Project 1

140313

1. make a scenario and build a bayesian network + conditional probability table

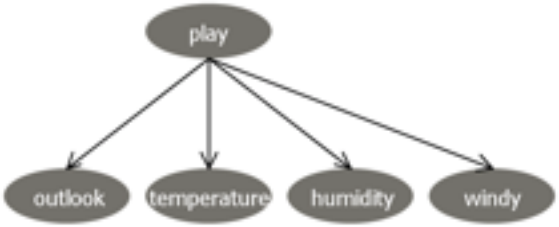
- use only nominal variable
- network.txt

```
@attribute play {yes, no}
…

@graph
play -> outlook
play -> temperature
play -> humidity
play -> windy
```

cpt.txt

```
P(target)
yes no
0.2 0.8
P(target | parents)
outlook humidity yes no
sunny high 0.2 0.8
```



2. implement data generator

- generate data from the network
- input: network.txt cpt.txt #ofInstances
- output: data.arff

Attribute
Relation
File
Format

Make an .arff file

@RELATION <dataset name>

@ATTRIBUTE <feature1 name> <feature1 type>
@ATTRIBUTE <feature2 name> <feature2 type>
@ATTRIBUTE <feature3 name> <feature3 type>
@ATTRIBUTE <feature4 name> <feature4 type>
…..
@ATTRIBUTE class <classes name>

@DATA
1,2,3,1, apple
3,2,0,1, book
……

data type
①numeric (real or integer numbers)
②<nominal-specification>
③string
④date [<date-format>]

More details:
http://www.cs.waikato.ac.nz/~ml/weka/arff.html

Header

Data
(feature values + class)

Data format for Weka (.ARFF)

Header

```
@relation heart-disease-simplified

@attribute age numeric
@attribute sex { female, male}
@attribute chest_pain_type { typ_angina, asympt, non_anginal,
    atyp_angina}
@attribute cholesterol numeric
@attribute exercise_induced_angina { no, yes}
@attribute class { present, not_present}

@data
```

Data
(CSV format)

```
63,male,typ_angina,233,no,not_present
67,male,asympt,286,yes,present
67,male,asympt,229,yes,present
38,female,non_anginal,?,no,not_present
```

Note: You can easily generate 'arff' file by adding a header to a usual CSV text file

3. implement network CPT calculator

- input1: network.txt
- input2: data.arff
- input3: test.txt
- ———————
- output: cpt_out.txt

- implement simple estimator

Weather Data

Outlook	Temp.	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

- A new day:

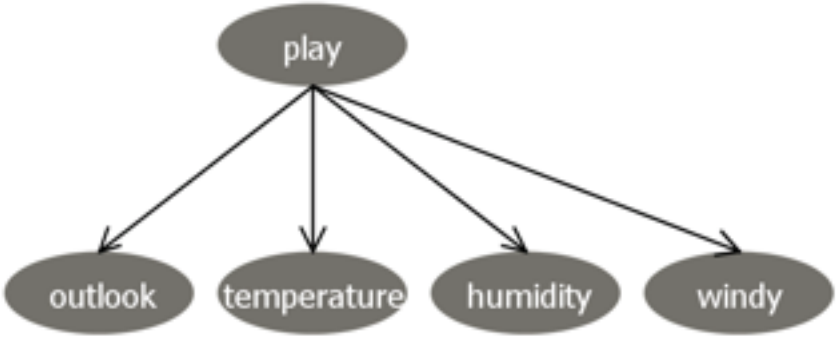
Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

Here we don't really have effects, but rather "**evidence**."

Naïve Bayes

Assuming the attributes are independent of each other, we have a Naïve Bayesian Network:

Outlook	Temp.	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No



P(play=yes)=9/14,

with Laplace correction:
P(play=yes)=9+1/14+2=0.625

In general, to make Laplace correction, we add an initial count (1) to the total of all instances with a given attribute value, and we add the number of distinct values of the same attribute to the total number of instances in the group.

Naïve Bayes

And to fill the Conditional Probability Tables we compute conditional probabilities for each node in form: Pr (attribute=value | parents values) for each combinations of attributes values in parent nodes

Outlook	Temp.	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

P(outlook=sunny|play=yes) =(2+1)/(9+3)=3/12

P(outlook=rainy|play=yes) =(3+1)/(9+3)=4/12

P(outlook=overcast| play=yes) =(4+1)/(9+3)=5/12

Sum is1

P(outlook=sunny|play=yes) =(2+1)/(9+3)=3/12

P(outlook=sunny|play=no) =(3+1)/(5+3)=4/8

Sum is NOT 1

- test.txt —> cpt_out.txt

target
==========

P(target)
yes no
0.2 0.8
P(target | parents)
outlook humidity yes no
sunny high 0.2 0.8

