Machine Learning 2: Naive Bayes

VL Sprachverarbeitung

Nils Reiter

May 12, 2022

Recap

- Classification: Sort instances into previously known groups
  - Major machine learning task type
- Evaluation
  - Baseline: Whatever we compare against (our decision, but needs to be convincing)
  - Intrinsic: Compare against a gold standard
    - Accuracy, precision, recall, f-score
  - Extrinsic: Evaluate downstream application

Today

1. More on classification
2. Our first classification/machine learning algorithm: Naive Bayes
   - ...and we need to talk about probabilities, again

Section 1

Machine Learning Basics

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?

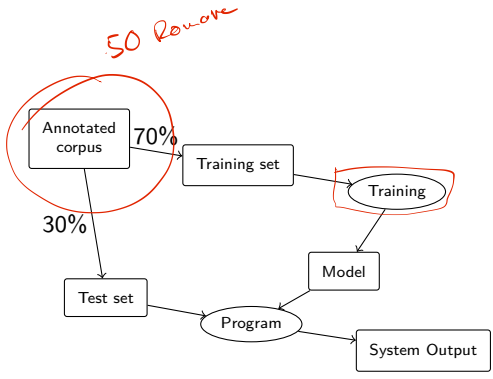
Two Parts

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?



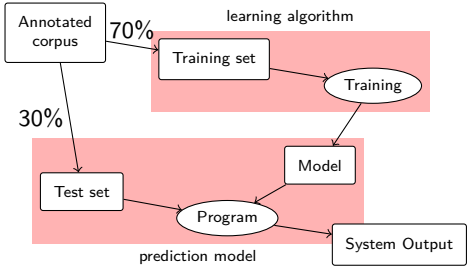
Two Parts

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?



Machine Learning

Classification

- Assigning *classes* to *objects/instances/items*
  - Words → parts of speech
  - Texts → genres
- Many algorithms available: Decision trees, support vector machines, naïve Bayes, neural networks, Bayesian networks, ...

logistic regression

Machine Learning

Classification

- Assigning *classes* to *objects/instances/items*
  - Words → parts of speech
  - Texts → genres
- Many algorithms available: Decision trees, support vector machines, naïve Bayes, neural networks, Bayesian networks, ...
- Libraries are available, not a technical challenge
- Challenges
  - Find helpful training data
  - Use an appropriate algorithm
  - Use it correctly
  - Interpret its results reasonably
  - Establish a realistic test set

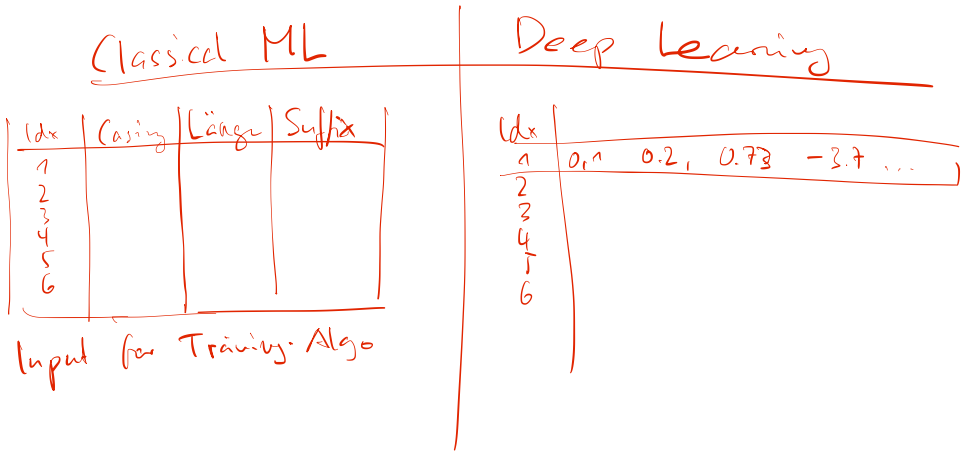
Machine Learning

Features (a.k.a. attributes, properties)

- Decision is based on features (= properties)
- The prediction model only sees feature values!
  - What's not encoded in a feature does not play a role
  - It's our job to provide useful features
- Playground for being creative
  - It helps to understand the problem/task/phenomenon

Deep Learning

- Most recent trend in NLP (recent: since ca. 2015)



Deep Learning

- Most recent trend in NLP (recent: since ca. 2015)
- Two components
  - Distributed input representation
    - Word embeddings: Each word type is represented as vector, derived from the words co-occurrences
    - ›Distributional semantics‹
    - No more manual feature engineering

Deep Learning

- Most recent trend in NLP (recent: since ca. 2015)
- Two components
  - Distributed input representation
    - Word embeddings: Each word type is represented as vector, derived from the words co-occurrences
    - ›Distributional semantics‹
    - No more manual feature engineering
  - Artificial neural networks
    - Large number of connected neurons that perform mathematical operations

