Data Mining and Knowledge Discovery Practice notes: Classification 2

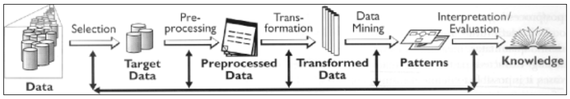
Data Mining and Knowledge Discovery: Practice Notes

Petra Kralj Novak

Petra.Kralj.Novak@ijs.si

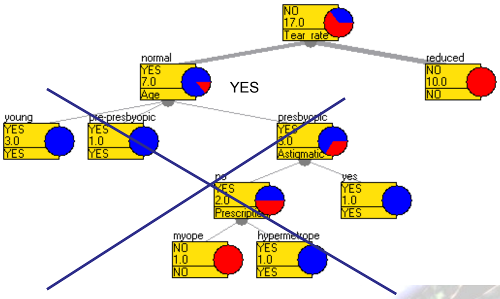
2016/11/16

Keywords

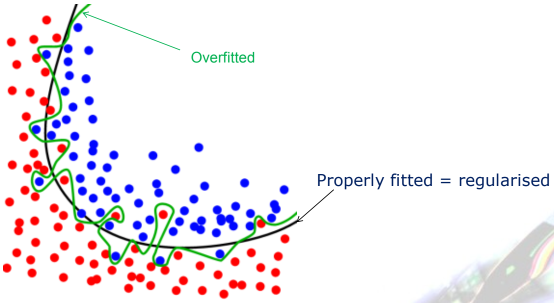


- Data
  - Attribute, example, attribute-value data, target variable, class, discretization
- Algorithms
  - Decision tree induction, entropy, information gain, overfitting, Occam's razor, model pruning, naïve Bayes classifier, KNN, association rules, support, confidence, numeric prediction, regression tree, model tree, heuristics vs. exhaustive search, predictive vs. descriptive DM
- Evaluation
  - Train set, test set, accuracy, confusion matrix, cross validation, true positives, false positives, ROC space, error, precision, recall

Overfitting and Model Pruning

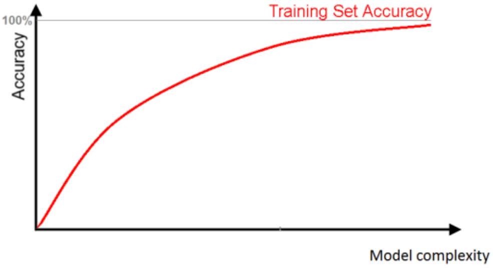


Overfitting & Model Pruning

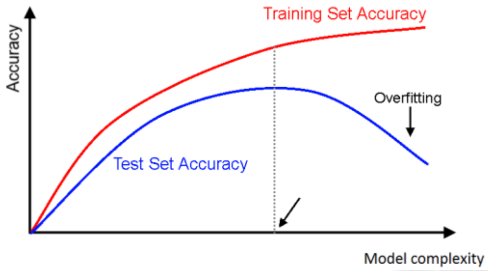


By Chabacano - Own work, GFDL, https://commons.wikimedia.org/w/index.php?curid=3610704

Model complexity and performance on train set



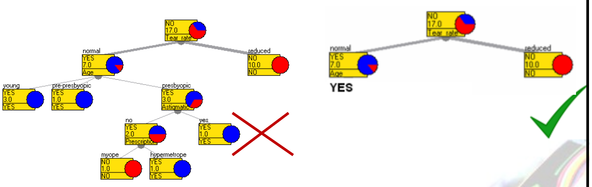
Performance on train and test set



Data Mining and Knowledge Discovery
Practice notes: Classification 2

Occam's raisor

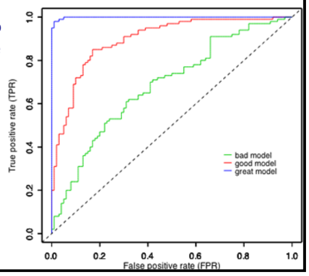
- Suppose there exist two explanations for a phenomena. In this case, the simpler one is usually better.



- Note: classifiers can/should also assign each prediction a confidence score.