- test.txt —> cpt_out.txt

```
target
==========
observation1
observation2
…
```

```
P(target | observations)
outlook humidity yes no
sunny high 0.2 0.8
sunny normal 0.3 0.7
…
P(target | observations, parents)
outlook humidity yes no
sunny high 0.2 0.8
sunny normal 0.3 0.7
…
```

- test.txt —> cpt_out.txt

```
target
==========
observation1 value1
observation2 value2
…
```

```
P(target | observation1 = value1, observation2 = value2)
outlook humidity yes no
sunny high 0.2 0.8
P(target | observation1 = value1, … , parents)
outlook humidity temperature yes no
sunny high high 0.2 0.8
sunny high mild 0.5 0.5
sunny high low 0.7 0.3
```

4. write a report (~max5page)

- Bayesian Network 개념
- 1, 2, 3 설명
- conditional independency 등 분석
- 간단한 실행 방법

- ~ midterm (~4/17)
- 구현 50% 보고서 50%
- hschun@bi.snu.ac.kr
- 주석 필수
- C/C++, Java, Matlab, Python etc.

appendix: Bayesian Network in WEKA

Structure Learning

Introduction to Weka

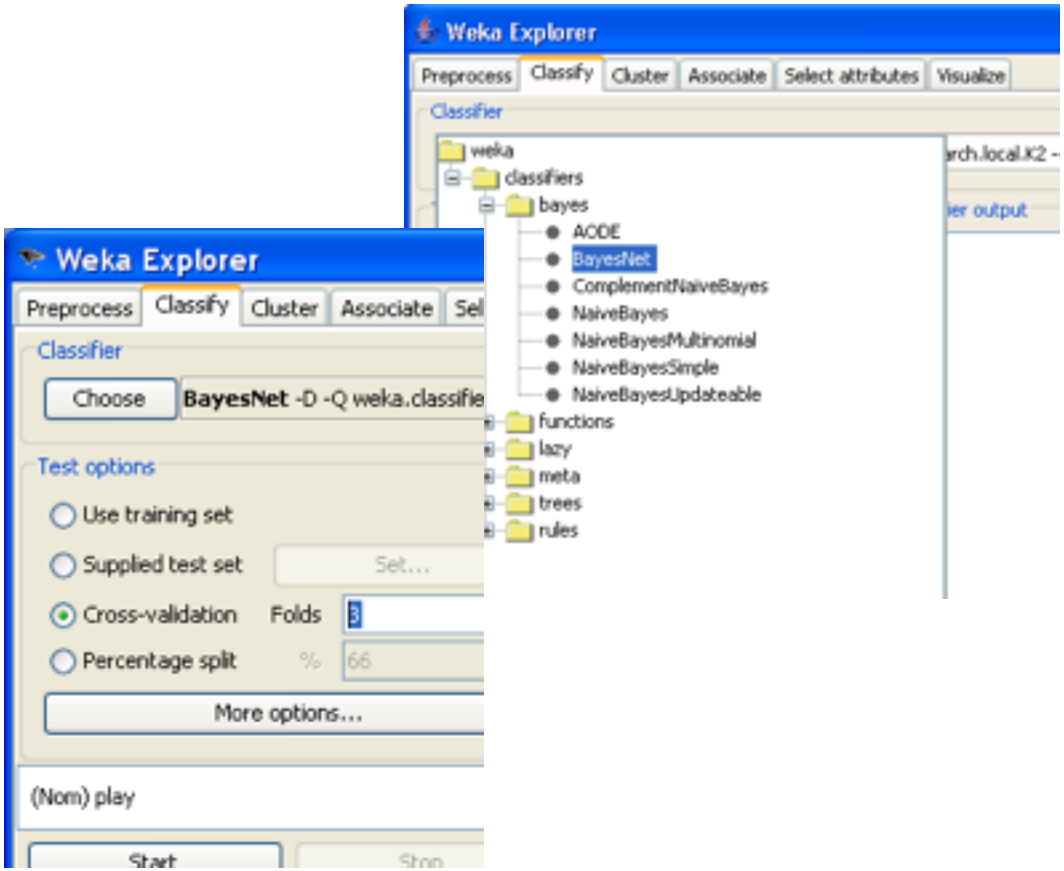
❖ Weka 3: Data Mining Software in Java

- Weka is a collection of machine learning algorithms for data mining tasks
- What you can do with weka?
  - data pre-processing, feature selection, classification, regression, clustering, association rules, and visualization

- Weka is an open source software issued under the GNU General Public License
- How to get? http://www.cs.waikato.ac.nz/ml/weka/ or just type 'Weka' in google.

