Recap

- Last Tuesday
  - Machine Learning Experiment
  - Convert data set into correct format
  - Run machine learning workflow
    - `weka.classifiers.Evaluation`
    - `weka.classifiers.bayes.NaiveBayes`

Lehrevaluation

- Antworten einsehbar in Ilias
- Vielen Dank für die Blumen 🎂

Lehrevaluation

- Antworten einsehbar in Ilias
- Vielen Dank für die Blumen
- Wichtigste Punkte aus Kommentaren
  - Tempo der Veranstaltung
  - Verhältnis Übungen und Klausur, Inhalt der Übungen
  - Vorlesungsmodus

Lehrevaluation

- Antworten einsehbar in Ilias
- Vielen Dank für die Blumen
- Wichtigste Punkte aus Kommentaren
  - Tempo der Veranstaltung
  - Verhältnis Übungen und Klausur, Inhalt der Übungen
  - Vorlesungsmodus
- Weitere Fragen oder Anmerkungen?

Naive Bayes

Sprachverarbeitung (VL + Ü)

Nils Reiter

May 23, 2023

INSTITUT FÜR
DIGITAL HUMANITIES
UNIVERSITÄT ZU KÖLN

Recap: Probabilities

- Probability: Ratio of events of interest to all possible events (within event space)
- Joint probability: Two events take place simultaneously
- Conditional probability: One event takes place under the assumption that another event took place
  - Can be calculated from joint and individual probabilities: $\boxed{p(A|B) = \frac{p(A,B)}{p(B)}}$
- Dependent and independent events

Conditional and Joint Probabilities

Example

Relation between **hair color** H and preferred **wake-up time** W[1]

$\downarrow W \; / \; H \rightarrow$	brown	red	sum
early	20	10	30
late	30	5	35
sum	50	15	65

Table: Experimental Results, Ω: Group of questioned people, $|\Omega| = 65$

[1] All numbers are made up.

Conditional and Joint Probabilities

Example

Relation between **hair color** H and preferred **wake-up time** W[1]

$\downarrow W$ / $H \rightarrow$	brown	red	sum
early	20	10	30
late	30	5	35
sum	50	15	65

Table: Experimental Results, Ω: Group of questioned people, $|\Omega| = 65$

- If we pick a random person, what's the probability that this person has brown hair?
- $$p(H = \text{brown}) = ? \quad \frac{50}{65}$$

[1] All numbers are made up.

Conditional and Joint Probabilities

Example

Relation between **hair color** H and preferred **wake-up time** W[1]

$\downarrow W$ / $H \rightarrow$	brown	red	sum
early	20	10	30
late	30	5	35
sum	50	15	65

Table: Experimental Results, Ω: Group of questioned people, $|\Omega| = 65$

$$\left.\begin{array}{ll} p(H = \text{brown}) = \frac{50}{65} & p(H = \text{red}) = \frac{15}{65} \\ p(W = \text{early}) = \frac{30}{65} & p(W = \text{late}) = \frac{35}{65} \end{array}\right\} \text{sums per row or column}$$

[1] All numbers are made up.

Conditional and Joint Probabilities

Example

Relation between **hair color** H and preferred **wake-up time** W[1]

$\downarrow W$ / $H \rightarrow$	brown	red	sum
early	20	10	30
late	30	5	35
sum	50	15	65

Table: Experimental Results, Ω: Group of questioned people, $|\Omega| = 65$

- Joint probability: $p(W = \text{late}, H = \text{brown}) = \frac{30}{65}$
  - Probability that someone has brown hair *and* prefers to wake up late
  - Denominator: Number of all items

[1] All numbers are made up.

Conditional and Joint Probabilities

Example

Relation between **hair color** H and preferred **wake-up time** W[1]

$\downarrow W \;/\; H \rightarrow$	brown	red	sum
early	20	10	30
late	30	5	35
sum	50	15	65

Table: Experimental Results, Ω: Group of questioned people, $|\Omega| = 65$

- Joint probability: $p(W = \text{late}, H = \text{brown}) = \frac{30}{65}$
  - Probability that someone has brown hair *and* prefers to wake up late
  - Denominator: Number of all items
- Conditional probability: $p(W = \text{late} | H = \text{brown}) = \frac{30}{50}$
  - Probability that one of the brown-haired participants prefers to wake up late
  - Denominator: Number of remaining items (after conditioned event has happened)

[1] All numbers are made up.

