Reflected Text Analysis beyond Linguistics

DGfS-CL fall school

Nils Reiter,
nils.reiter@ims.uni-stuttgart.de

Sept. 9-13, 2019

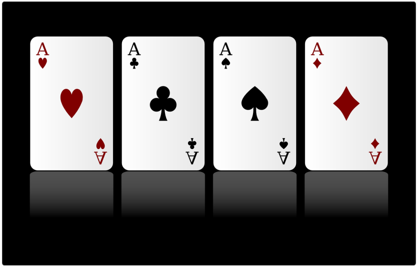
Part III

Automatisation and Machine Learning

Section 1

Probabilities

Basics: Cards



- 32 cards Ω (sample space)
- 4 'colors': $C = \{\clubsuit, \spadesuit, \diamondsuit, \heartsuit\}$
- 8 values: $V = \{7, 8, 9, 10, J, Q, K, A\}$
- Individual cards ('outcomes') are denoted with value and color: $8\heartsuit$

Basics

Events

- Generally, we draw cards from a (well shuffled) deck
- We define what events we are interested in
- An event can be any subset of the sample space Ω
- Events will be denoted with E

Basics

Events

- Generally, we draw cards from a (well shuffled) deck
- We define what events we are interested in
- An event can be any subset of the sample space Ω
- Events will be denoted with E

Examples

- 'We draw a heart eight' – $E = \{8\heartsuit\}$

Basics

Events

- Generally, we draw cards from a (well shuffled) deck
- We define what events we are interested in
- An event can be any subset of the sample space Ω
- Events will be denoted with E

Examples

- 'We draw a heart eight' – $E = \{8\heartsuit\}$
- 'We draw card with a diamond'

Basics

Events

- Generally, we draw cards from a (well shuffled) deck
- We define what events we are interested in
- An event can be any subset of the sample space Ω
- Events will be denoted with E

Examples

- 'We draw a heart eight' – $E = \{8\heartsuit\}$
- 'We draw card with a diamond' –
  $E = \{7\diamondsuit, 8\diamondsuit, 9\diamondsuit, 10\diamondsuit, J\diamondsuit, Q\diamondsuit, K\diamondsuit, A\diamondsuit\}$

Basics

Events

- Generally, we draw cards from a (well shuffled) deck
- We define what events we are interested in
- An event can be any subset of the sample space Ω
- Events will be denoted with E

Examples

- 'We draw a heart eight' – $E = \{8\heartsuit\}$
- 'We draw card with a diamond' – $E = \{7\diamondsuit, 8\diamondsuit, 9\diamondsuit, 10\diamondsuit, J\diamondsuit, Q\diamondsuit, K\diamondsuit, A\diamondsuit\}$
- 'We draw a queen'

Basics

Events

- Generally, we draw cards from a (well shuffled) deck
- We define what events we are interested in
- An event can be any subset of the sample space Ω
- Events will be denoted with E

Examples

- 'We draw a heart eight' – $E = \{8\heartsuit\}$
- 'We draw card with a diamond' – $E = \{7\diamondsuit, 8\diamondsuit, 9\diamondsuit, 10\diamondsuit, J\diamondsuit, Q\diamondsuit, K\diamondsuit, A\diamondsuit\}$
- 'We draw a queen' – $E = \{Q\clubsuit, Q\spadesuit, Q\diamondsuit, Q\heartsuit\}$

Basics

Events

- Generally, we draw cards from a (well shuffled) deck
- We define what events we are interested in
- An event can be any subset of the sample space Ω
- Events will be denoted with E

Examples

- 'We draw a heart eight' – $E = \{8\heartsuit\}$
- 'We draw card with a diamond' – $E = \{7\diamondsuit, 8\diamondsuit, 9\diamondsuit, 10\diamondsuit, J\diamondsuit, Q\diamondsuit, K\diamondsuit, A\diamondsuit\}$
- 'We draw a queen' – $E = \{Q\clubsuit, Q\spadesuit, Q\diamondsuit, Q\heartsuit\}$
- 'We draw a heart eight or diamond 10'

Basics

Events

- Generally, we draw cards from a (well shuffled) deck
- We define what events we are interested in
- An event can be any subset of the sample space Ω
- Events will be denoted with E

