Chair of Decision Sciences and Systems
TUM School of Computation, Information and Technology
Technical University of Munich

Business Analytics & Machine Learning
Homework sheet 4: Naïve Bayes – Solution

Prof. Dr. Martin Bichler
Julius Durmann, Markus Ewert, Yutong Chao, Dr. Mete Ahunbay

Exercise H4.1 *Living situation*

The following shows data of the living situation of different people, depending on their social context.

#	Job	Merital Status	Children	Living Situation
1	employed	married	yes	rent
2	employed	single	no	property
3	employed	married	no	rent
4	employed	married	yes	property
5	freelance	single	yes	property
6	freelance	single	no	rent
7	freelance	married	yes	property
8	freelance	married	no	property

a) Compute a priori probabilities for the classes "rent" and "property".

b) Compute all conditional probabilities for the conditions "rent" and "property".

c) Classify a single employee with child using Naive Bayes Classification.

Solution

In the following, the random variables are abreviated by their starting latters, i.e. "Job" = "J", "Merital Status" = "MS", "Children" = "C", "Living Situation" = "LS".

a) $\mathcal{P}(LS = rent) = 3/8$
$\mathcal{P}(LS = property) = 5/8$

b)
- Job:
$\mathcal{P}(J = employed|LS = rent) = 2/3$
$\mathcal{P}(J = freelance|LS = rent) = 1/3$
$\mathcal{P}(J = employed|LS = property) = 2/5$
$\mathcal{P}(J = freelance|LS = property) = 3/5$
- Merital Status:
$\mathcal{P}(MS = married|LS = rent) = 2/3$
$\mathcal{P}(MS = single|LS = rent) = 1/3$
$\mathcal{P}(MS = married|LS = property) = 3/5$
$\mathcal{P}(MS = single|LS = property) = 2/5$
- Children:
$\mathcal{P}(C = yes|LS = rent) = 1/3$
$\mathcal{P}(C = no|LS = rent) = 2/3$

$\mathcal{P}(C = yes|LS = property) = 3/5$
$\mathcal{P}(C = no|LS = property) = 2/5$

c) $\mathcal{P}(LS = rent|J = employed, MS = single, C = yes) =$
$1/\gamma \cdot \mathcal{P}(J = employed|LS = rent) \cdot \mathcal{P}(MS = single|LS = rent)$
$\cdot\, \mathcal{P}(C = yes|LS = rent) \cdot \mathcal{P}(LS = rent) =$
$1/\gamma \cdot 2/3 \cdot 1/3 \cdot 1/3 \cdot 3/8 = 1/\gamma \cdot 6/216 = 1/\gamma \cdot 1/36$

$\mathcal{P}(LS = property|J = employed, MS = single, C = yes) =$
$1/\gamma \cdot \mathcal{P}(J = employed|LS = property) \cdot \mathcal{P}(MS = single|LS = property)$
$\cdot\, \mathcal{P}(C = yes|LS = property) \cdot \mathcal{P}(LS = property) =$
$1/\gamma \cdot 2/5 \cdot 2/5 \cdot 3/5 \cdot 5/8 = 1/\gamma \cdot 60/1000 = 1/\gamma \cdot 3/50$

Thus, given the above data, a Naive Bayes Classifier would classify a single employee with child as a property owner. In a pure classifying task, the normalization constant γ does not have to be computed, as we are only interested in the maximum value, and not the specific probabilities.

Exercise H4.2 *To play or not to play?*

The following table contains data of past decisions, on whether or not to play, depending on weather conditions.

#	Outlook	Temperature	Humidity	Wind	Play
1	sunny	hot	high	false	no
2	sunny	hot	high	true	no
3	overcast	hot	high	false	yes
4	rainy	mild	high	false	yes
5	rainy	cool	normal	false	yes
6	rainy	cool	normal	true	no
7	overcast	cool	normal	true	yes
8	sunny	mild	high	false	no
9	sunny	cool	normal	false	yes
10	rainy	mild	normal	false	yes
11	sunny	mild	normal	true	yes
12	overcast	mild	high	true	yes
13	overcast	hot	normal	false	yes
14	rainy	mild	high	true	no

a) Use Naive Bayes Classification to decide on whether to play or not, given the following conditions:

#	Outlook	Temperature	Humidity	Wind	Play
15	sunny	mild	normal	false	??
16	rainy	hot	high	true	??
17	overcast	cool	normal	false	??

b) Create a bayesian network representing the assumptions of the Naive Bayes Classification from a).