Conditional and Joint Probabilities

Example

	brown $\frac{20}{65}$	red	margin
early	$p(W=e, H=b) = 0.31$	$p(W=e, H=r) = 0.15$	$p(W=e) = 0.46$
late	$p(W=l, H=b) = 0.46$	$p(W=l, H=r) = 0.08$	$p(W=l) = 0.54$
margin	$p(H=b) = 0.77$	$p(H=r) = 0.23$	$p(\Omega) = 1$

Table: (Joint) Probabilities, derived by dividing everything by $|\Omega|$

Conditional and Joint Probabilities

Example

	brown	red	margin
early	$p(W = e, H = b) = 0.31$	$p(W = e, H = r) = 0.15$	$p(W = e) = 0.46$
late	$p(W = l, H = b) = 0.46$	$p(W = l, H = r) = 0.08$	$p(W = l) = 0.54$
margin	$p(H = b) = 0.77$	$p(H = r) = 0.23$	$p(\Omega) = 1$

Table: (Joint) Probabilities, derived by dividing everything by $|\Omega|$

$$p(A|B) = \frac{p(A, B)}{p(B)} \quad \text{definition of conditional probabilities}$$

Conditional and Joint Probabilities

Example

	brown	red	margin
early	$p(W=e,H=b)=0.31$	$p(W=e,H=r)=0.15$	$p(W=e)=0.46$
late	$p(W=l,H=b)=0.46$	$p(W=l,H=r)=0.08$	$p(W=l)=0.54$
margin	$p(H=b)=0.77$	$p(H=r)=0.23$	$p(\Omega)=1$

Table: (Joint) Probabilities, derived by dividing everything by $|\Omega|$

$$p(A|B) = \frac{p(A,B)}{p(B)} \quad \text{definition of conditional probabilities}$$

$$p(W=\text{late}|H=\text{brown}) = \frac{30}{50} = 0.6 \quad \text{intuition from previous slide}$$

Conditional and Joint Probabilities

Example

	brown	red	margin
early	$p(W = e, H = b) = 0.31$	$p(W = e, H = r) = 0.15$	$p(W = e) = 0.46$
late	$p(W = l, H = b) = 0.46$	$p(W = l, H = r) = 0.08$	$p(W = l) = 0.54$
margin	$p(H = b) = 0.77$	$p(H = r) = 0.23$	$p(\Omega) = 1$

Table: (Joint) Probabilities, derived by dividing everything by $|\Omega|$

$$
\begin{aligned}
p(A|B) &= \frac{p(A, B)}{p(B)} && \text{definition of conditional probabilities} \\
p(W = \text{late}|H = \text{brown}) &= \frac{30}{50} = 0.6 && \text{intuition from previous slide} \\
&= \frac{p(W = \text{late}, H = \text{brown})}{p(H = \text{brown})} && \text{by applying definition}
\end{aligned}
$$

Conditional and Joint Probabilities

Example

	brown	red	margin
early	$p(W = e, H = b) = 0.31$	$p(W = e, H = r) = 0.15$	$p(W = e) = 0.46$
late	$p(W = l, H = b) = 0.46$	$p(W = l, H = r) = 0.08$	$p(W = l) = 0.54$
margin	$p(H = b) = 0.77$	$p(H = r) = 0.23$	$p(\Omega) = 1$

Table: (Joint) Probabilities, derived by dividing everything by $|\Omega|$

$$p(A|B) = \frac{p(A,B)}{p(B)} \quad \text{definition of conditional probabilities}$$

$$p(W = \text{late}|H = \text{brown}) = \frac{30}{50} = 0.6 \quad \text{intuition from previous slide}$$

$$= \frac{p(W = \text{late}, H = \text{brown})}{p(H = \text{brown})} \quad \text{by applying definition}$$

$$= \frac{0.46}{0.77} = 0.6$$

Multiple Conditions

- Joint probabilities can include more than two events
  $p(E_1, E_2, E_3, \dots)$
- Conditional probabilities can be conditioned on more than two events

$$p(A|B, C, D) = \frac{p(A, B, C, D)}{p(B, C, D)}$$

Multiple Conditions

- Joint probabilities can include more than two events $p(E_1, E_2, E_3, \ldots)$
- Conditional probabilities can be conditioned on more than two events

$$p(A|B, C, D) = \frac{p(A, B, C, D)}{p(B, C, D)}$$

- Chain rule

$$\begin{aligned} p(A, B, C, D) &= p(A|B, C, D)p(B, C, D) \\ &= p(A|B, C, D)p(B|C, D)p(C, D) \\ &= p(A|B, C, D) \times p(B|C, D) \times p(C|D) \times p(D) \end{aligned}$$

