UNIVERSITÄT
ZU KÖLN

Machine Learning 1: Naive Bayes

VL Sprachliche Informationsverarbeitung

Nils Reiter
nils.reiter@uni-koeln.de

November 21, 2024
Winter term 2024/25

INSTITUT FÜR
DIGITAL HUMANITIES
UNIVERSITÄT ZU KÖLN

Hausaufgabe 2

- Reden von Politiker:innen herunterladen
- Type-Token-Ratio berechnen
- Was kam raus?

Hausaufgabe 2

- Reden von Politiker:innen herunterladen
- Type-Token-Ratio berechnen
- Was kam raus?

Meine Kommentare zu den Ergebnissen

- Absolute Pfade in Programmcode ☹
- Setzen Sie keine Screenshots von Programmcode in Ihre Dokumente ☹
- Je länger der Text, desto geringer die TTR – deswegen besser STTR verwenden
- TTR *kann* eine interessante Unterschiede zeigen, aber meistens in Kombination mit anderen Indikatoren
- Lexikalische Varianz interagiert mit den Inhalten

SHK-Stelle am Bundesinstitut für Berufsbildung (BIBB)

Informationsextraktion aus Stellenanzeigen

- Mehrere Millionen Stellenanzeigen sollen mit Informationen zu Beruf, Tätigkeitsprofil und Kompetenzen angereichert werden
- Modellentwicklung mithilfe von LLMs auf hauseigener Serverinfrastruktur

- ab Frühjahr 2025
- Umfang: 19 Stunden/Woche
- Gehaltseinstufung: TVÖD-Bund E6
- Befristung: 2 Jahre
- weitere Informationen: **kai.krueger@bibb.de**
  Tel.: **+49 (0) 228 – 107 1580**

Introduction

- Probabilistic classification algorithm
- Makes independence assumption about features – 'naive'
- Reading Jurafsky/Martin (2023, Ch. 4)

Introduction

- Probabilistic classification algorithm
- Makes independence assumption about features – ‘naive’
- Reading Jurafsky/Martin (2023, Ch. 4)
- Nice intro to Bayesian statistics by Matt Parker and Hannah Fry Parker/Fry (2019)

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
    - There are $2^{|\Omega|}$ different subsets, i.e., $2^{|\Omega|}$ possible events
- 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
  - There are $2^{|\Omega|}$ different subsets, i.e., $2^{|\Omega|}$ possible events
- 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
  - There are $2^{|\Omega|}$ different subsets, i.e., $2^{|\Omega|}$ possible events
- 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
  - There are $2^{|\Omega|}$ different subsets, i.e., $2^{|\Omega|}$ possible events
- 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
  - There are $2^{|\Omega|}$ different subsets, i.e., $2^{|\Omega|}$ possible events
- 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
  - There are $2^{|\Omega|}$ different subsets, i.e., $2^{|\Omega|}$ possible events
- 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
  - There are $2^{|\Omega|}$ different subsets, i.e., $2^{|\Omega|}$ possible events
- 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 ten"

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
  - There are $2^{|\Omega|}$ different subsets, i.e., $2^{|\Omega|}$ possible events
- 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 ten" – $E = \{8\heartsuit, 10\diamondsuit\}$
- "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
  - There are $2^{|\Omega|}$ different subsets, i.e., $2^{|\Omega|}$ possible events
- 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 ten" – $E = \{8\heartsuit, 10\diamondsuit\}$
- "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 $p(E) = 1$: 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 $p(E) = 1$: 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, 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

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
  - $\lim_{\infty} f = p$
- In practice, we will often use relative frequencies as probabilities
- This establishes assumptions:
  - Data set is representative of the real world
  - We make a lot of observations (the more, the better we approximate real probabilities)

Basics

Joint Probability (Independent Events)

- 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 $8\heartsuit$
  - $p(E) = p(E_1, E_2) = p(E_1) * p(E_2) = \frac{1}{32} * \frac{1}{32} = 0.0156$

Basics

Joint Probability (Independent Events)

- 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 $X\heartsuit$
  - $p(E) = p(E_1, E_2) = p(E_1) * p(E_2) = \frac{1}{4} * \frac{1}{4} = 0.0625$

Basics

Joint Probability (Independent Events)

- 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

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 $8\heartsuit$ (without putting the first card back)
  - $\cancel{p(E_1, E_2) = p(E_1) * p(E_2)}$
  - 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 $p(E_1, E_2) = 0$

Basics II

Conditional Probability (Dependent Events)

- E: We draw $\heartsuit$ first (E_1), followed by:
  - E_2: Second card is $X\diamondsuit$
  - E_3: Second card is $X\heartsuit$
  - $p(E_1, E_2) = p(E_1) * p(E_2|E_1) = \frac{8}{32} * \frac{8}{31} = 0.064$
  - $p(E_1, E_3) = p(E_1) * p(E_3|E_1) = \frac{8}{32} * \frac{7}{31} = 0.056$

Conditional and Joint Probabilities

Example

Relation between **hair color** H and preferred **wake-up time** W (all numbers are made up.)

