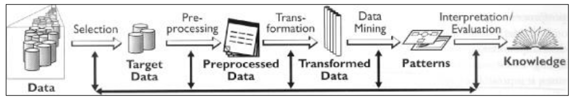
Data Mining and Knowledge Discovery: Practice Notes

Petra Kralj Novak
Petra.Kralj.Novak@ijs.si
8.11.2018

1

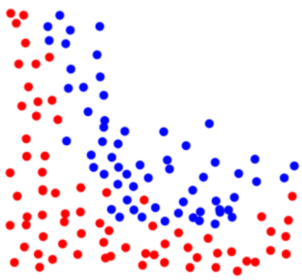
Keywords



- 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, AUC, error, precision, recall

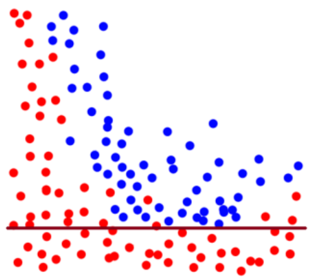
2

DT induction graphically



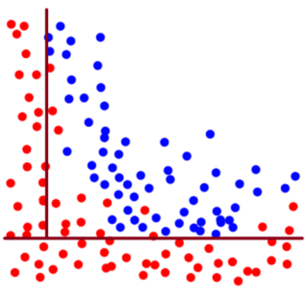
3

DT induction graphically



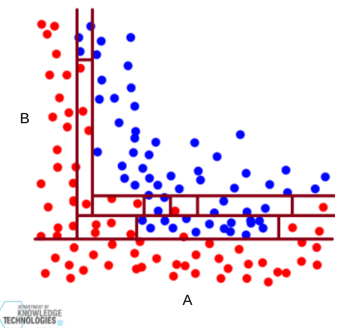
4

DT induction graphically



5

DT induction graphically

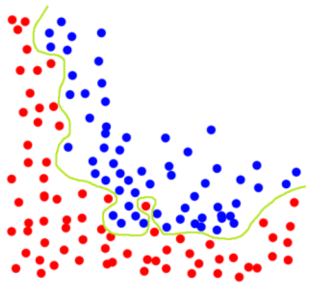


- Language bias: DTs can only make perpendicular splits
- No conditions like A>B

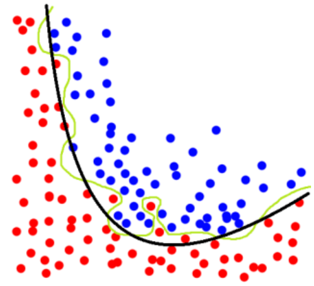
6

Data Mining and Knowledge Discovery
Practice notes 2

Models with other language biases can also overfit (e.g. SVM)

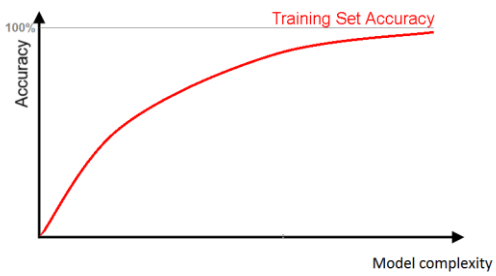


Models with other language biases can also overfit (e.g. SVM)

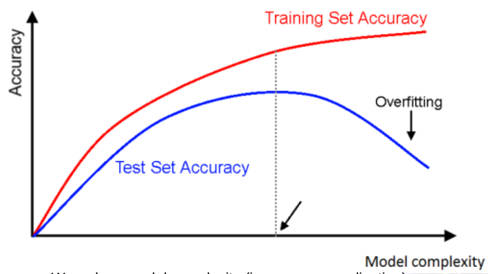


… and need to be generalized

Model complexity and performance on train set



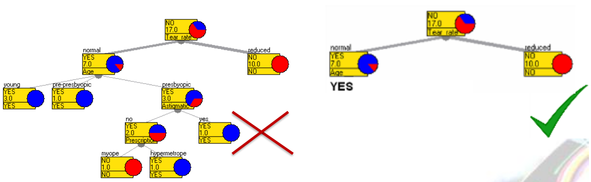
Performance on train and test set



We reduce model complexity (increase generalization)
- to avoid overfitting
- to get more interpretable models (sacrifice some accuracy)

Occam's raisor

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



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

Prediction confidence

- 6/7 examples in this leaf belong to the class Lenses=YES
- 1/7 belongs to the class Lenses=NO

$P(YES) = 6/7 = 0.86$

$P_{Laplace}(YES) = \frac{6+1}{7+2} = 0.78$

- 10/10 examples in this leaf belong to class Lenses=NO

$P(YES) = 0/10 = 0$

$P_{Laplace}(YES) = \frac{0+1}{10+2} = 0.08$

* Laplace probability estimate is explained at "Naïve Bayes".

