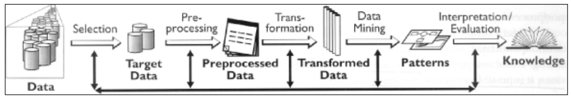
# Data Mining and Knowledge Discovery Practice notes: Classification 2

**Data Mining and Knowledge Discovery: Practice Notes**

Petra Kralj Novak
Petra.Kralj.Novak@ijs.si
2011/12/04

1

## Keywords




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

2

## Practice plan

- 2012/11/20: Predictive data mining 1
  - Decision trees
  - Evaluating classifiers 1: separate test set, confusion matrix, classification accuracy
  - Hands on Weka 1: Just a taste of Weka
- 2012/12/4: Predictive data mining 2
  - Discussion about decision trees
  - Naïve Bayes classifier
  - Evaluating classifiers 2: Cross validation
  - Numeric prediction
  - Hands on Weka 2: Classification and numeric prediction
- 2012/12/4: Descriptive data mining
  - Discussion on classification
  - Association rules
  - Hands on Weka 3: Descriptive data mining
  - Discussion about seminars and exam
- 2013/1/15: Written exam, seminar proposal discussion
- 2013/2/12: Data mining seminar presentations

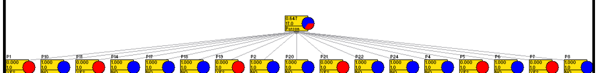
3

## Discussion about decision trees

- → How much is the information gain for the "attribute" Person? How would it perform on the test set?
- How do we compute entropy for a target variable that has three values? Lenses = {hard=4, soft=5, none=13}
- What would be the classification accuracy of our decision tree if we pruned it at the node *Astigmatic*?
- What are the stopping criteria for building a decision tree?
- How would you compute the information gain for a numeric attribute?

4

## Information gain of the "attribute" Person



On training set
- As many values as there are examples
- Each leaf has exactly one example
- E(1/1, 0/1) = 0 (entropy of each leaf is zero)
- The weighted sum of entropies is zero
- The information gain is maximum (as much as the entropy of the entire training set)

On testing set
- The values from the testing set do not appear in the tree

5

## Discussion about decision trees

- How much is the information gain for the "attribute" Person? How would it perform on the test set?
- → How do we compute entropy for a target variable that has three values? Lenses = {hard=4, soft=5, none=13}
- What would be the classification accuracy of our decision tree if we pruned it at the node *Astigmatic*?
- What are the stopping criteria for building a decision tree?
- How would you compute the information gain for a numeric attribute?

6

# Data Mining and Knowledge Discovery
# Practice notes: Classification 2

## Entropy{hard=4, soft=5, none=13}=

= E(4/22, 5/22, 13/22)

= $-\Sigma p_i * \log_2 p_i$

= $-4/22 * \log_2 4/22 - 5/22 * \log_2 5/22 - 13/22 * \log_2 13/22$

= 1.38

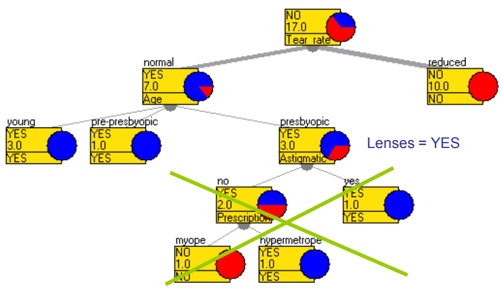


## Discussion about decision trees

- How much is the information gain for the "attribute" Person? How would it perform on the test set?
- How do we compute entropy for a target variable that has three values? Lenses = {hard=4, soft=5, none=13}
- → What would be the classification accuracy of our decision tree if we pruned it at the node Astigmatic?
- What are the stopping criteria for building a decision tree?
- How would you compute the information gain for a numeric attribute?

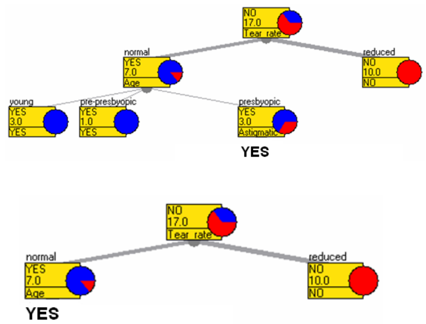


## Decision tree pruning






## These two trees are equivalent






## Classification accuracy of the pruned tree

| 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 |

Ca = (3+2)/ (3+2+2+0) = 71%


| | Predicted positive | Predicted negative |
|---|---|---|
| Actual positive | TP=3 | FN=0 |
| Actual negative | FP=2 | TN=2 |

## Discussion about decision trees

- How much is the information gain for the "attribute" Person? How would it perform on the test set?
- How do we compute entropy for a target variable that has three values? Lenses = {hard=4, soft=5, none=13}
- What would be the classification accuracy of our decision tree if we pruned it at the node Astigmatic?
- → What are the stopping criteria for building a decision tree?
- How would you compute the information gain for a numeric attribute?