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

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

Conditional and Joint Probabilities

Example

Relation between **hair color** H and preferred **wake-up time** W (all numbers are made up.)

$\downarrow W$ / $H \rightarrow$	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?

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

Conditional and Joint Probabilities

Example

Relation between **hair color** H and preferred **wake-up time** W (all numbers are made up.)

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

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

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

Conditional and Joint Probabilities

Example

Relation between **hair color** H and preferred **wake-up time** W (all numbers are made up.)

$\downarrow W$ / $H \rightarrow$	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 and Joint Probabilities

Example

Relation between **hair color** H and preferred **wake-up time** W (all numbers are made up.)

$\downarrow W$ / $H \rightarrow$	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)

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 $|\Omega|$

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

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 $|\Omega|$

$$\begin{aligned}
p(A|B) &= \frac{p(A,B)}{p(B)} && \text{definition of conditional probabilities}\\
p(W=\text{late}|H=\text{brown}) &= \frac{30}{50}=0.6 && \text{intuition from previous slide}\\
&= \frac{p(W=\text{late},H=\text{brown})}{p(H=\text{brown})} && \text{by applying definition}\\
&= \frac{0.46}{0.77}=0.6
\end{aligned}$$

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

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

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

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

- 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

- Probabilistic model (i.e., takes probabilities into account)
- Probabilities are estimated on training data (relative frequencies)
- Setup
  - A set of features f_i
  - A data set $x \in X$ (x is an individual instance, X the entire set)
  - The feature *value* is given as $f_i(x)$

Naive Bayes

Prediction Model

- Probabilistic model (i.e., takes probabilities into account)
- Probabilities are estimated on training data (relative frequencies)
- Setup
  - A set of features f_i
  - A data set $x \in X$ (x is an individual instance, X the entire set)
  - The feature *value* is given as $f_i(x)$

Example

- Feature representing "word length" f_6
- One data point is "dog"
- $f_6(``dog") = 3$

Naive Bayes

Prediction Model

- Probabilistic model (i.e., takes probabilities into account)
- Probabilities are estimated on training data (relative frequencies)
- Setup
  - A set of features f_i
  - A data set $x \in X$ (x is an individual instance, X the entire set)
  - The feature *value* is given as $f_i(x)$

Example

- Feature representing "word length" f_6
- One data point is "dog"
- $f_6(\text{``}dog\text{''}) = 3$

You can also think of f_6 as a function in a program:

```
def f6(x):
    return len(x)
```

Naive Bayes

Prediction Model

Intuition

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

Intuition

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
- $\arg\max_i e$: Select the argument i that maximizes the expression e

Naive Bayes

Prediction Model

```
def argmax(SET, EXPRESSION):
  arg = 0
  maxvalue = 0
  foreach i in SET:
    value = EXPRESSION(i)
    if value > maxvalue:
      arg = i
      maxvalue = value
  return arg
```

Intuition

We calculate the probability for each possible class c, given the
and we assign most probably class

- $f_n(x)$: Value of feature n for instance x
- $\arg\max_i e$: Select the argument i that maximizes the expression e

Naive Bayes

Prediction Model

```
def argmax(SET, EXPRESSION):
  arg = 0
  maxvalue = 0
  foreach i in SET:
    value = EXPRESSION(i)
    if value > maxvalue:
      arg = i
      maxvalue = value
  return arg
```

Intuition

We calculate the probability for each possible class c, given the and we assign most probably class

- $f_n(x)$: Value of feature n for instance x
- $\arg\max_i e$: Select the argument i that maximizes the expression e

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

Naive Bayes

Prediction Model

Intuition

We calculate the probability for each possible class c, given the … and we assign most probably class

```
def argmax(SET, EXPRESSION):
  arg = 0
  maxvalue = 0
  foreach i in SET:
    value = EXPRESSION(i)
    if value > maxvalue:
      arg = i
      maxvalue = value
  return arg
```

- $f_n(x)$: Value of feature n for instance x
- $\arg\max_i e$: Select the argument i that maximizes the expression e

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

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

Naive Bayes

Prediction Model

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

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

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

Naive Bayes

Prediction Model

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

Naive Bayes

Prediction Model

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

Naive Bayes

Prediction Model

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

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

Naive Bayes

Learning Algorithm

1. For each feature f_i
   - 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'

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

		C	
		true	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*

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

Prediction

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

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

Prediction

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

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

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	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 (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

Summary

- Probability theory
  - Probability: Fraction of positive over all possible events
  - Conditional probability: Restrict the space of possible events
- Naive Bayes
  - Probability-based classification algorithm
  - Assumes feature independence (therefore: “naive”)
  - Still used in many applications
    - E.g., spam classification

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/.

Parker, Matt/Hannah Fry (2019). *Bayesian Statistics with Hannah Fry*. URL: https://www.youtube.com/watch?v=7GgLSnQ48os.