Examples

- 'We draw a heart eight' – $E = \{8\heartsuit\}$
- 'We draw card with a diamond' – $E = \{7\diamondsuit, 8\diamondsuit, 9\diamondsuit, 10\diamondsuit, J\diamondsuit, Q\diamondsuit, K\diamondsuit, A\diamondsuit\}$
- 'We draw a queen' – $E = \{Q\clubsuit, Q\spadesuit, Q\diamondsuit, Q\heartsuit\}$
- 'We draw a heart eight or diamond 10' – $E = \{8\heartsuit, 10\diamondsuit\}$
- 'We draw any card'

Basics

Events

- Generally, we draw cards from a (well shuffled) deck
- We define what events we are interested in
- An event can be any subset of the sample space Ω
- Events will be denoted with E

Examples

- 'We draw a heart eight' – $E = \{8\heartsuit\}$
- 'We draw card with a diamond' –
  $E = \{7\diamondsuit, 8\diamondsuit, 9\diamondsuit, 10\diamondsuit, J\diamondsuit, Q\diamondsuit, K\diamondsuit, A\diamondsuit\}$
- 'We draw a queen' – $E = \{Q\clubsuit, Q\spadesuit, Q\diamondsuit, Q\heartsuit\}$
- 'We draw a heart eight or diamond 10' – $E = \{8\heartsuit, 10\diamondsuit\}$
- 'We draw any card' – $E = \Omega$

Basics
Probabilities

- Probability $p(E)$: Likelihood, that a certain event ($E \subset \Omega$) happens
  - $0 \leq p \leq 1$
  - $p(E) = 0$: Impossible event $\qquad$ $p(E) = 1$: Certain event
  - $p(E) = 0.000001$: Very unlikely event

Basics
Probabilities

- Probability $p(E)$: Likelihood, that a certain event ($E \subset \Omega$) happens
  - $0 \leq p \leq 1$
  - $p(E) = 0$: Impossible event $\qquad$ $p(E) = 1$: Certain event
  - $p(E) = 0.000001$: Very unlikely event

Example

- If all outcomes are equally likely: $p(E) = \frac{|E|}{|\Omega|}$
- $p(\{8\heartsuit\}) = \frac{1}{32}$
- $p(\{Q\clubsuit, Q\spadesuit, Q\diamondsuit, Q\heartsuit\}) = \frac{4}{32}$
- $p(\Omega) = 1$ (must happen, certain event)

Basics I

Probability and Relative Frequency

- Probability (p): Theoretical concept, idealisation
  - Expectation
- Relative Frequency (f): Concrete measure
  - Normalised number of *observed* events
  - E.g., after 10 times drawing a card (with returning and shuffling), we counted the event ♠ eight times: $f(\{x♠\}) = \frac{8}{10}$
- For large numbers of drawings, relative frequency approximates the probability
  - $\lim_{\infty} f = p$

Basics

Joint Probability (Independent Events)

- We are often interested in multiple events (and their relation)
- E: We draw 8♡ two times in a row
    - E_1: First card is 8♡
    - E_2: Second card is 8♡
    - $p(E) = p(E_1, E_2) = p(E_1) * p(E_2) = \frac{1}{32} * \frac{1}{32} = 0.0156$

Basics

Joint Probability (Independent Events)

- We are often interested in multiple events (and their relation)
- E: We draw 8♡ two times in a row
  - E_1: First card is 8♡
  - E_2: Second card is 8♡
  - $p(E) = p(E_1, E_2) = p(E_1) * p(E_2) = \frac{1}{32} * \frac{1}{32} = 0.0156$
- E: We draw ♡ two times in a row
  - E_1: First card is X♡
  - E_2: Second card is X♡
  - $p(E) = p(E_1, E_2) = p(E_1) * p(E_2) = \frac{1}{4} * \frac{1}{4} = 0.0625$

Basics

Joint Probability (Independent Events)

