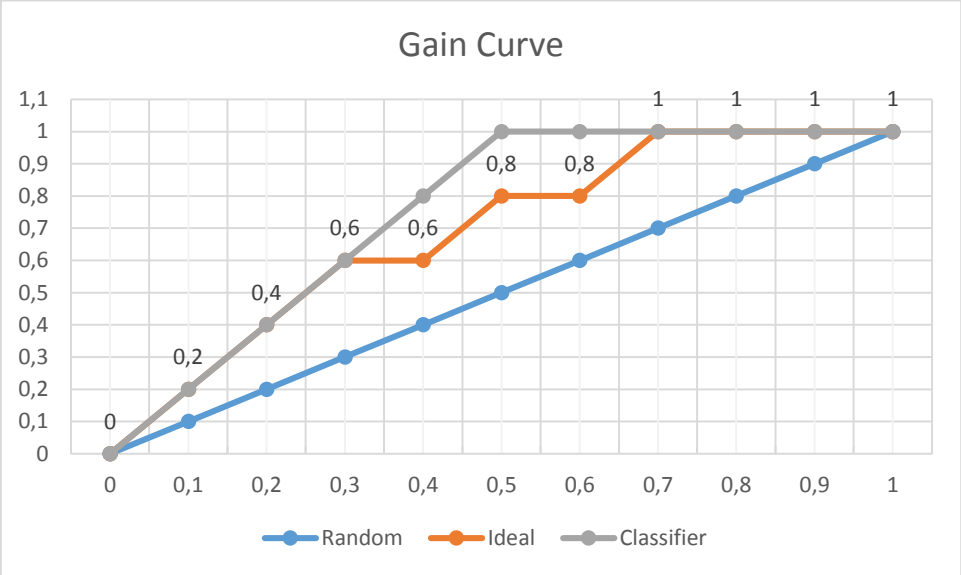
Solution Area 3

a)

- 5 parts/selections/blocks/Abschnitte/Teile
- Class Distribution is kept constant
- N-1/4 are used for training the model, 1 is used for testing
- 5 steps/iterations/Schritte
- Correct plan/each selection is test set in one step

Dataset is divided into 5 sections. The class distribution is kept constant. Each section is used once for testing and 4 times for training.

b)



c)

The measured performance will increase, because the fraction of instances that is used for training will be increased (90% vs. 80%).

(A) Predictive performance
- Increase
- More training instances

(B) Computational performance
- Decrease
- 10 models instead of 5

d) Yes. The order of instances decides in which section an instance is packed. So will be different instances in each section. Generally we expect different results.

- Yes
- Differing sections, differing model
- Generally different results

Downloaded by mayk use (mehmetemre5252@hotmail.com)

Area 4: Logistic Regression [30 P]

You are an advisor of the German Federal Ministry of Health and want to investigate the effect of workspace smoking bans (ban or no ban) on the smoking behavior of indoor workers (smoker or non-smoker), controlling for age and education. You use the following variables to estimate a **logic regression model**:

Variable	Range	Explanation
Smoker*	{1,0}	{“smoker”, “non-smoker”}
Smokeban	{1,0}	{“smoke ban”, “no smoke ban”}
Age	{18,…,88}	{“18 years of age”,…,”88 years of age”}
Edu	{1,2,3}	{“high school graduation”, “BSc from college”, “MSc from college”}**

***Hint 1: Smoker is the dependent variable.**
****Hint 2: The variable “edu” is the highest degree of academic education. The three degrees are mutually exclusive: A worker’s degree is either a high school graduation (edu=1), a BSc from college (edu=2), or a MSc from college (edu=3).**

a) Which attribute(s) do you need to preprocess to be able to run a regression and explain how you would preprocess this/these in one to two sentences.

You run a logistic regression and obtain the following results:

Variable	Coefficient	p-Value
Intercept	-0.821	2.57e-11
Smokeban	-0.23	0.00149
Age	-0.001	0.72184
Dummy(edu=2)	-0.918	<2e-16
Dummy(edu=3)	-1.342	<2e-16

b) Provide a formula to explain the relation between the odds and the coefficient of an independent variable.

c) Based on the formula in c), explain the meaning of the estimated coefficient value of “age” and “dummy(edu=3)” in one to two sentences each.

d) You receive a McFadden R^2 of 0.03874. Explain the meaning of McFadden R^2 and you corresponding value in two to three sentences. Give on possible reason for the obtained R^2 value.

e) Assume that this is data from a panel and the smoking behavior of each individual has been analyzed multiple times in recent years. How would you deal with this change of data in your model?

Downloaded by mayk use (mehmetemre5252@hotmail.com)

Solution Area 4

a) The ordinal attribute “du” has to be preprocessed in order to run a regression. It has to be split up in three binary dummies, from which one has to be left out of the regression model.

b) $\frac{p(x_{ij}+1)}{1-p(x_{ij}+1)} = e^{\beta_j} * \frac{p(x_{ij})}{1-p(x_{ij})}$

c) $\frac{p(age+1)}{1-p(age+1)} = 0.999 * \frac{p(age)}{1-p(age)}$

An increase in age of one year leads to an increase in the odds of being a smoker b a factor of 0.999. This is almost no fall in the odds of being a smoker.

$$\frac{p(MSc\ college)}{1-p(MSc\ college)} = 0.261 * \frac{p(high\ school)}{1-p(high\ school)}$$

An increase in education from only having a high school graduation to having a MSc from college raises the odds of being a smoker by a factor of 0.261.
This is a very strong fall in the odds of being a smoker.

d) The McFadden R^2 is distributed between 0 and 1 and measures the fit of a logistic model. Values between 0.2 and 0.4 resemble a very good fit of the model. Our value of 0.03874 signifies not a good fit.
Many more explaining variables exist, which have not been included in the model e.g. gender, race, type of work (…)

e) Endogeneit1y becomes an issue and you should control for influence of the identity of the panel participants and the year.

Downloaded by mayk use (mehmetemre5252@hotmail.com)