How confident

Person	Age	Prescription	Astigmatic	Tear_Rate	Lenses
P3	young	hypermetrope	no	normal	**YES**
P9	pre-presbyopic	myope	no	normal	**YES**
P12	pre-presbyopic	hypermetrope	no	reduced	**NO**
P13	pre-presbyopic	myope	yes	normal	**YES**
P15	pre-presbyopic	hypermetrope	yes	normal	**NO**
P16	pre-presbyopic	hypermetrope	yes	reduced	**NO**
P23	presbyopic	hypermetrope	yes	normal	**NO**

Assign to each example in the test set a confidence score* of assigning it to the class "Lenses=YES".

* Use Laplace estimate

How confident

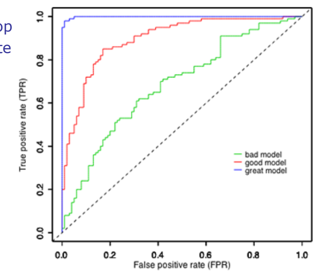
Person	Age	Prescription	Astigmatic	Tear_rate	Actual Lenses	Predicted P(Lenses=YES)
P3	young	hypermetrope	no	normal	YES	0.78
P9	pre-presbyopic	myope	no	normal	YES	0.78
P12	pre-presbyopic	hypermetrope	no	reduced	NO	0.08
P13	pre-presbyopic	myope	yes	normal	YES	0.78
P15	pre-presbyopic	hypermetrope	yes	normal	NO	0.78
P16	pre-presbyopic	hypermetrope	yes	reduced	NO	0.08
P23	presbyopic	hypermetrope	yes	normal	NO	0.78

Sort descending

Person	Age	Prescription	Astigmatic	Tear_rate	Actual Lenses	Predicted P(Lenses=YES)
P3	young	hypermetrope	no	normal	YES	0.78
P9	pre-presbyopic	myope	no	normal	YES	0.78
P13	pre-presbyopic	myope	yes	normal	YES	0.78
P15	pre-presbyopic	hypermetrope	yes	normal	NO	0.78
P23	presbyopic	hypermetrope	yes	normal	NO	0.78
P12	pre-presbyopic	hypermetrope	no	reduced	NO	0.08
P16	pre-presbyopic	hypermetrope	yes	reduced	NO	0.08

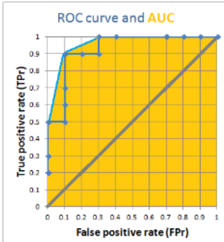
ROC curve and AUC

- **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.



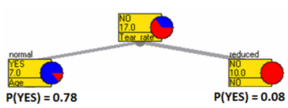
How confident

Person	Age	Prescription	Astigmatic	Tear_rate	Actual Lenses	Predicted P(Lenses=YES)
P3	young	hypermetrope	no	normal	YES	0.78
P9	pre-presbyopic	myope	no	normal	YES	0.78
P13	pre-presbyopic	myope	yes	normal	YES	0.78
P15	pre-presbyopic	hypermetrope	yes	normal	NO	0.78
P23	presbyopic	hypermetrope	yes	normal	NO	0.78
P12	pre-presbyopic	hypermetrope	no	reduced	NO	0.08
P16	pre-presbyopic	hypermetrope	yes	reduced	NO	0.08

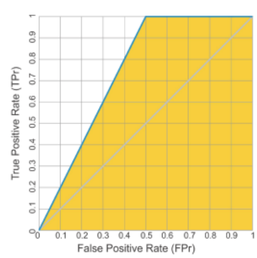
Possible classifiers:
100 % confident → TP=0, FP=0 → TPr= 0, FPr=0
78 % confident → TP=3, FP=2 → TPr= 3/3 = 1, FPr= 2/4 = 0.5
8 % confident → TP=3, FP=4 → TPr= 3/3 = 1, FPr= 4/4 = 1
0 % confident → TP=3, FP=4 → TPr= 3/3 = 1, FPr= 4/4 = 1
TPr = |correctly classified positives| / |all positives|
FPr = |negatives classified as positives| / |all negatives|