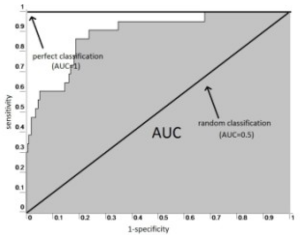
ROC curve and AUC

- **Receiver Operating Characteristic curve** (or ROC curve) is a plot of the true positive rate (TPr=Sensitivity=Recall) against the false positive rate (FPr) for different possible cutpoints.
- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve to the top left corner, the more accurate the classifier.
- The diagonal represents a baseline classifier.



AUC - Area Under (ROC) Curve

- Performance is measured by the area under the ROC curve (AUC). An area of 1 represents a perfect classifier; an area of 0.5 represents a worthless classifier.
- The area under the curve (AUC) is equal to the probability that a classifier will rank a randomly chosen positive example higher than a randomly chosen negative example.

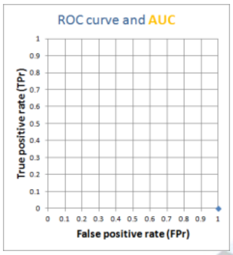


ROC curve and AUC

	Actual class	Confidence classifier forclass Y	FP	TP	FPr	TPr
P1	Y	1				
P2	Y	1				
P3	Y	0.95				
P4	Y	0.9				
P5	Y	0.9				
P6	N	0.85				
P7	Y	0.8				
P8	Y	0.6				
P9	Y	0.55				
P10	Y	0.55				
P11	N	0.3				
P12	N	0.25				
P13	Y	0.25				
P14	N	0.2				
P15	N	0.1				
P16	N	0.1				
P17	N	0.1				
P18	N	0				
P19	N	0				
P20	N	0				

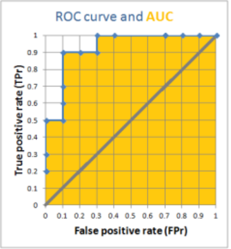
ROC curve and AUC

	Actual class	Confidence classifier forclass Y	FP	TP	FPr	TPr
P1	Y	1	0	2	0	0.2
P2	Y	1	0	2	0	0.2
P3	Y	0.95	0	3	0	0.3
P4	Y	0.9	0	5	0	0.5
P5	Y	0.9	0	5	0	0.5
P6	N	0.85	1	5	0.1	0.5
P7	Y	0.8	1	6	0.1	0.6
P8	Y	0.6	1	7	0.1	0.7
P9	Y	0.55	1	9	0.1	0.9
P10	Y	0.55	1	9	0.1	0.9
P11	N	0.3	2	9	0.2	0.9
P12	N	0.25	3	9	0.3	0.9
P13	Y	0.25	3	10	0.3	1
P14	N	0.2	4	10	0.4	1
P15	N	0.1	7	10	0.7	1
P16	N	0.1	7	10	0.7	1
P17	N	0.1	7	10	0.7	1
P18	N	0	8	10	0.8	1
P19	N	0	9	10	0.9	1
P20	N	0	10	10	1	1



ROC curve and **AUC**

	Actual class	Confidence classifier forclass Y	FPr	TPr
P1	Y	1	0	0.2
P2	Y	1	0	0.2
P3	Y	0.95	0	0.3
P4	Y	0.9	0	0.5
P5	Y	0.9	0	0.5
P6	N	0.85	0.1	0.5
P7	Y	0.8	0.1	0.6
P8	Y	0.6	0.1	0.7
P9	Y	0.55	0.1	0.9
P10	Y	0.55	0.1	0.9
P11	N	0.3	0.2	0.9
P12	N	0.25	0.3	0.9
P13	Y	0.25	0.3	1
P14	N	0.2	0.4	1
P15	N	0.1	0.7	1
P16	N	0.1	0.7	1
P17	N	0.1	0.7	1
P18	N	0	0.8	1
P19	N	0	0.9	1
P20	N	0	1	1



Area under curve
AUC = 0.93

Data Mining and Knowledge Discovery Practice notes: Classification 2

Predicting with Naïve Bayes

Given

- Attribute-value data with nominal target variable

Induce

- Build a Naïve Bayes classifier and estimate its performance on new data

Naïve Bayes classifier

$$P(c \mid a_1, a_2, \ldots a_n) = P(c)\prod_i \frac{P(c \mid a_i)}{P(c)}$$

Assumption: conditional independence of attributes given the class.

Will the spider catch these two ants?