Solution

a)

- The task is to classify datapoints #15, #16 and #17, with possible lables "yes" and "no", using Naive Bayes Classifier. First, build the likelihood tables for each feature:

Outlook	Play = no	Play = yes
sunny	$3/5$	$2/9$
overcast	$0/5$	$4/9$
rainy	$2/5$	$3/9$

Temperature	Play = no	Play = yes
hot	$2/5$	$2/9$
mild	$2/5$	$4/9$
cool	$1/5$	$3/9$

Humidity	Play = no	Play = yes
high	$4/5$	$3/9$
normal	$1/5$	$6/9$

Wind	Play = no	Play = yes
false	$2/5$	$6/9$
true	$3/5$	$3/9$

In the following, the random variables are abbreviated by their first letter, i.e. "Play" = "P", "Outlook" = "O", "Temperature" = "T", "Humidity" = "H", "Wind" = "W".

- #15:

$$\mathcal{P}(P = no|\underbrace{O = sunny, T = mild, H = normal, W = false}_{=:e_{15}})$$
$$= 1/\gamma \cdot \mathcal{P}(O = sunny|P = no) \cdot \mathcal{P}(T = mild|P = no) \cdot \mathcal{P}(H = normal|P = no)$$
$$\cdot \mathcal{P}(W = false|P = no) \cdot \mathcal{P}(P = no)$$
$$= 1/\gamma \cdot 3/5 \cdot 2/5 \cdot 1/5 \cdot 2/5 \cdot 5/14$$
$$= 1/\gamma \cdot 60/8750 \approx 0.00686$$
$$\mathcal{P}(P = yes|O = sunny, T = mild, H = normal, W = false)$$
$$= 1/\gamma \cdot \mathcal{P}(O = sunny|P = yes) \cdot \mathcal{P}(T = mild|P = yes) \cdot \mathcal{P}(H = normal|P = yes)$$
$$\cdot \mathcal{P}(W = false|P = yes) \cdot \mathcal{P}(P = yes)$$
$$= 1/\gamma \cdot 2/9 \cdot 4/9 \cdot 6/9 \cdot 6/9 \cdot 9/14$$
$$= 1/\gamma \cdot 2592/91854 \approx 0.0282$$

With $\gamma = 0.00686 + 0.0282 = 0.03506$, the corresponding probabilities are
$\mathcal{P}(P = no|e_{15}) = 0.00686/0.03506 \approx 0.196$, and
$\mathcal{P}(P = yes|e_{15}) = 0.0282/0.03506 \approx 0.804$

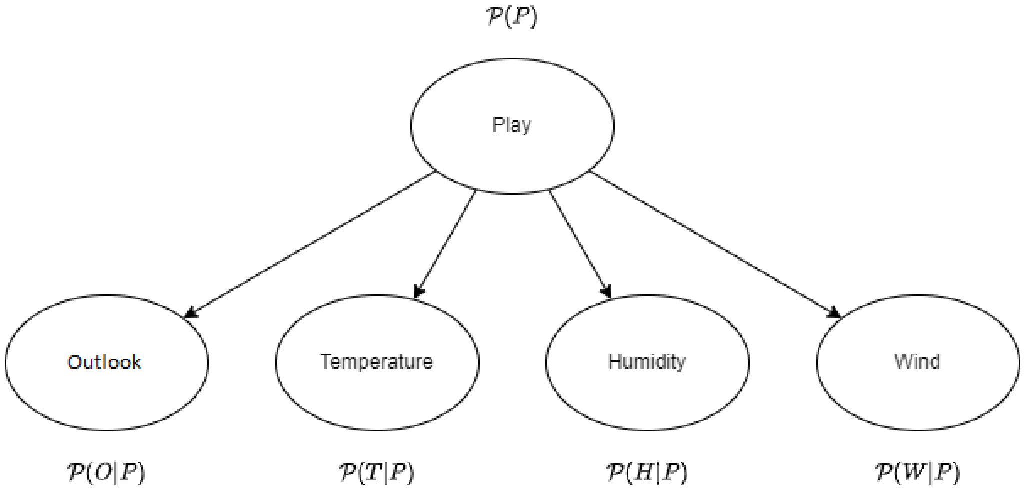
- #16: (similar to #15)