Classifier to ROC

Possible classifiers:
100 % confident → TP=0, FP=0 → TPr= 0, FPr=0
78 % confident → TP=3, FP=2 → TPr= 3/3 = 1, FPr= 2/4 = 0.5
8 % confident → TP=3, FP=4 → TPr= 3/3 = 1, FPr= 4/4 = 1
0 % confident → TP=3, FP=4 → TPr= 3/3 = 1, FPr= 4/4 = 1



AUC of our classifier is 0.75.
An AUC close to 0.5 is a bad AUC.

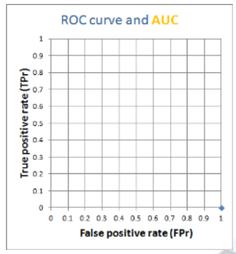


ROC curve and AUC

	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				

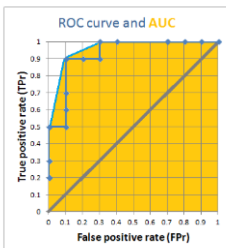
ROC curve and AUC

	Actual class	Classifier confidence 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 (the convex) Curve
AUC = 0.96

Predicting with Naïve Bayes

Given
- Attribute-value data with nominal attributes and target variable

Classify new instances with a Naïve Bayes classifier and estimate its performance on new data

Naïve Bayes classifier

Assumption: conditional independence of attributes given the class.

$$\text{classification} = \text{argmax}_{c_i} P(c_i) \prod_{j=1}^{n} P(v_j|c_i)$$

$$P(c_i|v_1, v_2, ..., v_n) \propto P(c_i) \prod_{j=1}^{n} P(v_j|c_i)$$

$c_1, c_2, ..., c_k$ classes
$P(c_1), P(c_2), ..., P(c_k)$ prior probabilities of classes
$v_1, v_1, ..., v_n$ attribute values

Naïve Bayes classifier

$$P(c_i|v_1, v_2, ..., v_n) \propto P(c_i) \prod_{j=1}^{n} P(v_j|c_i)$$

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

$$P(c_i|v_1, v_2, ..., v_n) \propto P(c_i)\prod_{j=1}^{n} P(v_j|c_i)$$

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 = "$Color = white$"
v_2 = "$Time = night$"
$c_1 = YES$
$c_2 = NO$

$$\begin{aligned} P(C_1|v_1, v_2) = & \\ &= P(\text{YES}|C = w, T = n) \\ &= P(\text{YES}) \cdot P(C = w|\text{YES}) \cdot P(T = n|\text{YES}) \\ &= \frac{1}{2} \cdot \frac{1}{3} \cdot \frac{1}{3} \\ &= \frac{1}{18} \end{aligned}$$

$$\begin{aligned} P(C_2|v_1, v_2) = & \\ &= P(\text{NO}|C = w, T = n) \\ &= P(\text{NO}) \cdot P(C = w|\text{NO}) \cdot P(T = n|\text{NO}) \\ &= \frac{1}{2} \cdot \frac{1}{3} \cdot 1 \\ &= \frac{1}{6} \end{aligned}$$

Estimating probability

Relative frequency

- **P(e) = |e| /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|Class=NO) =
  = 0/3 = 0

|e| … number times an event e happened
n … number of trials
k … number of possible outcomes

Relative frequency vs. Laplace estimate

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

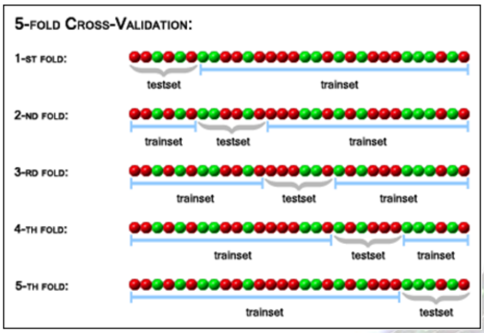
n(c) … number of examples where c is true
N … number of all examples
k … number of possible events

Laplace estimate

- Assumes uniform prior distribution over the probabilities for each possible event
- **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?