# Data Mining and Knowledge Discovery
# Practice notes: Classification 2

## Stopping criteria for building a decision tree

- ID3
  - "Pure" nodes (entropy =0)
  - Out of attributes
- J48 (C4.5)
  - Minimum number of instances in a leaf constraint

13

## Discussion about decision trees

- How much is the information gain for the "attribute" Person? How would it perform on the test set?
- How do we compute entropy for a target variable that has three values? Lenses = {hard=4, soft=5, none=13}
- What would be the classification accuracy of our decision tree if we pruned it at the node Astigmatic?
- What are the stopping criteria for building a decision tree?
- How would you compute the information gain for a numeric attribute?

14

## Information gain of a numeric attribute

| Age | Lenses |
|---|---|
| 67 | YES |
| 52 | YES |
| 63 | NO |
| 26 | YES |
| 65 | NO |
| 23 | YES |
| 65 | NO |
| 25 | YES |
| 26 | YES |
| 57 | NO |
| 49 | NO |
| 23 | YES |
| 39 | NO |
| 55 | NO |
| 53 | NO |
| 38 | NO |
| 67 | YES |
| 54 | NO |
| 29 | YES |
| 46 | NO |
| 44 | YES |
| 32 | NO |
| 39 | NO |
| 45 | YES |

15

## Information gain of a numeric attribute

| Age | Lenses |
|---|---|
| 67 | YES |
| 52 | YES |
| 63 | NO |
| 26 | YES |
| 65 | NO |
| 23 | YES |
| 65 | NO |
| 25 | YES |
| 26 | YES |
| 57 | NO |
| 49 | NO |
| 23 | YES |
| 39 | NO |
| 55 | NO |
| 53 | NO |
| 38 | NO |
| 67 | YES |
| 54 | NO |
| 29 | YES |
| 46 | NO |
| 44 | YES |
| 32 | NO |
| 39 | NO |
| 45 | YES |

**Sort by Age** →

| Age | Lenses |
|---|---|
| 23 | YES |
| 23 | YES |
| 25 | YES |
| 26 | YES |
| 26 | YES |
| 29 | YES |
| 32 | NO |
| 38 | NO |
| 39 | NO |
| 39 | NO |
| 44 | YES |
| 45 | YES |
| 46 | NO |
| 49 | NO |
| 52 | YES |
| 53 | NO |
| 54 | NO |
| 55 | NO |
| 57 | NO |
| 63 | NO |
| 65 | NO |
| 65 | NO |
| 67 | YES |
| 67 | YES |

16

## Information gain of a numeric attribute

| Age | Lenses |
|---|---|
| 67 | YES |
| 52 | YES |
| 63 | NO |
| 26 | YES |
| 65 | NO |
| 23 | YES |
| 65 | NO |
| 25 | YES |
| 26 | YES |
| 57 | NO |
| 49 | NO |
| 23 | YES |
| 39 | NO |
| 55 | NO |
| 53 | NO |
| 38 | NO |
| 67 | YES |
| 54 | NO |
| 29 | YES |
| 46 | NO |
| 44 | YES |
| 32 | NO |
| 39 | NO |
| 45 | YES |

**Sort by Age** →

| Age | Lenses |
|---|---|
| 23 | YES |
| 23 | YES |
| 25 | YES |
| 26 | YES |
| 26 | YES |
| 29 | YES |
| 32 | NO |
| 38 | NO |
| 39 | NO |
| 39 | NO |
| 44 | YES |
| 45 | YES |
| 46 | NO |
| 49 | NO |
| 52 | YES |
| 53 | NO |
| 54 | NO |
| 55 | NO |
| 57 | NO |
| 63 | NO |
| 65 | NO |
| 65 | NO |
| 67 | YES |
| 67 | YES |