$$\mathcal{P}(P = no|\underbrace{O = rainy, T = hot, H = high, W = true}_{=:e_{16}})$$
$$= 1/\gamma \cdot \mathcal{P}(O = rainy|P = no) \cdot \mathcal{P}(T = hot|P = no) \cdot \mathcal{P}(H = high|P = no)$$
$$\cdot \mathcal{P}(W = true|P = no) \cdot \mathcal{P}(P = no)$$
$$= 1/\gamma \cdot 2/5 \cdot 2/5 \cdot 4/5 \cdot 3/5 \cdot 5/14$$
$$= 1/\gamma \cdot 240/8750 \approx 0.0274$$
$$\mathcal{P}(P = yes|O = rainy, T = hot, H = high, W = true)$$
$$= 1/\gamma \cdot \mathcal{P}(O = rainy|P = yes) \cdot \mathcal{P}(T = hot|P = yes) \cdot \mathcal{P}(H = high|P = yes)$$
$$\cdot \mathcal{P}(W = true|P = yes) \cdot \mathcal{P}(P = yes)$$
$$= 1/\gamma \cdot 3/9 \cdot 2/9 \cdot 3/9 \cdot 3/9 \cdot 9/14$$
$$= 1/\gamma \cdot 486/91854 \approx 0.00529$$

With $\gamma = 0.00529 + 0.0274 = 0.03269$, this leads to
$\mathcal{P}(P = no|e_{16}) = 0.0274/0.03269 \approx 0.84$, and
$\mathcal{P}(P = yes|e_{16}) = 0.00529/0.03269 \approx 0.16.$

- #17: (Zero Frequency Problem)
  Since there are no occurences of "$O = overcast|P = no$", the corresponding likelihood is zero, i.e. $\mathcal{P}(O = overcast|P = no) = 0$. This will always result in $\mathcal{P}(P = no|O = overcast, \ldots) = 0$. To fix this, one can add "+1" to each entry of the frequency tables, and then build the corresponding likelihood tables.
  $\mathcal{P}(P = no|\underbrace{O = overcast, T = cool, H = normal, W = false}_{=:e_{17}})$
  = $^1/_\gamma \cdot \mathcal{P}(O = overcast|P = no) \cdot \mathcal{P}(T = cool|P = no) \cdot \mathcal{P}(H = normal|P = no)$
  $\cdot \mathcal{P}(W = false|P = no) \cdot \mathcal{P}(P = no)$
  = $^1/_\gamma \cdot {}^{(0+1)}/_{(5+3)} \cdot {}^{(1+1)}/_{(5+3)} \cdot {}^{(1+1)}/_{(5+2)} \cdot {}^{(2+1)}/_{(5+2)} \cdot {}^{5}/_{14}$
  = $^1/_\gamma \cdot {}^{60}/_{43904} \approx 0.00137$
  $\mathcal{P}(P = yes|O = overcast, T = cool, H = normal, W = false)$
  = $^1/_\gamma \cdot \mathcal{P}(O = overcast|P = yes) \cdot \mathcal{P}(T = cool|P = yes) \cdot \mathcal{P}(H = normal|P = yes)$
  $\cdot \mathcal{P}(W = false|P = yes) \cdot \mathcal{P}(P = yes)$
  = $^1/_\gamma \cdot {}^{(4+1)}/_{(9+3)} \cdot {}^{(3+1)}/_{(9+3)} \cdot {}^{(6+1)}/_{(9+2)} \cdot {}^{(6+1)}/_{(9+2)} \cdot {}^{9}/_{14}$
  = $^1/_\gamma \cdot {}^{8820}/_{243936} \approx 0.0362$
  With $\gamma = 0.00137 + 0.0362 = 0.03757$, this leads to
  $\mathcal{P}(P = no|e_{17}) = {}^{0.00137}/_{0.03757} \approx 0.036$, and
  $\mathcal{P}(P = yes|e_{17}) = {}^{0.0362}/_{0.03757} \approx 0.964$.

b) As independent variables are an underlying assumption of Naive Bayes Classifier, the corresponding Bayesian Network never exceeds depth 1, and has the following shape:


$\mathcal{P}(P)$
Play
Outlook
Temperature
Humidity
Wind
$\mathcal{P}(O|P)$
$\mathcal{P}(T|P)$
$\mathcal{P}(H|P)$
$\mathcal{P}(W|P)$