Deep Learning

- Most recent trend in NLP (recent: since ca. 2015)
- Two components
  - Distributed input representation
    - Word embeddings: Each word type is represented as vector, derived from the words co-occurrences
    - ›Distributional semantics‹
    - No more manual feature engineering
  - Artificial neural networks
    - Large number of connected neurons that perform mathematical operations
- Manually crafted features provide generalization
  - Neural networks need to learn this generalization in other ways

  $\Rightarrow$ Typically, we need more training data and compute power

Deep Learning

Impact

Consumption	CO_2e (lbs)
Air travel, 1 person, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experiments	78,468
Transformer (big)	192
w/ neural arch. search	626,155

Figure: Estimated CO_2 consumption from training common deep learning models (Strubell et al., 2019)

Emma Strubell/Ananya Ganesh/Andrew McCallum (2019). »Energy and Policy Considerations for Deep Learning in NLP«. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 3645–3650

Section 2

Naive Bayes

Introduction

- Probabilistic classification algorithm
- Makes independence assumption about features – ›naive‹
- Reading JM20, 56 ff.

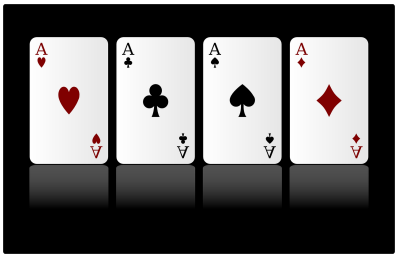
Introduction

- Probabilistic classification algorithm
- Makes independence assumption about features – ›naive‹
- Reading JM20, 56 ff.
- Nice intro to Bayesian statistics by Matt Parker and Hannah Fry
  `https://www.youtube.com/watch?v=7GgLSnQ48os`

Subsection 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 $p(E) = 1$: 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 $\quad$ $p(E) = 1$: 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: $f(\{x\spadesuit\}) = \frac{8}{10}$
- 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: $f(\{x\spadesuit\}) = \frac{8}{10}$
- For large numbers of drawings, relative frequency approximates the probability
  - $\lim_{\infty} f = p$
- In practice, we will often use relative frequency as probability
- This assumes that the data set on which we measure relative frequency is representative etc.

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 $X\heartsuit$
  - $p(E) = p(E_1, E_2) = p(E_1) * p(E_2) = \frac{1}{4} * \frac{1}{4} = 0.0625$
- So far, events have been 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 $8\heartsuit$
  - ~~$p(E_1, E_2) = p(E_1) * p(E_2)$~~
  - 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 ♡ two times in a row
  - E_1: First card is $X\heartsuit$
  - E_2: First card is $X\diamondsuit$
  - E_3: Second card is $X\heartsuit$
  - $p(E_1, E_3) = p(E_1) * p(E_3|E_1) = \frac{8}{32} * \frac{7}{31} = 0.056$
  - $p(E_2, E_3) = p(E_2) * p(E_3|E_2) = \frac{8}{32} * \frac{8}{31} = 0.064$

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

[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

$P(W=\text{late} \mid H=\text{brown})$

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

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

$p(A, B, C, D)$
$= p(B, A, D, C)$

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

Subsection 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

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

Naive Bayes

Prediction Model

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
- $\text{argmax}_i\, e$: 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

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
- $\text{argmax}_i\, e$: 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

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

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

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

denominator is constant, so we skip it

$$\propto p(f_1|f_2, \ldots, f_n, c) \times p(f_2|f_3, \ldots, f_n, c) \times \cdots \times p(c)$$

Naive Bayes

Prediction Model

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

denominator is constant, so we skip it

$$\propto p(f_1|f_2, \ldots, f_n, c) \times p(f_2|f_3, \ldots, f_n, c) \times \cdots \times p(c)$$

Now we – naively – assume feature independence

$$= p(f_1|c) \times p(f_2|t) \times \cdots \times p(c)$$

Naive Bayes

Prediction Model

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

Naive Bayes

Prediction Model

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

denominator is constant, so we skip it

$$\propto p(f_1|f_2, \ldots, f_n, c) \times p(f_2|f_3, \ldots, f_n, c) \times \cdots \times p(c)$$

Now we – naively – assume feature independence

$$= p(f_1|c) \times p(f_2|t) \times \cdots \times p(c)$$

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

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:

$v(f_i)$	C (classes) c_1	c_2	...	c_m
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., $p(a|c_1) = \frac{3}{3+5+0}$ $\quad$ $p(b|c_2) = \frac{7}{2+7+1}$

Subsection 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*

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

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

Prediction

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

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

$$= \cdots \times \frac{5}{50} \times \cdots \times \frac{15}{50} \times \cdots = \ldots \quad 0.2$$

$$p\left(\text{false} \,\middle|\, \begin{bmatrix} \text{casino} & 0 \\ \vdots & \vdots \end{bmatrix}\right) \propto \ldots \quad p(\text{casino} = 0|\text{false}) < \quad 0.3$$

3. Assign the class with the higher probability

demo

Subsection 4

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)!

Section 3

Summary

Summary

- Probabilities
  - Joint and conditional probabilities
  - Probabilities and relative frequencies
- ›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)