Recap: Decision Tree, Problem Gambling

- Decision tree
  - Classification method
  - Transparent for humans (for limited number of features)
  - Core idea:
    - Repeatedly split the data set using features, until sub sets are »pure«
    - Split according to information gain of the features
- Problem Gambling
  - Text classification problem
  - Non-linguistic use-case, with criteria grounded in medicinal diagnostics
  - BERT: First »large language model«
  - Pre-training / fine-tuning paradigm

Wahlen zum Europäischen Parlament

- Sonntag, 9. Juni, 9:00-18:00 Uhr
  - 96 Abgeordnete aus Deutschland
  - Listenwahl (d.h. Parteien)
- Relevant für uns, weil (praktisch alle) IT-Themen EU-Themen sind
  - Künstliche Intelligenz
  - Digitale Märkten
  - Chatkontrolle
  - Datenschutzgrundverordnung
  - Und natürlich: Klimakrise

Naive Bayes

Sprachverarbeitung (VL + Ü)

Nils Reiter

June 6, 2024

UNIVERSITÄT ZU KÖLN
INSTITUT FÜR DIGITAL HUMANITIES
UNIVERSITÄT ZU KÖLN

Introduction and Overview

- Second machine learning method (after decision trees)
- Probabilistic method (i.e., probabilities are involved)
- Feature-based method

Basic Probability Theory

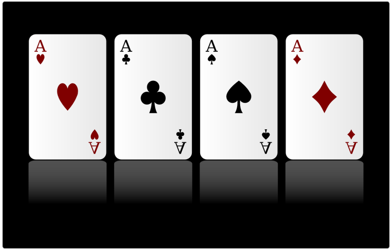
Naive Bayes Algorithm

Example: Spam Classification

Section 1

Basic Probability Theory

Example: 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 $p(E)$: Ratio of size of E to size of Ω (Laplace)
  - $0 \leq p \leq 1$
  - $p(E) = 0$: Impossible event $\qquad p(E) = 1$: Certain event
  - $p(E) = 0.000001$: Very unlikely event

Basics

Probabilities

- Probability $p(E)$: Ratio of size of E to size of Ω (Laplace)
  - $0 \leq p \leq 1$
  - $p(E) = 0$: Impossible event $p(E) = 1$: Certain event
  - $p(E) = 0.000001$: Very unlikely event

Example

- If all outcomes are equally likely: $p(E) = \frac{|E|}{|\Omega|}$
- $p(\{8\heartsuit\}) = \frac{1}{32}$
- $p(\{9\clubsuit, 9\spadesuit, 9\diamondsuit, 9\heartsuit\}) = \frac{4}{32}$
- $p(\Omega) = 1$ (must happen, certain event)

Basics

Probability and Relative Frequency

- Probability p: Theoretical concept, idealization, expectation
- Relative Frequency f: Concrete measure
  - Normalised number of *observed* events

Example

After 10 cards (with returning and shuffling), the event ♠ took place 8 times: $f(\{\spadesuit\}) = \frac{8}{10}$

Basics

Probability and Relative Frequency

- Probability p: Theoretical concept, idealization, expectation
- Relative Frequency f: Concrete measure
  - Normalised number of *observed* events

Example

After 10 cards (with returning and shuffling), the event ♠ took place 8 times: $f(\{♠\}) = \frac{8}{10}$

- For large numbers of drawings, relative frequency approximates the probability
  - $\lim_{\infty} f = p$

Basics

Probability and Relative Frequency

- Probability p: Theoretical concept, idealization, expectation
- Relative Frequency f: Concrete measure
  - Normalised number of *observed* events

Example

After 10 cards (with returning and shuffling), the event ♠ took place 8 times: $f(\{♠\}) = \frac{8}{10}$

- For large numbers of drawings, relative frequency approximates the probability
  - $\lim_{\infty} f = p$
- In practice, we will often use determine probabilities by counting relative frequencies
  - Assumption: Frequency is measured on representative and large data set

Independent Events

Joint Probability

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

Independent Events

Joint Probability

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

Independent Events

Joint Probability

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

Dependent Events

Conditional Probability

- 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 $8\heartsuit$
  - ~~$p(E_1, E_2) = p(E_1) * p(E_2)$~~
  - This no longer works, because the events are not independent
    - Obvious: Only one $8\heartsuit$ in the game, and $p(E_2)$ has to express that it might be gone

Dependent Events

Conditional Probability

- 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 $8\heartsuit$
  - ~~$p(E_1, E_2) = p(E_1) * p(E_2)$~~
  - This no longer works, because the events are not independent
    - Obvious: Only one $8\heartsuit$ in the game, and $p(E_2)$ has to express that it might be gone
  - This is done 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 $p(E) = 0$

Dependent Events

Conditional Probability

A less obvious example:

- We draw two cards in a row
- $E_{\heartsuit}$: Card is $X\heartsuit$
- $E_{\diamondsuit}$: Card is $X\diamondsuit$

Dependent Events

Conditional Probability

A less obvious example:

- We draw two cards in a row
- $E_\heartsuit$: Card is $X\heartsuit$
- $E_\diamondsuit$: Card is $X\diamondsuit$

$$\begin{aligned} p(E_\heartsuit, E_\heartsuit) &= p(E_\heartsuit) * p(E_\heartsuit | E_\heartsuit) \\ &= \end{aligned}$$

Dependent Events

Conditional Probability

A less obvious example:

- We draw two cards in a row
- $E_{\heartsuit}$: Card is $X\heartsuit$
- $E_{\diamondsuit}$: Card is $X\diamondsuit$

$$\begin{aligned} p(E_{\heartsuit}, E_{\heartsuit}) &= p(E_{\heartsuit}) * p(E_{\heartsuit} | E_{\heartsuit}) \\ &= \frac{8}{32} * \end{aligned}$$