**Define possible splitting points** →

| Age | Lenses |
|---|---|
| 23 | YES |
| 23 | YES |
| 25 | YES |
| 26 | YES |
| 26 | YES |
| 29 | YES |
| 32 | NO |
| 38 | NO |
| 39 | NO |
| 39 | NO |
| 44 | YES |
| 45 | YES |
| 46 | NO |
| 49 | NO |
| 52 | YES |
| 53 | NO |
| 54 | NO |
| 55 | NO |
| 57 | NO |
| 63 | NO |
| 65 | NO |
| 65 | NO |
| 67 | YES |
| 67 | YES |

17

## Information gain of a numeric attribute

| Age | Lenses | Split point |
|---|---|---|
| 23 | YES | |
| 23 | YES | |
| 25 | YES | |
| 26 | YES | |
| 26 | YES | |
| 29 | YES | → **30.5** |
| 32 | NO | |
| 38 | NO | |
| 39 | NO | |
| 39 | NO | → **41.5** |
| 44 | YES | |
| 45 | YES | → **45.5** |
| 46 | NO | |
| 49 | NO | → **50.5** |
| 52 | YES | → **52.5** |
| 53 | NO | |
| 54 | NO | |
| 55 | NO | |
| 57 | NO | |
| 63 | NO | |
| 65 | NO | |
| 65 | NO | → **66** |
| 67 | YES | |
| 67 | YES | |

18

# Data Mining and Knowledge Discovery
# Practice notes: Classification 2

## Information gain of a numeric attribute

| Age | Lenses |
|---|---|
| 23 | YES |
| 23 | YES |
| 25 | YES |
| 26 | YES |
| 26 | YES |
| 29 | YES |
| 32 | NO |
| 38 | NO |
| 39 | NO |
| 39 | NO |
| 44 | YES |
| 45 | YES |
| 46 | NO |
| 49 | NO |
| 52 | YES |
| 53 | NO |
| 54 | NO |
| 55 | NO |
| 57 | NO |
| 63 | NO |
| 65 | NO |
| 65 | NO |
| 67 | YES |
| 67 | YES |

30.5
41.5
45.5
50.5
52.5
66

Age
<30.5 >=30.5
6/24 18/24
E(6/6 , 0/6) = 0
E(5/18 , 13/18) = 0.85

19

## Information gain of a numeric attribute

| Age | Lenses |
|---|---|
| 23 | YES |
| 23 | YES |
| 25 | YES |
| 26 | YES |
| 26 | YES |
| 29 | YES |
| 32 | NO |
| 38 | NO |
| 39 | NO |
| 39 | NO |
| 44 | YES |
| 45 | YES |
| 46 | NO |
| 49 | NO |
| 52 | YES |
| 53 | NO |
| 54 | NO |
| 55 | NO |
| 57 | NO |
| 63 | NO |
| 65 | NO |
| 65 | NO |
| 67 | YES |
| 67 | YES |

30.5
41.5
45.5
50.5
52.5
66

**E(S) = E(11/24 , 13/24) = 0.99**

Age
<30.5 >=30.5
6/24 18/24
E(6/6 , 0/6) = 0
E(5/18 , 13/18) = 0.85

**InfoGain (S, Age$_{30.5}$)=**

**= E(S) - $\sum p_v$ E(pv)**

**= 0.99 – (6/24*0 + 18/24*0.85)**

**= 0.35**

20

## Information gain of a numeric attribute

| Age | Lenses |
|---|---|
| 23 | YES |
| 23 | YES |
| 25 | YES |
| 26 | YES |
| 26 | YES |
| 29 | YES |
| 32 | NO |
| 38 | NO |
| 39 | NO |
| 39 | NO |
| 44 | YES |
| 45 | YES |
| 46 | NO |
| 49 | NO |
| 52 | YES |
| 53 | NO |
| 54 | NO |
| 55 | NO |
| 57 | NO |
| 63 | NO |
| 65 | NO |
| 65 | NO |
| 67 | YES |
| 67 | YES |