Bayes Law

$$p(B|A) = \frac{p(A,B)}{p(A)}$$ Definition

$$= \frac{p(A|B)\,p(B)}{p(A)}$$ Kettenregel

$$\boxed{p(B|A)} = \frac{p(A,B)}{p(A)} = \boxed{\frac{p(A|B)p(B)}{p(A)}}$$

Allows reordering of conditional probabilities

- Follows directly from above definitions

Section 1

Machine Learning Algorithms

Introduction

- What is machine learning?
  - Method to find patterns, hidden structures and undetected relations in data

Introduction

- What is machine learning?
  - Method to find patterns, hidden structures and undetected relations in data
- It's all around us: Stock market transactions, search engines, surveillance, data-driven research & science, ...

Introduction

- What is machine learning?
  - Method to find patterns, hidden structures and undetected relations in data
- It's all around us: Stock market transactions, search engines, surveillance, data-driven research & science, ...
- Why is it interesting for text analysis?
  - Rule-based approaches ›don't scale‹ – hard to maintain for real texts

Introduction

- What is machine learning?
  - Method to find patterns, hidden structures and undetected relations in data
- It's all around us: Stock market transactions, search engines, surveillance, data-driven research & science, ...
- Why is it interesting for text analysis?
  - Rule-based approaches ›don't scale‹ – hard to maintain for real texts
  - Big data analyses
    - Automatic prediction of phenomena
    - Statements about 1000 texts more representative than about 10
    - Canonisation, Euro-centrism
  - Insights into data
    - By inspecting features and making error analysis

Two Parts

Prediction Model

- How do we make predictions on data instances?
- E. g.: how do we assign a part of speech tag for a word?

Learning Algorithm

- How do we create a prediction model, given annotated data?
- E. g.: how do we create a system for assigning a part of speech tag for a word?

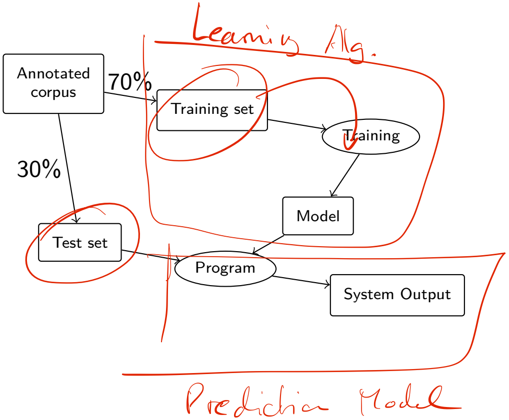
Two Parts

Prediction Model

- How do we make predictions on data instances?
- E. g.: how do we assign a part of speech tag for a word?

Learning Algorithm

- How do we create a prediction model, given annotated data?
- E. g.: how do we create a system for assigning a part of speech tag for a word?



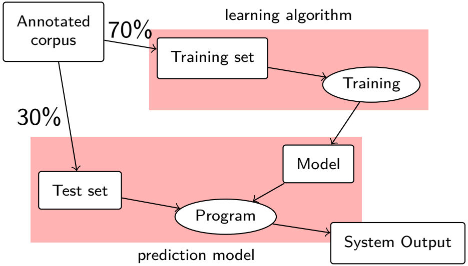
Two Parts

Prediction Model

- How do we make predictions on data instances?
- E. g.: how do we assign a part of speech tag for a word?

Learning Algorithm

- How do we create a prediction model, given annotated data?
- E. g.: how do we create a system for assigning a part of speech tag for a word?



Section 2

Naive Bayes Algorithm

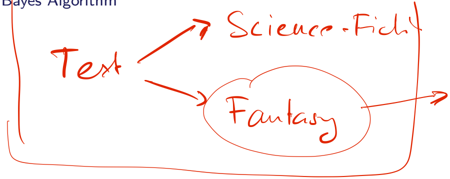
Naive Bayes

Prediction Model

- Probabilistic model (i.e., takes probabilities into account)
- Probabilities are estimated on training data (relative frequencies)

Naive Bayes

Prediction Model



Idea: We calculate the probability for each possible class c, given the feature values of the item x, and we assign most probably class

Naive Bayes

Prediction Model

Idea: We calculate the probability for each possible class c, given the feature values of the item x, and we assign most probably class

- $\boxed{f_n(x):}$ Value of feature n for instance x
- $\boxed{\operatorname{argmax}_i e:}$ Select the argument i that maximizes the expression e