WEKA Exercise 1. Bayesian network for weather data with default parameters

- **Preprocess** tab
  - ***Open file*** *weather.nominal.arff*
  - Perform ***filters*** (if needed): discretize or replace missing values
- **Classify** tab
  - ***Classifier->choose->classifiers->bayes->BayesNet***
  - Click on row with selected classifier, change Laplace correction (initial count) to 1 (instead of 0.5) in the ***option row for the estimator***
  - ***Cross-validation*** change to ***3 folds*** (since we have only 14 instances, with 10 folds cross validation we will have test groups of size less than 2, which makes the classifier less reliable). Press ***Start***

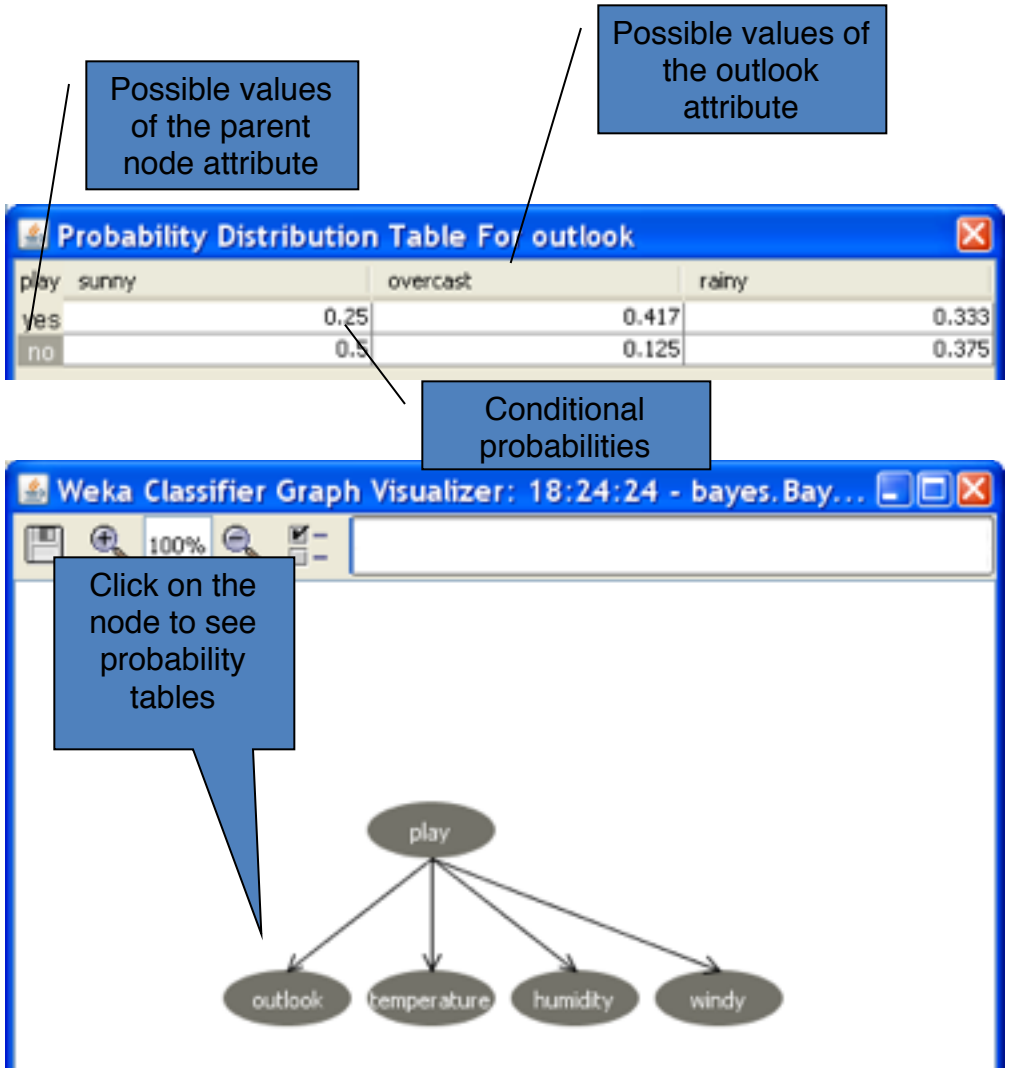


WEKA Exercise 2. Examining the output

- In the history box, right-click and choose ***visualize graph***. Check that probabilities in CPT correspond to what we calculated before (clicking on the graph node brings table of conditional probabilities)
- Naïve Bayes? Study parameters of the program. Click on ***choose*** line again.
- Save this model in file weather.xml for later use
- Click on ***searchAlgorithm*** row. Default parameters are:

  initAsNaiveBayes=true

  maxNrOfParents=1
- Change ***maxNrOfParents=2***. Run. Visualize graph
- Change to ***initAsNaiveBayes=false***. Run. Visualize graph. Change back to true.