30.5
41.5
45.5
50.5
52.5
66

Age
<30.5 >=30.5

InfoGain (S, Age$_{30.5}$) = 0.35

Age
<41.5 >=41.5

Age
<45.5 >=45.5

Age
<50.5 >=50.5

Age
<52.5 >=52.5

Age
<66 >=66

21

## Practice plan

- 2012/11/20: Predictive data mining 1
  - Decision trees
  - Evaluating classifiers 1: separate test set, confusion matrix, classification accuracy
  - Hands on Weka 1: Just a taste of Weka
- 2012/12/4: Predictive data mining 2
  - Discussion about decision trees
  - Naïve Bayes classifier
  - Evaluating classifiers 2: Cross validation
  - Numeric prediction
  - Hands on Weka 2: Classification and numeric prediction
- 2012/12/4: Descriptive data mining
  - Discussion on classification
  - Association rules
  - Hands on Weka 3: Descriptive data mining
  - Discussion about seminars and exam
- 2013/1/15: Written exam, seminar proposal discussion
- 2013/2/12: Data mining seminar presentations

22

## 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

23

## Naïve Bayes classifier

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

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 |

24

# Data Mining and Knowledge Discovery
# Practice notes: Classification 2

## 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 |

$v_1 = \text{"}Color = white\text{"}$

$v_2 = \text{"}Time = night\text{"}$

$c_1 = YES$

$c_2 = NO$

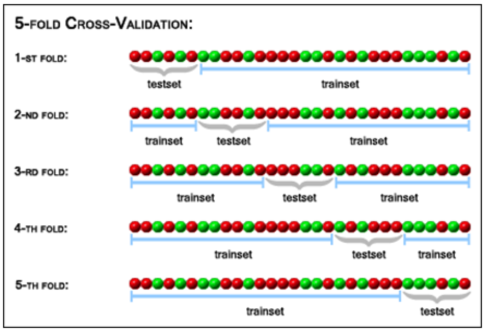
$$p(c_1|v_1, v_2) =$$
$$p(Caught = YES|Color = white, Time = night) =$$
$$p(Caught = YES) * \frac{p(Caught = YES|Color = white)}{p(Caught = YES)} * \frac{p(Caught = YES|Time = night)}{p(Caught = YES)} =$$
$$\frac{1}{2} * \frac{\frac{1}{2}}{\frac{1}{2}} * \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{4}$$

25

## K-fold cross validation

1. The sample set is partitioned into K subsets ("folds") of about equal size
2. Of the K subsets, 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 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.

26



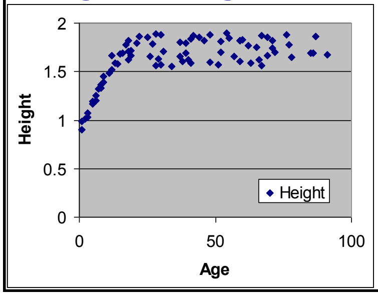

27

## Numeric prediction

28

## Example

- data about 80 people: Age and Height




| Age | Height |
|---|---|
| 3 | 1.03 |
| 5 | 1.19 |
| 6 | 1.26 |
| 9 | 1.39 |
| 15 | 1.69 |
| 19 | 1.67 |
| 22 | 1.86 |
| 25 | 1.85 |
| 41 | 1.59 |
| 48 | 1.60 |
| 54 | 1.90 |
| 71 | 1.82 |
| … | … |

29

## Test set

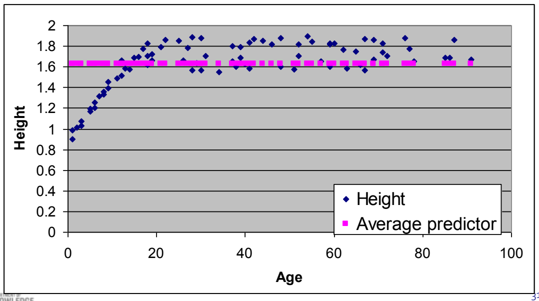
| Age | Height |
|---|---|
| 2 | 0.85 |
| 10 | 1.4 |
| 35 | 1.7 |
| 70 | 1.6 |

30

# Data Mining and Knowledge Discovery
# Practice notes: Classification 2

## Baseline numeric predictor

- Average of the target variable






## Baseline predictor: prediction

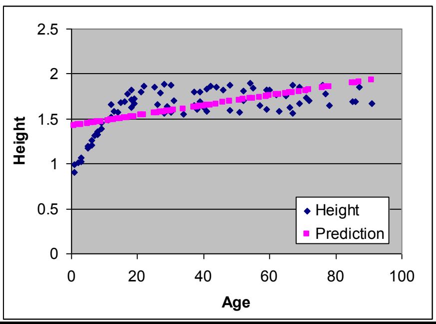
Average of the target variable is 1.63

| Age | Height | Baseline |
|---|---|---|
| 2 | 0.85 | |
| 10 | 1.4 | |
| 35 | 1.7 | |
| 70 | 1.6 | |