- We are often interested in multiple events (and their relation)
- E: We draw 8♡ two times in a row
  - E_1: First card is 8♡
  - E_2: Second card is 8♡
  - $p(E) = p(E_1, E_2) = p(E_1) * p(E_2) = \frac{1}{32} * \frac{1}{32} = 0.0156$
- E: We draw ♡ two times in a row
  - E_1: First card is X♡
  - E_2: Second card is X♡
  - $p(E) = p(E_1, E_2) = p(E_1) * p(E_2) = \frac{1}{4} * \frac{1}{4} = 0.0625$
- So far, events have been independent
  - because we return and re-shuffle the cards all the time
  - Drawing 8♡ the first time has no influence on the second drawing

Basics

Joint Probability (Independent Events)

- We are often interested in multiple events (and their relation)
- E: We draw 8♡ two times in a row
  - E_1: First card is 8♡
  - E_2: Second card is 8♡
  - $p(E) = p(E_1, E_2) = p(E_1) * p(E_2) = \frac{1}{32} * \frac{1}{32} = 0.0156$
- E: We draw ♡ two times in a row
  - E_1: First card is X♡
  - E_2: Second card is X♡
  - $p(E) = p(E_1, E_2) = p(E_1) * p(E_2) = \frac{1}{4} * \frac{1}{4} = 0.0625$
- So far, events have been independent
  - because we return and re-shuffle the cards all the time
  - Drawing 8♡ the first time has no influence on the second drawing

Basics I

Conditional Probability (Dependent Events)

- We no longer return the card
- E: We draw $8\heartsuit$ two times in a row
  - E_1: First card is $8\heartsuit$
  - E_2: Second card is $8\heartsuit$
  - $\cancel{p(E_1, E_2) = p(E_1) * p(E_2)}$
  - This no longer works, because the events are not independent
  - There is only one $8\heartsuit$ in the game, and $p(E_2)$ has to take into account that it might be gone already
  - This is expressed with the notion of conditional probability
  - $p(E_1, E_2) = p(E_1) * p(E_2|E_1)$
    - $p(E_2|E_1) = 0$, therefore $p(E) = 0$

Basics II

Conditional Probability (Dependent Events)

- E: We draw $\heartsuit$ two times in a row
  - E_1: First card is $X\heartsuit$
  - E_2: Second card is $X\heartsuit$
  - $p(E_1, E_2) = p(E_1) * p(E_2|E_1) = \frac{8}{32} * \frac{7}{31} = 0.056$

Conditional and Joint Probabilities

Example

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

	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]

	brown	red	sum
early	20	10	30
late	30	5	35
sum	50	15	65

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

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

[1] All numbers are made up

Conditional and Joint Probabilities

Example

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

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

Conditional and Joint Probabilities

Example

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

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

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 $|\Omega|$

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 $|\Omega|$

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

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 $|\Omega|$

$$
\begin{aligned}
p(A|B) &= \frac{p(A,B)}{p(B)} \\
p(W=l|H=b) &= \frac{30}{50} = 0.6 \quad \text{from previous slide}
\end{aligned}
$$

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 $|\Omega|$

$$
\begin{aligned}
p(A|B) &= \frac{p(A,B)}{p(B)} \\
p(W=l|H=b) &= \frac{30}{50} = 0.6 \quad \text{from previous slide} \\
&= \frac{p(W=l, H=b)}{p(H=b)} \quad \text{by applying equation above}
\end{aligned}
$$

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 $|\Omega|$

$$\begin{aligned}
p(A|B) &= \frac{p(A,B)}{p(B)} \\
p(W=l|H=b) &= \frac{30}{50} = 0.6 \quad \text{from previous slide} \\
&= \frac{p(W=l,H=b)}{p(H=b)} \quad \text{by applying equation above} \\
&= \frac{0.46}{0.77} = 0.6
\end{aligned}$$

Conditional and Joint Probabilities

Random Variables

- W and H: Random variables
- Generally:
  - Random variables are functions $X : \Omega \to R$
  - Random variables map events to numbers
    - (and numbers can be assigned to categories)
- Conceptually, features can be considered as random variables

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

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

Bayes Law

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

Allows reordering of conditional probabilities

- Follows directly from above definitions

Section 2

Naive Bayes

Naive Bayes

Prediction Model

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

Naive Bayes

Prediction Model

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

(i.e., we calculate the probability for each possible class c, given the feature values of the item x, and we assign most probably class)
In our case:

$$prediction(x) = \underset{c \in \{\clubsuit\spadesuit\heartsuit\diamondsuit\}}{\operatorname{argmax}}\, p(c|f_1(x), f_2(x), \ldots, f_n(x))$$

- argmax: Select the argument that maximizes the expression
- How exactly do we calculate $p(c|f_1(x), f_2(x), \ldots, f_n(x))$?

