Classification

- Part 3 -

Outline

1. What is Classification?
2. K-Nearest-Neighbors
3. Decision Trees
4. Regression
5. Naïve Bayes

6. Naïve Bayes

- Probabilistic classification technique based on Bayes theorem
  - widely used and especially successful at classifying texts
- Goal: Estimate the most probable class label for a given record
- Probabilistic formulation of the classification task:
  - consider each attribute and class label as random variables
  - given a record with attributes $(A_1, A_2,\ldots,A_n)$,
    the goal is to find the class C that maximizes the conditional probability

$$P(C \mid A_1, A_2,\ldots,A_n)$$

- Example: Should we play golf?
  - P(Play=yes | Outlook=rainy, Temperature=cool)
  - P(Play=no | Outlook=rainy, Temperature=cool)
- Question: How to estimate these probabilities given training data?

Bayes Theorem

- Thomas Bayes (1701-1761)
  - British mathematician and priest
  - tried to formally prove the existence of God
- Bayes Theorem

$$P(C|A) = \frac{P(A|C)P(C)}{P(A)}$$

- useful in situations where P(C|A) is unknown
  while P(A|C), P(A) and P(C) are known or easy to estimate

Bayes Theorem: Evidence Formulation

- Prior probability of event H:
  - probability of event <u>before</u> evidence is seen
  - we play golf in 70% of all cases ➔ P(H) = 0.7
- Posterior probability of event H:
  - probability of event ***after*** evidence is seen
  - evidence: It is windy and raining ➔ P(H | E) = 0.2
- Probability of event *H* given evidence *E*:

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)}$$

Applying Bayes Theorem to the Classification Task

Evidence = record

Class-conditional probability of evidence

$$P(C \mid A) = \frac{P(A \mid C)P(C)}{P(A)}$$

Prior probability of class

Class

Prior probability of evidence

1. Compute the probability P(C | A) for all values of C using Bayes theorem.
   - P(A) is the same for all classes. Thus, we just need to estimate P(C) and P(A|C)
2. Choose value of C that maximizes P(C | A).

Example:

$$P(\text{Play}=\text{yes} \mid \text{Outlook}=\text{rainy}, \text{Temp}=\text{cool}) = \frac{P(\text{Outlook}=\text{rainy}, \text{Temp}=\text{cool} \mid \text{Play}=\text{yes})P(\text{Play}=\text{yes})}{P(\text{Outlook}=\text{rainy}, \text{Temp}=\text{cool})}$$

$$P(\text{Play}=\text{no} \mid \text{Outlook}=\text{rainy}, \text{Temp}=\text{cool}) = \frac{P(\text{Outlook}=\text{rainy}, \text{Temp}=\text{cool} \mid \text{Play}=\text{no})P(\text{Play}=\text{no})}{P(\text{Outlook}=\text{rainy}, \text{Temp}=\text{cool})}$$

Estimating the Prior Probability P(C)

– The prior probability $P(C_j)$ for each class is estimated by

1. counting the records in the training set that are labeled with class C_j
2. dividing the count by the overall number of records

– Example:

- P(Play=no) = 5/14
- P(Play=yes) = 9/14

Training 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**

Estimating the Class-Conditional Probability P(A | C)

– Naïve Bayes assumes that all attributes are *statistically independent*
  - knowing the value of one attribute says nothing about the value of another
  - this independence assumption is almost never correct!
  - but … this scheme works well in practice

– The independence assumption allows the joint probability P(A | C) to be reformulated as the product of the individual probabilities $P(A_i | C_j)$:

$$P(A_1, A_2, \ldots, A_n \mid C_j) = \prod P(A_n | C_j) = P(A_1 | C_j) \times P(A_2 | C_j) \times \ldots \times P(A_n | C_j)$$

P(Outlook=rainy, Temperature=cool | Play=yes) = P(Outlook=rainy | Play=yes) × P(Temperature=cool | Play=yes)

– Result: The probabilities $P(A_i | C_j)$ for all A_i and C_j can be estimated directly from the training data

Estimating the Probabilities $P(A_i | C_j)$

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	**2/9**	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