## Linear Regression Model

Height = 0.0056 * Age + 1.4181





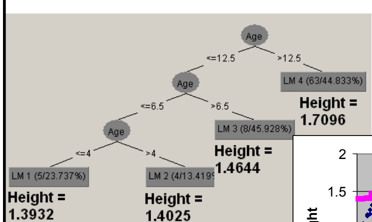

## Linear Regression: prediction

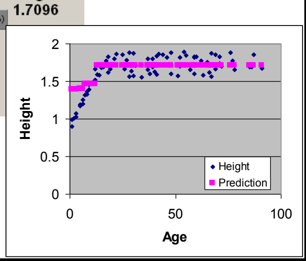
Height = 0.0056 * Age + 1.4181

| Age | Height | Linear regression |
|---|---|---|
| 2 | 0.85 | |
| 10 | 1.4 | |
| 35 | 1.7 | |
| 70 | 1.6 | |

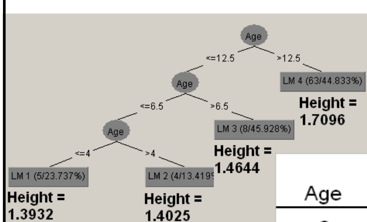


## Regression tree



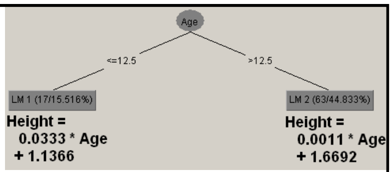

## Regression tree: prediction

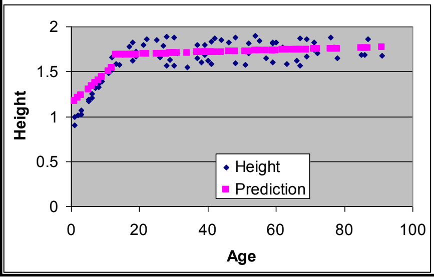



| Age | Height | Regression tree |
|---|---|---|
| 2 | 0.85 | |
| 10 | 1.4 | |
| 35 | 1.7 | |
| 70 | 1.6 | |

# Data Mining and Knowledge Discovery
# Practice notes: Classification 2

## Model tree

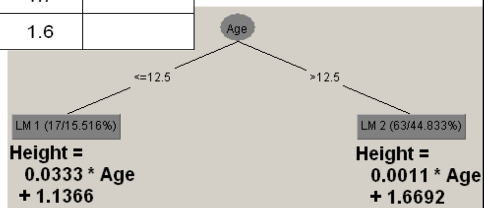



37

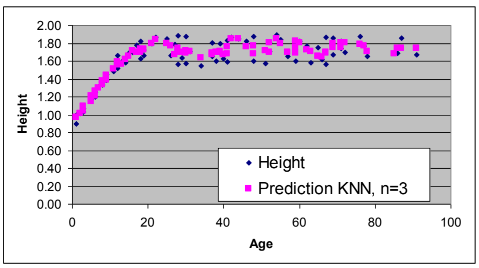
## Model tree: prediction

| Age | Height | Model tree |
|---|---|---|
| 2 | 0.85 | |
| 10 | 1.4 | |
| 35 | 1.7 | |
| 70 | 1.6 | |




## KNN – K nearest neighbors

- Looks at K closest examples (by non-target attributes) and predicts the average of their target variable
- In this example, K=3




39

## KNN prediction

| Age | Height |
|---|---|
| 1 | 0.90 |
| 1 | 0.99 |
| 2 | 1.01 |
| 3 | 1.03 |
| 3 | 1.07 |
| 5 | 1.19 |
| 5 | 1.17 |

| Age | Height | kNN |
|---|---|---|
| 2 | 0.85 | |
| 10 | 1.4 | |
| 35 | 1.7 | |
| 70 | 1.6 | |

40

## KNN prediction

| Age | Height |
|---|---|
| 8 | 1.36 |
| 8 | 1.33 |
| 9 | 1.45 |
| 9 | 1.39 |
| 11 | 1.49 |
| 12 | 1.66 |
| 12 | 1.52 |
| 13 | 1.59 |
| 14 | 1.58 |

| Age | Height | kNN |
|---|---|---|
| 2 | 0.85 | |
| 10 | 1.4 | |
| 35 | 1.7 | |
| 70 | 1.6 | |