How WEKA infers a structure of the network

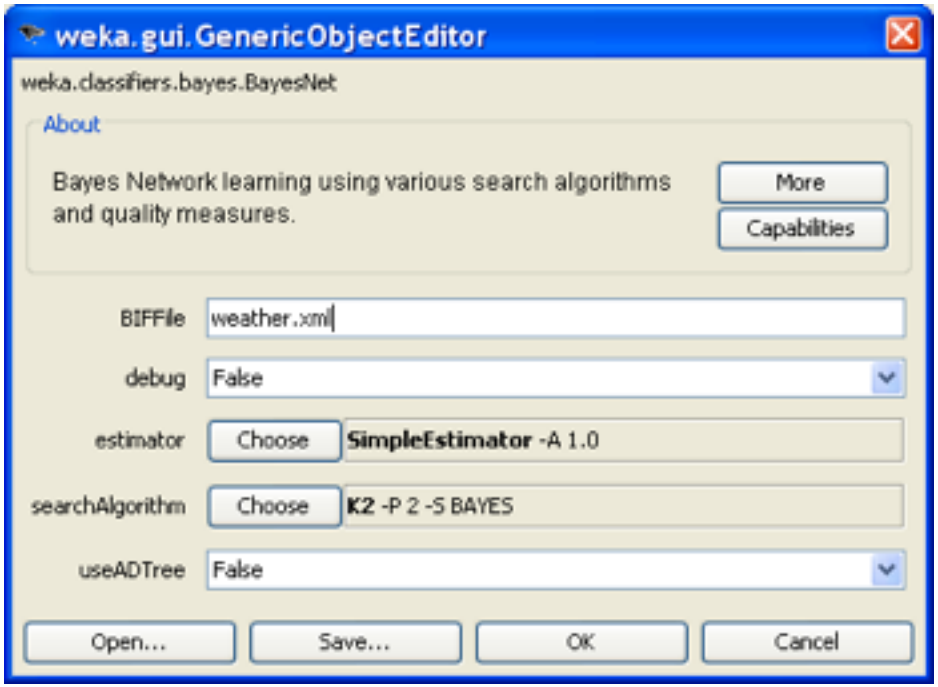
- The nodes correspond to the attributes
- Learning the structure is to find edges
  - Searching through the possible edges sets
  - For each set estimate the conditional probability tables from the data
  - Estimate quality of the network as the probability of obtaining the set of data given this network

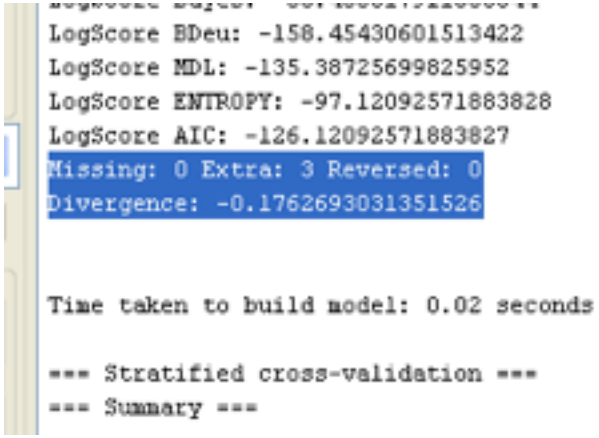
WEKA Search Algorithms. Example

- By default: K2.
- Starts with a given ordering of attributes.
- Adds one node in order and considers adding edges from each previously added node to a new node.
- Then it adds the edge which maximizes the network score.
- The number of parents is restricted to a predefined maximum.
- *The Markov blanket* of a node includes all its parents, children and children parents. It is proven, that a given node is conditionally dependent only on nodes in its Markov blanket. So the edge is added from the class node to the node which is not in its Markov blanket. Otherwise the value of this attribute would be irrelevant for the class.

WEKA Exercise 3. Improving the network supplied as a file

- Bring in the window of ***classifier's options***
  - Type in the ***BIFF file*** box: *weather.xml*
- ***Run***
- In the output window find the comparison between two networks:
  - the supplied and inferred by machine learning.





WEKA Exercise 4. Structure supplied by the user

- Bring in the window of ***classifier's options***
- In ***searchAlgorithm*** row press ***Choose*** button
- Choose search->fixed->FromFile. OK
- Press ***searchAlgorithm*** row to define parameters
- Type in the ***BIFF file*** box: *weather.xml (Do NOT use the button Open…)*
- ***Run***
- Check that WEKA has produced the Naïve Bayes, as it was supplied in your file