The probabilities $P(A_i | C_j)$ are estimated by

1. counting how often an attribute value appears together with class C_j
2. dividing the count by the overall number of records belonging to class C_j

Example:
2 times "Yes" together with "Outlook=sunny" out of altogether 9 "Yes" examples
➔ p(Outlook=sunny|Yes) = 2/9

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

Classifying a New Day

Unseen record

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

Class-conditional probability of the evidence

$$P(yes \mid E) = P(Outlook = Sunny \mid yes)$$
$$\times P(Temperature = Cool \mid yes)$$
$$\times P(Humidity = High \mid yes)$$
$$\times P(Windy = True \mid yes)$$
$$\times \frac{P(yes)}{P(E)}$$

Probability of class "yes" given the evidence

Prior probability of class "yes"

Prior probability of evidence

$$= \frac{\frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14}}{P(E)}$$

Classifying a New Day: Weigh the Evidence!

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

– A new day:

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

Prior probability

Evidence

Likelihood of the two classes

For "yes" = $2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14 = 0.0053$

For "no" = $3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0.0206$

Conversion into a probability by normalization:

P("yes") = 0.0053 / (0.0053 + 0.0206) = 0.205

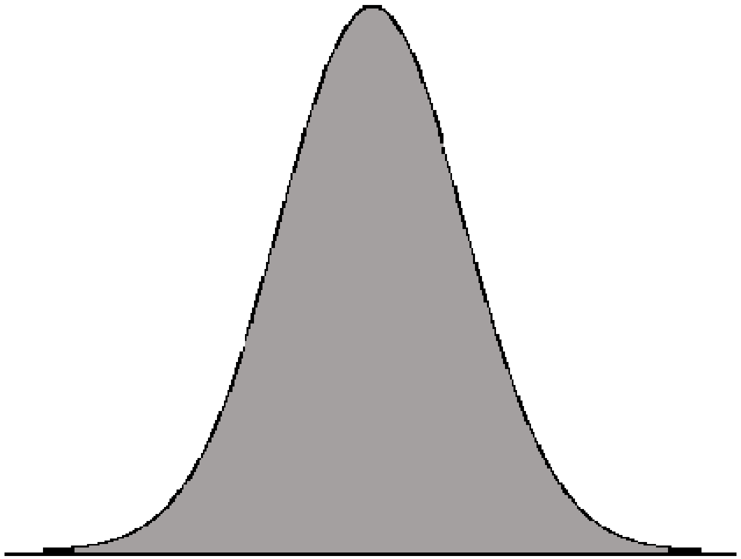
Choose Maximum

P("no") = 0.0206 / (0.0053 + 0.0206) = 0.795

Handling Numerical Attributes

– Option 1:
  Discretize numerical attributes before learning classifier.
  - Temp= 37°C ➔ “Hot”
  - Temp= 21°C ➔ “Mild”

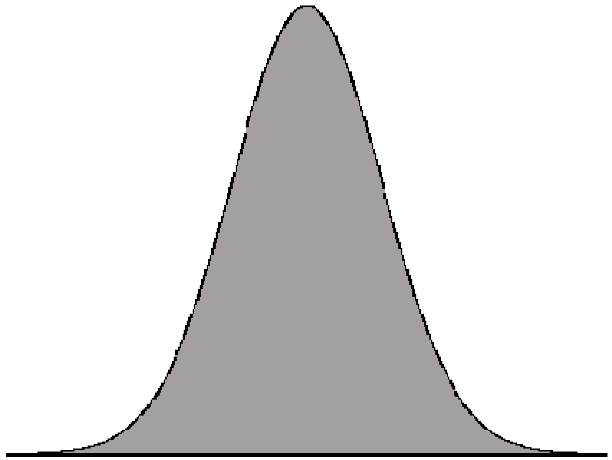
– Option 2:
  Make assumption that numerical attributes have a normal distribution given the class.
  - use training data to estimate parameters of the distribution (e.g., mean and standard deviation)
  - once the probability distribution is known, it can be used to estimate the conditional probability $P(A_i|C_j)$



Handling Numerical Attributes

– The probability density function for the normal distribution is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



– It is defined by two parameters:

- *Sample mean μ*

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- *Standard deviation σ*

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2}$$

– Both parameters can be estimated from the training data

First Example

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Normal distribution:

$$P(A_i|c_j) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(A_i-\mu\ )^2}{2\sigma_{ij}^2}}$$

One for each (A_i , c_i) pair

For (Income, Class=No):

If Class=No

sample mean = 110

sample variance = 2975

π = (125 + 100+70+….+75) /7 = 110

σ^2 = [(125-110)^2 +(100-110)^2 +…….+(75-110)^2] / (7-1=6)

$15^2+10^2+40^2+10^2+50^2+110^2+35^2$ = 225+100+1600+100+2500+12100+1225=17850

17850/6 =2975

σ = 54.54

First Example

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Normal distribution:

$$P(A_i|c_j) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(A_i-\mu\ \ )^2}{2\sigma_{ij}^2}}$$

One for each (A_i, c_j) pair

For (Income, Class=No):

If Class=No

sample mean = 110

sample variance = 2975

$$P(Income = 120 \mid No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

Second Example

Outlook			Temperature		Humidity		Windy			Play	
	Yes	*No*	*Yes*	*No*	*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	64, 68,	65, 71,	65, 70,	70, 85,	False	6	2	9	5
Overcast	4	0	69, 70,	72, 80,	70, 75,	90, 91,	True	3	3		
Rainy	3	2	72, ...	85, ...	80, ...	95, ...					
Sunny	2/9	3/5	μ =73	μ =75	μ =79	μ =86	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	σ =6.2	σ =7.9	σ =10.2	σ =9.7	True	3/9	3/5		
Rainy	3/9	2/5									

Example calculation:

$$f(temp = 66 \mid yes) = \frac{1}{\sqrt{2\pi}6.2} e^{-\frac{(66-73)^2}{2*6.2^2}} = 0.0340$$

Classifying a New Day

Unseen record

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

Likelihood of "yes" = $2/9 \times 0.0340 \times 0.0221 \times 3/9 \times 9/14 = 0.000036$

Likelihood of "no" = $3/5 \times 0.0291 \times 0.0380 \times 3/5 \times 5/14 = 0.000136$

P("yes") = 0.000036 / (0.000036 + 0. 000136) = 20.9%

P("no") = 0.000136 / (0.000036 + 0. 000136) = 79.1%

But note: Some numeric attributes are not normally distributed and you may thus need to choose a different probability density function or use discretization

Handling Missing Values

- Missing values may occur in training and in unseen classification records
- Training: Record is not included into frequency count for attribute value-class combination
- Classification: Attribute will be omitted from calculation
  - Example:

Unseen record

Outlook	Temp.	Humidity	Windy	Play
?	Cool	High	True	?

Likelihood of "yes" = $3/9 \times 3/9 \times 3/9 \times 9/14 = 0.0238$

Likelihood of "no" = $1/5 \times 4/5 \times 3/5 \times 5/14 = 0.0343$

P("yes") = 0.0238 / (0.0238 + 0.0343) = 41%

P("no") = 0.0343 / (0.0238 + 0.0343) = 59%

The Zero-Frequency Problem

- What if an attribute value doesn't occur with every class value?
  (e.g. no "Outlook = overcast" for class "no")
  - class-conditional probability will be zero!

$$P[Out. = overc. \mid no] = \frac{0}{5} = 0$$

- Problem: Posterior probability will also be zero!
  No matter how likely the other values are!