- Color = white, Time = night
- Color = black, Size = large, Time = day

Color	Size	Time	Caught
black	large	day	YES
white	small	night	YES
black	small	day	YES
red	large	night	NO
black	large	night	NO
white	large	night	NO

Naïve Bayes classifier -example

Color	Size	Time	Caught
black	large	day	YES
white	small	night	YES
black	small	day	YES
red	large	night	NO
black	large	night	NO
white	large	night	NO

$v_1 = \text{“}Color = white\text{”}$

$v_2 = \text{“}Time = night\text{”}$

$c_1 = YES$

$c_2 = NO$

$$p(c_1|v_1, v_2) =$$

$$p(Caught = YES|Color = white, Time = night) =$$

$$p(Caught = YES) * \frac{p(Caught = YES|Color = white)}{p(Caught = YES)} * \frac{p(Caught = YES|Time = night)}{p(Caught = YES)} =$$

$$\frac{1}{2} * \frac{\frac{1}{2}}{\frac{1}{2}} * \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{4}$$

Estimating probability

Relative frequency

- **P(c) = n(c) /N**
- A disadvantage of using relative frequencies for probability estimation arises with small sample sizes, especially if they are either very close to zero, or very close to one.
- In our spider example:
  P(Time=day|caught=NO) =
  = 0/3 = 0

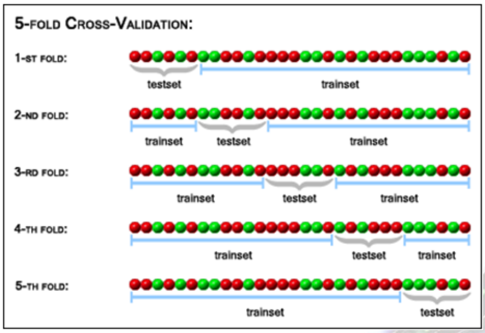
n(c) … number of examples where c is true
N … number of all examples
k … number of classes

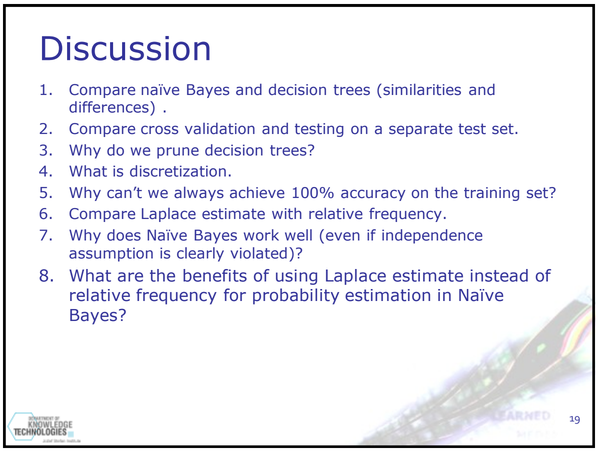
Laplace estimate

- Assumes uniform prior distribution of k classes
- **P(c) = (n(c) + 1) / (N + k)**
- In our spider example: P(Time=day|caught=NO) = (0+1)/(3+2) = 1/5
- With lots of evidence approximates relative frequency
- If there were 300 cases when the spider didn’t catch ants at night: P(Time=day|caught=NO) = (0+1)/(300+2) = 1/302 = 0.003
- With Laplace estimate probabilities can never be 0.

K-fold cross validation

1. The sample set is partitioned into K subsets ("folds") of about equal size
2. A single subset is retained as the validation data for testing the model (this subset is called the "testset"), and the remaining K - 1 subsets together are used as training data ("trainset").
3. A model is trained on the trainset and its performance (accuracy or other performance measure) is evaluated on the testset
4. Model training and evaluation is repeated K times, with each of the K subsets used exactly once as the testset.
5. The average of all the accuracy estimations obtained after each iteration is the resulting accuracy estimation.





Discussion

1. Compare naïve Bayes and decision trees (similarities and differences) .
2. Compare cross validation and testing on a separate test set.
3. Why do we prune decision trees?
4. What is discretization.
5. Why can’t we always achieve 100% accuracy on the training set?
6. Compare Laplace estimate with relative frequency.
7. Why does Naïve Bayes work well (even if independence assumption is clearly violated)?
8. What are the benefits of using Laplace estimate instead of relative frequency for probability estimation in Naïve Bayes?

19