Naive Bayes

Prediction Model

```
def argmax(SET, EXP):
  arg = 0
  max = 0
  foreach i in SET:
    val = EXP(i)
    if val > max:
      arg = i
      max = val
  return arg
```

Idea: We calculate the probability for each possible class c, given the f
item x, and we assign most probably class

- $f_n(x)$: Value of feature n for instance x
- $\operatorname{argmax}_i e$: Select the argument i that maximizes the expression e

Naive Bayes

Prediction Model

```
def argmax(SET, EXP):
  arg = 0
  max = 0
  foreach i in SET:
    val = EXP(i)
    if val > max:
      arg = i
      max = val
  return arg
```

Idea: We calculate the probability for each possible class c, given the f
item x, and we assign most probably class

- $f_n(x)$: Value of feature n for instance x
- $\operatorname{argmax}_i e$: Select the argument i that maximizes the expression e

	Name	Embarked	Cab
1212	Smith, May	S	15

$$\text{prediction}(x) = \operatorname*{argmax}_{c \in C} p(c|f_1(x), f_2(x), \ldots, f_n(x))$$

$p(0 | \text{Name} = \text{Smith}, \text{Embark} = S, \text{Cab} = 15)$

$p(1 | \text{Name} = \text{Smith}, \text{Emb} = S, \text{Cab} = 15)$

Naive Bayes

Prediction Model

```
def argmax(SET, EXP):
  arg = 0
  max = 0
  foreach i in SET:
    val = EXP(i)
    if val > max:
      arg = i
      max = val
  return arg
```

Idea: We calculate the probability for each possible class c, given the f
item x, and we assign most probably class

- $f_n(x)$: Value of feature n for instance x
- $\operatorname{argmax}_i e$: Select the argument i that maximizes the expression e

$$\text{prediction}(x) = \operatorname*{argmax}_{c \in C} p(c | f_1(x), f_2(x), \ldots, f_n(x))$$

How do we calculate $p(c | f_1(x), f_2(x), \ldots, f_n(x))$?

Naive Bayes

Prediction Model

$$p(c|f_1,\ldots,f_n) = \frac{p(c, f_1, \ldots, f_n)}{p(f_1,\ldots,f_n)} = \frac{p(f_1,\ldots,f_n,c)}{p(f_1,\ldots,f_n)}$$

$$= \frac{p(f_1|f_2,\ldots,f_n,c) \cdot p(f_2|f_3,\ldots,f_n,c) \cdot \ldots \cdot p(c)}{p(f_1,\ldots,f_n)}$$

naive $\Rightarrow$

$$= \frac{p(f_1|c) \cdot p(f_2|c) \cdot p(f_3|c) \cdots p(c)}{p(f_1,\ldots,f_n)}$$

Naive Bayes

Prediction Model

$$p(c|f_1,\ldots,f_n) = \frac{p(c,f_1,f_2,\ldots,f_n)}{p(f_1,f_2,\ldots,f_n)}$$

Naive Bayes

Prediction Model

$$p(c|f_1,\ldots,f_n) \quad = \quad \frac{p(c,f_1,f_2,\ldots,f_n)}{p(f_1,f_2,\ldots,f_n)} = \frac{p(f_1,f_2,\ldots,f_n,c)}{p(f_1,f_2,\ldots,f_n)}$$

Naive Bayes

Prediction Model

$$p(c|f_1,\ldots,f_n) = \frac{p(c,f_1,f_2,\ldots,f_n)}{p(f_1,f_2,\ldots,f_n)} = \frac{p(f_1,f_2,\ldots,f_n,c)}{p(f_1,f_2,\ldots,f_n)}$$

Application of chain rule

$$= \frac{p(f_1|f_2,\ldots,f_n,c)\times p(f_2|f_3,\ldots,f_n,c)\times\cdots\times p(c)}{p(f_1,f_2,\ldots,f_n)}$$

Naive Bayes

Prediction Model

$$p(c|f_1,\ldots,f_n) = \frac{p(c,f_1,f_2,\ldots,f_n)}{p(f_1,f_2,\ldots,f_n)} = \frac{p(f_1,f_2,\ldots,f_n,c)}{p(f_1,f_2,\ldots,f_n)}$$

Application of chain rule

$$= \frac{p(f_1|f_2,\ldots,f_n,c)\times p(f_2|f_3,\ldots,f_n,c)\times\cdots\times p(c)}{p(f_1,f_2,\ldots,f_n)}$$