Dependent Events

Conditional Probability

A less obvious example:

- We draw two cards in a row
- $E_\heartsuit$: Card is $X\heartsuit$
- $E_\diamondsuit$: Card is $X\diamondsuit$

$$\begin{aligned} p(E_\heartsuit, E_\heartsuit) &= p(E_\heartsuit) * p(E_\heartsuit | E_\heartsuit) \\ &= \frac{8}{32} * \frac{7}{31} = 0.056 \end{aligned}$$

Dependent Events

Conditional Probability

A less obvious example:

- We draw two cards in a row
- $E_{\heartsuit}$: Card is $X\heartsuit$
- $E_{\diamondsuit}$: Card is $X\diamondsuit$

$$
\begin{aligned}
p(E_{\heartsuit}, E_{\heartsuit}) &= p(E_{\heartsuit}) * p(E_{\heartsuit}|E_{\heartsuit}) \\
&= \frac{8}{32} * \frac{7}{31} = 0.056 \\
p(E_{\diamondsuit}, E_{\heartsuit}) &= p(E_{\diamondsuit}) * p(E_{\heartsuit}|E_{\diamondsuit}) \\
&=
\end{aligned}
$$

Dependent Events

Conditional Probability

A less obvious example:

- We draw two cards in a row
- $E_\heartsuit$: Card is $X\heartsuit$
- $E_\diamondsuit$: Card is $X\diamondsuit$

$$
\begin{aligned}
p(E_\heartsuit, E_\heartsuit) &= p(E_\heartsuit) * p(E_\heartsuit | E_\heartsuit) \\
&= \frac{8}{32} * \frac{7}{31} = 0.056 \\
p(E_\diamondsuit, E_\heartsuit) &= p(E_\diamondsuit) * p(E_\heartsuit | E_\diamondsuit) \\
&= \frac{8}{32} * \frac{8}{31} = 0.064
\end{aligned}
$$

Conditional and Joint Probabilities

Another Example

- Setup: We make a survey in a street in Cologne
- We count four types of events in two random variables:
  - Person has brown hair ($H = B$)
  - Person has red hair ($H = R$)
  - Person likes to wake up late ($W = L$)
  - Person likes to wake up early ($W = E$)

Conditional and Joint Probabilities

Another Example

- Setup: We make a survey in a street in Cologne
- We count four types of events in two random variables:
  - Person has brown hair ($H = B$)
  - Person has red hair ($H = R$)
  - Person likes to wake up late ($W = L$)
  - Person likes to wake up early ($W = E$)
- Assumption: B / R and L / E are mutually exclusive
  - I.e., a single person cannot have red *and* brown hair
- A single person can be encoded with two symbols (e.g., »BL«)
  - ⚠ But this combination is not unique – in contrast to the cards example
- All following numbers are made up

Conditional and Joint Probabilities

Example

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

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

Table: Survey Results, Ω: Group of questioned people

Conditional and Joint Probabilities

Example

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

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

Table: Survey Results, Ω: Group of questioned people

If we pick a random person, what's the probability that this person has brown hair?

$$p(H = \text{brown}) =$$

Conditional and Joint Probabilities

Example

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

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

Table: Survey Results, Ω: Group of questioned people

If we pick a random person, what's the probability that this person has brown hair?

$$p(H = \text{brown}) = \frac{50}{65}$$

Conditional and Joint Probabilities

Example

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

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

Table: Survey Results, Ω: Group of questioned people

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

Conditional and Joint Probabilities

Example

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

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

Table: Survey Results, Ω: Group of questioned people

- 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 and Joint Probabilities

Example

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

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

Table: Survey Results, Ω: Group of questioned people

- 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)

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 $|\Omega| = 65$

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 $|\Omega| = 65$

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

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 $|\Omega| = 65$

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

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

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 $|\Omega| = 65$

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

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

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

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 $|\Omega| = 65$

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

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

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

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

Section 2

Naive Bayes Algorithm

Naive Bayes

- Probabilistic model (i.e., takes probabilities into account)
- Probabilities are estimated on training data (relative frequencies)
- Reading — Jurafsky/Martin (2023, Chapter 4)

Two Parts

- Prediction model: How does the model make predictions on new instances?
- Learning algorithm: How is the model created based on annotated data?

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

- $f_n(x)$: 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
- $\text{argmax}_i\, e$: 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
- $\text{argmax}_i\, e$: Select the argument i that maximizes the expression e

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

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
- $\text{argmax}_i\, e$: Select the argument i that maximizes the expression e

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

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

Naive Bayes

Prediction Model

Definition of conditional probabilities

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

Naive Bayes

Prediction Model

Definition of conditional probabilities

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

Naive Bayes

Prediction Model

Definition of conditional probabilities

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

Naive Bayes

Prediction Model

Definition of conditional probabilities

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

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

Definition of conditional probabilities

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

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

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

Naive Bayes

Prediction Model

From previous slide

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

Naive Bayes

Prediction Model

From previous slide

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

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.
It's not a mistake to apply the denominator, but it's also not necessary.

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*

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

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.
It's not a mistake to apply the denominator, but it's also not necessary.

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		

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‹
  - »Bag of Words« representation

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

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

...

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

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*

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

Prediction

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

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

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

Prediction

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

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

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

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

3. Assign the class with the higher probability

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	C false
love	1	0	35
love	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

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
- Theoretical justification: Some of the probability space is left unused, for events (= words) that we haven't seen yet

References I

Jurafsky, Dan/James H. Martin (2023). *Speech and Language Processing*. 3rd ed. Draft of Janaury 7, 2023. Prentice Hall. URL: https://web.stanford.edu/~jurafsky/slp3/.