$$P[no \mid E] = 0$$

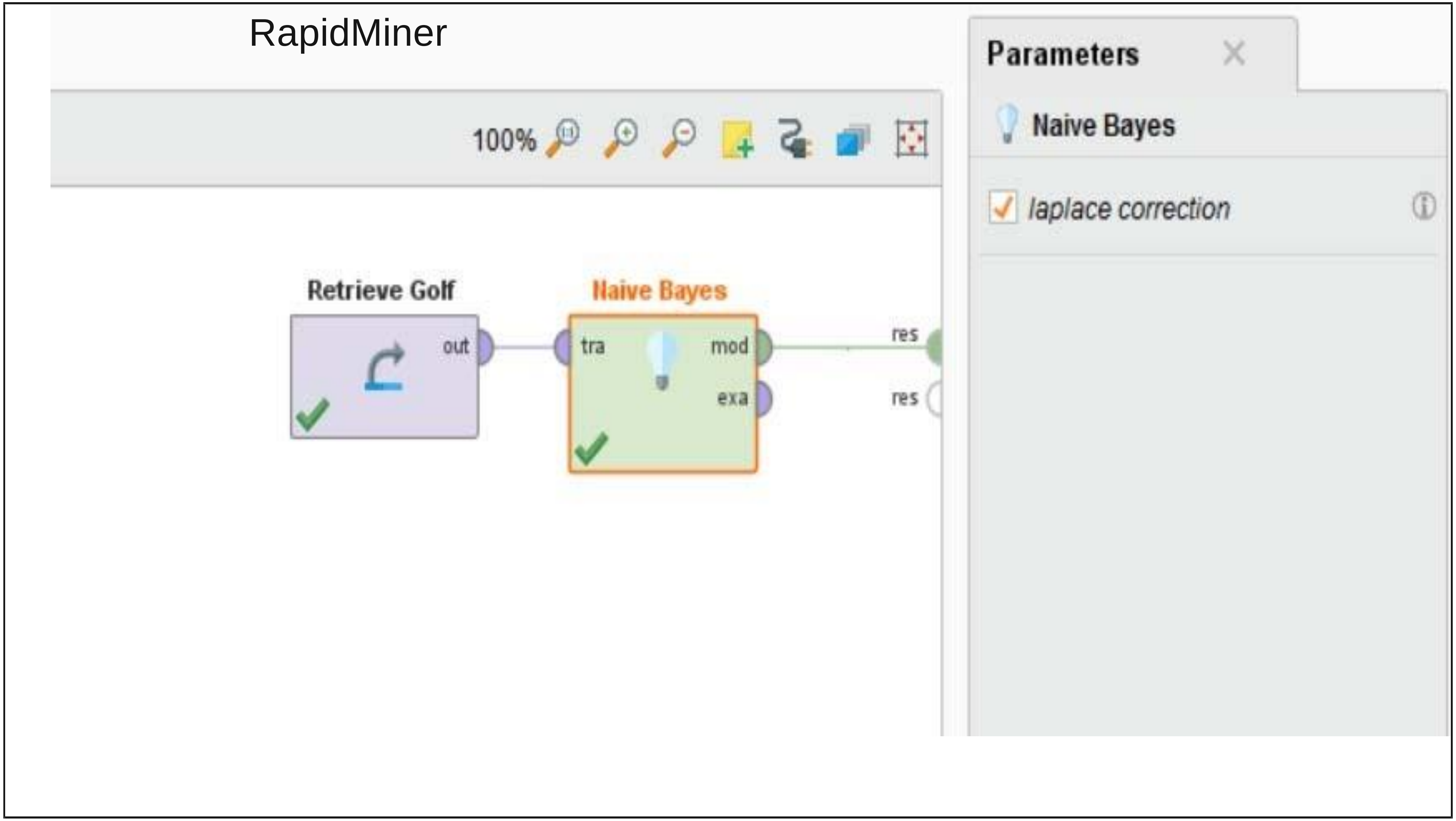
- Remedy: Add 1 to the count for every attribute value-class combination (*Laplace Estimator*)
- Result: Probabilities will never be zero!
  also: stabilizes probability estimates