Now we – naively – assume feature independence

$$= \frac{p(f_1|c)\times p(f_2|c)\times\cdots\times p(c)}{p(f_1,f_2,\ldots,f_n)}$$

Naive Bayes

Prediction Model

$p(\text{Smith}|1) \times p(S|1) \times p(15|1)$

$p(0 \mid f_1 = \text{Smith}, f_2 = S, f_3 = 15)$

$p(1 \mid f_1 = \text{Smith}, f_2 = S, f_3 = 15)$

$p(\text{Smith}|0) \times p(S|0) \times p(15|0)$

From previous slide

$$p(c|f_1, \ldots, f_n) \quad = \quad \frac{p(f_1|c) \times p(f_2|c) \times \cdots \times p(c)}{p(f_1, f_2, \ldots, f_n)}$$

Naive Bayes

Prediction Model

From previous slide

$$p(c|f_1, \ldots, f_n) \quad = \quad \frac{p(f_1|c) \times p(f_2|t) \times \cdots \times p(c)}{p(f_1, f_2, \ldots, f_n)}$$

Skip denominator, because it's constant*

$$\text{prediction}(x) \quad = \quad \operatorname*{argmax}_{c \in C} p(f_1(x)|c) \times p(f_2(x)|c) \times \cdots \times p(c)$$

Naive Bayes

Prediction Model

* This is a hack: The largest number in $\langle 2, 6, 3 \rangle$ is the second. This doesn't change when we divide every number by the same (constant) number. The largest of $\langle 1, 3, 1.5 \rangle$ is the second, and the largest of $\langle 0.2, 0.6, 0.3 \rangle$ is also the second.

From previous slide

$$p(c|f_1, \ldots, f_n) = \frac{p(f_1|c) \times p(f_2|t) \times \cdots \times p(c)}{p(f_1, f_2, \ldots, f_n)}$$

Skip denominator, because it's constant*

$$\text{prediction}(x) = \operatorname*{argmax}_{c \in C} p(f_1(x)|c) \times p(f_2(x)|c) \times \cdots \times p(c)$$

Naive Bayes

Prediction Model

* This is a hack: The largest number in $\langle 2, 6, 3\rangle$ is the second. This doesn't change when we divide every number by the same (constant) number. The largest of $\langle 1, 3, 1.5\rangle$ is the second, and the largest of $\langle 0.2, 0.6, 0.3\rangle$ is also the second.

From previous slide

$$p(c|f_1, \ldots, f_n) = \frac{p(f_1|c) \times p(f_2|t) \times \cdots \times p(c)}{p(f_1, f_2, \ldots, f_n)}$$

Skip denominator, because it's constant*

$$\text{prediction}(x) = \operatorname*{argmax}_{c \in C} p(f_1(x)|c) \times p(f_2(x)|c) \times \cdots \times p(c)$$

Where do we get $p(f_i(x)|c)$? – Training!

Naive Bayes

Learning Algorithm

1. For each feature $f_i \in F$
   - Count frequency tables from the training set:

		C (classes)			
		c_1	c_2	...	c_m
$v(f_i)$	a	3	2	...	
	b	5	7	...	
	c	0	1	...	
	$\sum$	8	10		

Embarked

	0	1
S	100	50
C	30	130
Q	38	38
Σ	168	218

$$p(S|0) = \frac{100}{168}$$

2. Calculate conditional probabilities
   - Divide each number by the sum of the entire column
     - E.g., $p(a|c_1) = \frac{3}{3+5+0}$ $\quad p(b|c_2) = \frac{7}{2+7+1}$

Section 3

Example: Spam Classification

Training

- Data set: 100 e-mails, manually classified as spam or not spam (50/50)
  - Classes $C = \{\text{true}, \text{false}\}$
- Features: Presence of each of these tokens (manually selected): ›casino‹, ›enlargement‹, ›meeting‹, ›profit‹, ›super‹, ›text‹, ›xxx‹

casino		C true	C false
	1	45	25
	0	5	25
	$\sum$	50	50

text		C true	C false
	1	15	35
	0	35	15
	$\sum$	50	50

...