Naive Bayes
Prediction Model

$$p(c|f_1, \ldots, f_n) \quad =$$

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

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

Naive Bayes

Prediction Model

$$
\begin{aligned}
p(c|f_1,\dots,f_n) &= \frac{p(c,f_1,f_2,\dots,f_n)}{p(f_1,f_2,\dots,f_n)} = \frac{p(f_1,f_2,\dots,f_n,c)}{p(f_1,f_2,\dots,f_n)} \\
& \text{denominator is constant, so we skip it} \\
&\propto p(f_1|f_2,\dots,f_n,c)p(f_2|f_3,\dots,f_n,c)\dots p(c)
\end{aligned}
$$

Naive Bayes

Prediction Model

$$
\begin{aligned}
p(c|f_1,\dots,f_n) &= \frac{p(c,f_1,f_2,\dots,f_n)}{p(f_1,f_2,\dots,f_n)} = \frac{p(f_1,f_2,\dots,f_n,c)}{p(f_1,f_2,\dots,f_n)} \\
& \quad \text{denominator is constant, so we skip it} \\
&\propto p(f_1|f_2,\dots,f_n,c)p(f_2|f_3,\dots,f_n,c)\dots p(c) \\
& \quad \text{Now we assume feature independence} \\
&= p(f_1|c)p(f_2|t)\dots p(c)
\end{aligned}
$$

Naive Bayes

Prediction Model

$$
\begin{aligned}
p(c|f_1,\ldots,f_n) &= \frac{p(c,f_1,f_2,\ldots,f_n)}{p(f_1,f_2,\ldots,f_n)} = \frac{p(f_1,f_2,\ldots,f_n,c)}{p(f_1,f_2,\ldots,f_n)} \\
& \quad \text{denominator is constant, so we skip it} \\
&\propto p(f_1|f_2,\ldots,f_n,c)p(f_2|f_3,\ldots,f_n,c)\ldots p(c) \\
& \quad \text{Now we assume feature independence} \\
&= p(f_1|c)p(f_2|t)\ldots p(c) \\
prediction(x) &= \operatorname*{argmax}_{c \in C} p(f_1(x)|c)p(f_2(x)|c)\ldots p(c)
\end{aligned}
$$

How do we get $p(f_i(x)|c)$? This is what the model has stored!

Naive Bayes

Learning Algorithm

- Very simple
  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		

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

Naive Bayes

Data set

$$\begin{aligned} D_{train} &= \{7\clubsuit, A\spadesuit, Q\spadesuit, K\spadesuit, J\spadesuit, 3\spadesuit, \\ &\quad 5\diamondsuit, 8\diamondsuit, 7\diamondsuit, 3\heartsuit, 7\heartsuit, 5\heartsuit\} \end{aligned}$$

Naive Bayes – Example Task

Feature f_1: Number?

		$\clubsuit$	$\spadesuit$	$\heartsuit$	$\diamondsuit$
		C (classes)			
$v(f_1)$	y	1	1	3	3
	n	0	4	0	0
	$\sum$	1	5	3	3

$$p(f_1 = y|\diamondsuit) = 1 \qquad p(f_1 = n|\diamondsuit) = 0$$
$$p(f_1 = y|\spadesuit) = \frac{1}{5} \qquad p(f_1 = n|\spadesuit) = \frac{4}{5}$$

Naive Bayes – Example Task

Feature f_2: Color?

$$
\begin{array}{ccccccc}
 & & & \multicolumn{4}{c}{C \text{ (classes)}} \\
 & & & \clubsuit & \spadesuit & \heartsuit & \diamondsuit \\
\hline
 & b & & 0 & 0 & 3 & 3 \\
v(f_2) & r & & 1 & 5 & 0 & 0 \\
\hline
 & \sum & & 1 & 5 & 3 & 3
\end{array}
$$

$$p(f_2 = r|\spadesuit) = 0 \qquad p(f_2 = b|\spadesuit) = 1$$

$$p(f_2 = r|\diamondsuit) = 1 \qquad p(f_2 = b|\diamondsuit) = 0$$

Naive Bayes – Example Task

Feature f_3: Odd/Even/Face?