$$\text{Original}: P(A_i \mid C) = \frac{N_{ic}}{N_c}$$

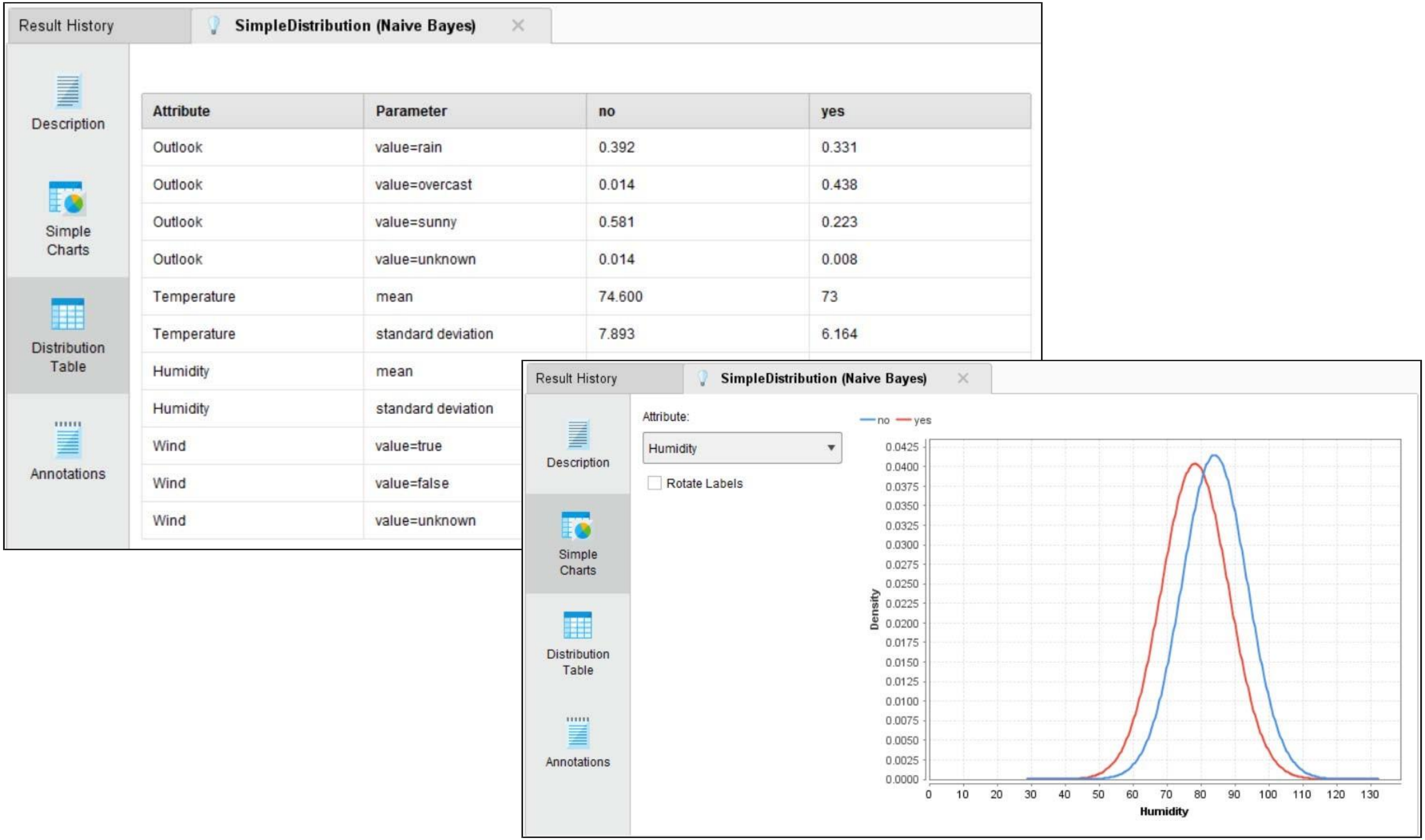
$$\text{Laplace}: P(A_i \mid C) = \frac{N_{ic} + 1}{N_c + |V_i|}$$

$|V_i|$ number of values

Naïve Bayes in RapidMiner



Naïve Bayes in RapidMiner: Probability Distribution Table



Characteristics of Naïve Bayes

- Naïve Bayes works surprisingly well for many classification tasks
  - even if independence assumption is clearly violated
  - Why? Because classification doesn't require accurate probability estimates as long as maximum probability is assigned to correct class
- Robust to isolated noise points as they will be averaged out
- Robust to irrelevant attributes as $P(A_i | C)$ distributed uniformly for A_i
- Adding too many redundant attributes can cause problems
  - Solution: Select attribute subset as Naïve Bayes often works better with just a fraction of all attributes
- Technical advantages
  - Learning Naïve Bayes classifiers is computationally cheap as probabilities can be estimated doing one pass over the training data
  - Storing the probabilities does not require a lot on memory