Table: Extracted frequencies for features ›casino‹ and ›text‹

$p(\text{casino}=1 \mid \text{true}) = \frac{45}{50}$

$p(\text{casino}=1 \mid \text{false}) = \frac{25}{50}$

Prediction

1. Extract word presence information from new text
2. Calculate the probability for *each possible class*

$$p\left(\text{true}\;\middle|\;\begin{bmatrix}\text{casino} & 0\\ \text{enlargement} & 0\\ \text{meeting} & 1\\ \text{profit} & 0\\ \text{super} & 0\\ \text{text} & 1\\ \text{xxx} & 1\end{bmatrix}\right)$$

Prediction

1. Extract word presence information from new text
2. Calculate the probability for *each possible class*

$$p\left(\text{true} \;\middle|\; \begin{bmatrix} \text{casino} & 0 \\ \text{enlargement} & 0 \\ \text{meeting} & 1 \\ \text{profit} & 0 \\ \text{super} & 0 \\ \text{text} & 1 \\ \text{xxx} & 1 \end{bmatrix}\right) \propto \begin{matrix} p(\text{casino} = 0|\text{true}) & \times \\ p(\text{enlargement} = 0|\text{true}) & \times \\ p(\text{meeting} = 1|\text{true}) & \times \\ p(\text{profit} = 0|\text{true}) & \times \\ p(\text{super} = 0|\text{true}) & \times \\ \boxed{p(\text{text} = 1|\text{true})} & \times \\ p(\text{xxx} = 1|\text{true}) & \times \\ p(\text{true}) & \end{matrix}$$

Prediction

1. Extract word presence information from new text
2. Calculate the probability for *each possible class*

$$
\begin{aligned}
p\left(\text{true} \;\middle|\; \begin{bmatrix} \text{casino} & 0 \\ \text{enlargement} & 0 \\ \text{meeting} & 1 \\ \text{profit} & 0 \\ \text{super} & 0 \\ \text{text} & 1 \\ \text{xxx} & 1 \end{bmatrix}\right) \propto \quad & \begin{array}{ll} p(\text{casino} = 0|\text{true}) & \times \\ p(\text{enlargement} = 0|\text{true}) & \times \\ p(\text{meeting} = 1|\text{true}) & \times \\ p(\text{profit} = 0|\text{true}) & \times \\ p(\text{super} = 0|\text{true}) & \times \\ p(\text{text} = 1|\text{true}) & \times \\ p(\text{xxx} = 1|\text{true}) & \end{array} \\
= \quad & \cdots \times \frac{5}{50} \times \cdots \times \frac{15}{50} \times \cdots = \ldots
\end{aligned}
$$

Prediction

1. Extract word presence information from new text
2. Calculate the probability for *each possible class*

$$
p\left(\text{true} \,\middle|\, \begin{bmatrix} \text{casino} & 0 \\ \text{enlargement} & 0 \\ \text{meeting} & 1 \\ \text{profit} & 0 \\ \text{super} & 0 \\ \text{text} & 1 \\ \text{xxx} & 1 \end{bmatrix}\right) \propto \begin{array}{ll} p(\text{casino} = 0|\text{true}) & \times \\ p(\text{enlargement} = 0|\text{true}) & \times \\ p(\text{meeting} = 1|\text{true}) & \times \\ p(\text{profit} = 0|\text{true}) & \times \\ p(\text{super} = 0|\text{true}) & \times \\ p(\text{text} = 1|\text{true}) & \times \\ p(\text{xxx} = 1|\text{true}) & \end{array}
$$

$$
= \cdots \times \frac{5}{50} \times \cdots \times \frac{15}{50} \times \cdots = \ldots
$$

$$
p\left(\text{false} \,\middle|\, \begin{bmatrix} \text{casino} & 0 \\ \vdots & \vdots \end{bmatrix}\right) \propto \ldots
$$

p(casino=0|false) ×
p(text=1|false) ×
...
p(false)

3. Assign the class with the higher probability

Subsection 1

Problems with Zeros

Danger

		C	
		true	false
love	1	0	35
	0	50	15
	$\sum$	50	50

- What happens in this situation to the prediction?

Danger

		C	
		true	false
love	1	0	35
	0	50	15
	$\sum$	50	50

- What happens in this situation to the prediction?
- At some point, we need to multiply with $p(\text{love} = 1|\text{true}) = 0$
- This leads to a total probability of zero (for this class), irrespective of the other features
  - Even if another feature would be a perfect predictor!

$\rightarrow$ Smoothing (as before)!

Smoothing

- Whenever multiplication is involved, zeros are dangerous
- Smoothing is used to avoid zeros
- Different possibilities
- Simple: Add something to the probabilities
  - $\frac{x_i+1}{N+1}$
  - This leads to values slightly above zero

Weka

Section 4

Summary

Summary

Two Areas: Prediction Model

Learning Algorithm

Naive Bayes

- Naive: Features Independent

- Argmax