41

## KNN prediction

| Age | Height |
|---|---|
| 30 | 1.57 |
| 30 | 1.88 |
| 31 | 1.71 |
| 34 | 1.55 |
| 37 | 1.65 |
| 37 | 1.80 |
| 38 | 1.60 |
| 39 | 1.69 |
| 39 | 1.80 |

| Age | Height | kNN |
|---|---|---|
| 2 | 0.85 | |
| 10 | 1.4 | |
| 35 | 1.7 | |
| 70 | 1.6 | |

42

# Data Mining and Knowledge Discovery
# Practice notes: Classification 2

## KNN prediction

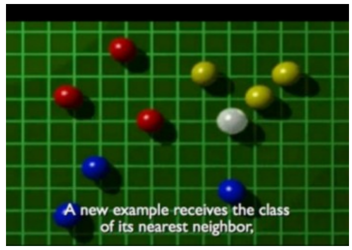
| Age | Height |
|---|---|
| 67 | 1.56 |
| 67 | 1.87 |
| 69 | 1.67 |
| 69 | 1.86 |
| 71 | 1.74 |
| 71 | 1.82 |
| 72 | 1.70 |
| 76 | 1.88 |

| Age | Height | kNN |
|---|---|---|
| 2 | 0.85 | |
| 10 | 1.4 | |
| 35 | 1.7 | |
| 70 | 1.6 | |

43

## KNN video

- http://videolectures.net/aaai07_bosch_knnc




44

## Which predictor is the best?

| Age | Height | Baseline | Linear regression | Regression tree | Model tree | kNN |
|---|---|---|---|---|---|---|
| 2 | 0.85 | 1.63 | 1.43 | 1.39 | 1.20 | 1.00 |
| 10 | 1.4 | 1.63 | 1.47 | 1.46 | 1.47 | 1.44 |
| 35 | 1.7 | 1.63 | 1.61 | 1.71 | 1.71 | 1.67 |
| 70 | 1.6 | 1.63 | 1.81 | 1.71 | 1.75 | 1.77 |

45

## Evaluating numeric prediction

| Performance measure | Formula |
|---|---|
| mean-squared error | $\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{n}$ |
| root mean-squared error | $\sqrt{\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{n}}$ |
| mean absolute error | $\frac{\lvert p_1-a_1\rvert+\ldots+\lvert p_n-a_n\rvert}{n}$ |
| relative squared error | $\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\ldots+(a_n-\bar{a})^2}$, where $\bar{a}=\frac{1}{n}\sum_i a_i$ |
| root relative squared error | $\sqrt{\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\ldots+(a_n-\bar{a})^2}}$ |
| relative absolute error | $\frac{\lvert p_1-a_1\rvert+\ldots+\lvert p_n-a_n\rvert}{\lvert a_1-\bar{a}\rvert+\ldots+\lvert a_n-\bar{a}\rvert}$ |
| correlation coefficient | $\frac{S_{PA}}{\sqrt{S_P S_A}}$, where $S_{PA}=\frac{\sum_i(p_i-\bar{p})(a_i-\bar{a})}{n-1}$, $S_p=\frac{\sum_i(p_i-\bar{p})^2}{n-1}$, and $S_A=\frac{\sum_i(a_i-\bar{a})^2}{n-1}$ |

| Numeric prediction | Classification |
|---|---|
| **Data**: attribute-value description | |
| **Target variable**: Continuous | **Target variable**: Categorical (nominal) |
| **Evaluation**: cross validation, separate test set, … | |
| **Error**: MSE, MAE, RMSE, … | **Error**: 1-accuracy |
| **Algorithms**: Linear regression, regression trees,… | **Algorithms**: Decision trees, Naïve Bayes, … |
| **Baseline predictor**: Mean of the target variable | **Baseline predictor**: Majority class |

47

## Discussion

1. Compare naïve Bayes and decision trees (similarities and differences) .
2. Can KNN be used for classification tasks?
3. Compare KNN and Naïve Bayes.
4. Compare decision trees and regression trees.
5. Consider a dataset with a target variable with five possible values:
   1. non sufficient
   2. sufficient
   3. good
   4. very good
   5. excellent
   1. Is this a classification or a numeric prediction problem?
   2. What if such a variable is an attribute, is it nominal or numeric?
6. Compare cross validation and testing on a different test set.
7. Why do we prune decision trees?
8. List 3 numeric prediction methods.

48