$$
\begin{array}{cc|cccc}
 & & \multicolumn{4}{c}{C \text{ (classes)}} \\
 & & \clubsuit & \spadesuit & \heartsuit & \diamondsuit \\
\hline
 & o & 1 & 1 & 3 & 2 \\
v(f_3) & e & 0 & 0 & 0 & 1 \\
 & f & 0 & 4 & 0 & 0 \\
\hline
 & \sum & 1 & 5 & 3 & 3
\end{array}
$$

$$
p(f_3 = o|\spadesuit) = \frac{1}{5} \quad p(f_3 = e|\spadesuit) = 0 \quad p(f_3 = f|\spadesuit) = \frac{4}{5}
$$

$$
p(f_3 = o|\diamondsuit) = \frac{2}{3} \quad p(f_3 = e|\diamondsuit) = \tfrac{1}{3} \quad p(f_3 = f|\diamondsuit) = 0
$$

Naive Bayes – Example Task

Prediction

$$\mathit{prediction}(K\spadesuit) \quad = \quad \operatorname*{argmax}_{c \in \{\spadesuit\clubsuit\heartsuit\diamondsuit\}} \; p(c|n,b,f) \quad \text{features extracted from } K\spadesuit$$

Naive Bayes – Example Task

Prediction

$$
\begin{aligned}
prediction(K\spadesuit) &= \underset{c \in \{\spadesuit\clubsuit\heartsuit\diamondsuit\}}{\operatorname{argmax}} \; p(c|n,b,f) \quad \text{features extracted from } K\spadesuit \\
p(\clubsuit|n,b,f) &= p(f_1 = n|\clubsuit) * p(f_2 = b|\clubsuit) * p(f_3 = f|\clubsuit) \\
&= 0
\end{aligned}
$$

Naive Bayes – Example Task

Prediction

$$
\begin{aligned}
prediction(K\spadesuit) &= \underset{c \in \{\spadesuit\clubsuit\heartsuit\diamondsuit\}}{\operatorname{argmax}} \; p(c|n, b, f) \quad \text{features extracted from } K\spadesuit \\
p(\clubsuit|n, b, f) &= p(f_1 = n|\clubsuit) * p(f_2 = b|\clubsuit) * p(f_3 = f|\clubsuit) \\
&= 0 \\
p(\heartsuit|n, b, f) &= p(f_1 = n|\heartsuit) * p(f_2 = b|\heartsuit) * p(f_3 = f|\heartsuit) \\
&= 0
\end{aligned}
$$

Naive Bayes – Example Task

Prediction

$$
\begin{aligned}
\textit{prediction}(K\spadesuit) &= \operatorname*{argmax}_{c \in \{\spadesuit\clubsuit\heartsuit\diamondsuit\}} \; p(c|n,b,f) \quad \text{features extracted from } K\spadesuit \\
p(\clubsuit|n,b,f) &= p(f_1 = n|\clubsuit) * p(f_2 = b|\clubsuit) * p(f_3 = f|\clubsuit) \\
&= 0 \\
p(\heartsuit|n,b,f) &= p(f_1 = n|\heartsuit) * p(f_2 = b|\heartsuit) * p(f_3 = f|\heartsuit) \\
&= 0 \\
p(\spadesuit|n,b,f) &= p(f_1 = n|\spadesuit) * p(f_2 = b|\spadesuit) * p(f_3 = f|\spadesuit) \\
&= \frac{4}{5} * 1 * \frac{4}{5} = 0.64
\end{aligned}
$$

Naive Bayes – Example Task

Prediction

$$
\begin{aligned}
prediction(K\spadesuit) &= \operatorname*{argmax}_{c \in \{\spadesuit\clubsuit\heartsuit\diamondsuit\}} p(c|n,b,f) \quad \text{features extracted from } K\spadesuit \\
p(\clubsuit|n,b,f) &= p(f_1 = n|\clubsuit) * p(f_2 = b|\clubsuit) * p(f_3 = f|\clubsuit) \\
&= 0 \\
p(\heartsuit|n,b,f) &= p(f_1 = n|\heartsuit) * p(f_2 = b|\heartsuit) * p(f_3 = f|\heartsuit) \\
&= 0 \\
p(\spadesuit|n,b,f) &= p(f_1 = n|\spadesuit) * p(f_2 = b|\spadesuit) * p(f_3 = f|\spadesuit) \\
&= \frac{4}{5} * 1 * \frac{4}{5} = 0.64 \\
p(\diamondsuit|n,b,f) &= \ldots = 0
\end{aligned}
$$

Naive Bayes – Example Task

Prediction

$$
\begin{aligned}
prediction(K\spadesuit) &= \underset{c \in \{\spadesuit\clubsuit\heartsuit\diamondsuit\}}{\operatorname{argmax}} \; p(c|n, b, f) \quad \text{features extracted from } K\spadesuit \\
p(\clubsuit|n, b, f) &= p(f_1 = n|\clubsuit) * p(f_2 = b|\clubsuit) * p(f_3 = f|\clubsuit) \\
&= 0 \\
p(\heartsuit|n, b, f) &= p(f_1 = n|\heartsuit) * p(f_2 = b|\heartsuit) * p(f_3 = f|\heartsuit) \\
&= 0 \\
p(\spadesuit|n, b, f) &= p(f_1 = n|\spadesuit) * p(f_2 = b|\spadesuit) * p(f_3 = f|\spadesuit) \\
&= \frac{4}{5} * 1 * \frac{4}{5} = 0.64 \\
p(\diamondsuit|n, b, f) &= \ldots = 0
\end{aligned}
$$

We predict $\spadesuit$

Naive Bayes – Example Task

Prediction

$$
\begin{aligned}
prediction(6\diamondsuit) &= \operatorname*{argmax}_{c\in\{\spadesuit\clubsuit\heartsuit\diamondsuit\}} p(c|y,r,e) \\
p(\clubsuit|y,r,e) &= p(f_1 = y|\clubsuit) * p(f_2 = r|\clubsuit) * p(f_3 = e|\clubsuit) \\
&= 0 \\
p(\heartsuit|y,r,e) &= p(f_1 = y|\heartsuit) * p(f_2 = r|\heartsuit) * p(f_3 = e|\heartsuit) \\
&= 1 * 1 * 0 = 0 \\
p(\diamondsuit|y,r,e) &= p(f_1 = y|\diamondsuit) * p(f_2 = r|\diamondsuit) * p(f_3 = e|\diamondsuit) \\
&= 1 * 1 * \frac{1}{3} = \frac{1}{3}
\end{aligned}
$$

We predict $\diamondsuit$

Naive Bayes – Example Task

Prediction

$$
\begin{aligned}
prediction(K\diamondsuit) &= \underset{c \in \{\spadesuit\clubsuit\heartsuit\diamondsuit\}}{\text{argmax}} \; p(c|n,r,f) \\
p(\clubsuit|n,r,f) &= p(f_1 = n|\clubsuit) * p(f_2 = r|\clubsuit) * p(f_3 = f|\clubsuit) \\
&= 0 \\
p(\heartsuit|n,r,f) &= p(f_1 = n|\heartsuit) * p(f_2 = r|\heartsuit) * p(f_3 = f|\heartsuit) \\
&= 0 \\
p(\diamondsuit|n,r,f) &= p(f_1 = n|\diamondsuit) * p(f_2 = r|\diamondsuit) * p(f_3 = f|\diamondsuit) \\
&= 0
\end{aligned}
$$

Oops, all probabilities are zero

Naive Bayes
Smoothing

- Whenever multiplication is involved, zeros are dangerous
- Smoothing is used to avoid zeros
- Different possibilities
- Simple: Add something to the probabilities
  - $\frac{x_i+a}{N+ad}$
  - E.g., $p(f_3 = e|\spadesuit) = \frac{0+1}{4+1*4}$
  - This leads to values slightly above zero

Naive Bayes

- 'Naive': Assuming feature independence is usually wrong
  - Even in our toy example, f_1 and f_3 are highly dependent
- Pros
  - Easy to implement, fast
  - Small models
- Cons
  - Naive: Feature dependence not modeled
  - Fragile for unseen data (without smoothing)