Universität Stuttgart
KI – Institute for Artificial Intelligence
Analytic Computing

Machine Learning
6 Bayesin Classification

Prof. Dr. Steffen Staab
Nadeen Fatallah
Osama Mohamed
Daniel Frank
Yi Wang
Akram Sadat Hosseini
Tim Schneider
Rodrigo Lopez

https://www.ki.uni-stuttgart.de/

Intermezzo:

Discrete Probabilistic Reasoning

Inference

„Inferences are steps in reasoning, moving from premises to logical consequences...“

(https://en.wikipedia.org/w/index.php?title=Inference&oldid=953649900)

„Statistical inference uses mathematics to draw conclusions in the presence of **uncertainty**.“

(https://en.wikipedia.org/w/index.php?title=Inference&oldid=953649900)

Sources of uncertainty

- Machine learning:
  - generalizes from (rather) small number of observations (i.e. samples)
  - to the general (including unseen objects)

- Observations are uncertain, because
  - observations may be flawed (e.g. defect or inaccurate sensor)
  - reality may be uncertain
    - inherently uncertain (e.g. quantum dynamics)
    - incomplete knowledge best modelled by uncertainty (no Laplace‘s demon)
      - including: modeling with latent variables

Probabilities and Random Variables

- For a random variable X with discrete $dom(X) = \Omega$ we write: $\forall x \in \Omega : 0 \leq P(X = x) \leq 1$

$$\sum_{x \in \Omega} P(X = x) = 1$$

- Example: A dice can take values $\Omega = \{1, 2, 3, 4, 5, 6\}$
  - X is the random variable representing a dice throw.
  - $P(X = 1) \in [0, 1]$ is the probability that X takes value 1, i.e. event „1“ happens
- A random variable is a map from a measurable space to a domain (sample space). It introduces a probability measure on the domain („assigns a probability to each possible value“ or “assigns a probability to each possible, maybe complex event“)

Atomic and non-atomic events

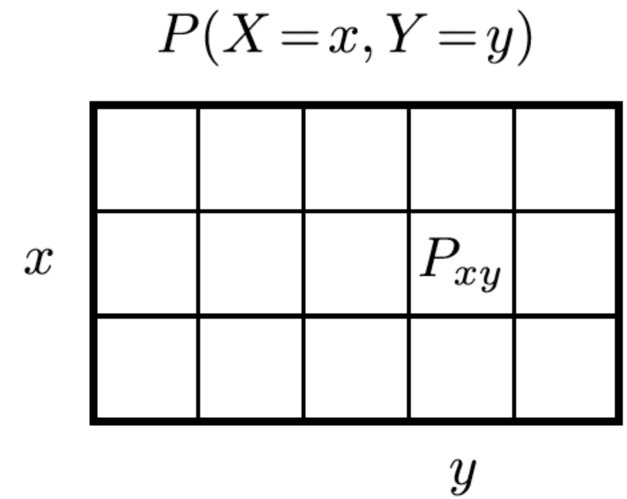
- Atomic or elementary event of X with $\Omega = \{1, 2, 3, 4, 5, 6\}$
  - Represented by a singleton, e.g. {3}
- Non-atomic event
  - E.g. seeing an even number {2,4,6}

Notation

- $P(X = 1) \in [0, 1]$ denotes a specific probability of an event
- $P(X)$ denotes the probability distribution (function over Ω)
  - For the dice example, $P(X)$ describes the distribution over the 6 possible atomic events
- Implementation over discrete random variables as array: $\left[\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right]$
- Non-atomic events $A = \{x_1, \ldots, x_k\}$ have probability $\Sigma_{x \in A} P(X = x)$ or shorthand $\Sigma_A P(X)$
- We also write $\Sigma_{x \in dom(X)} P(X = x) = \Sigma_X P(X)$

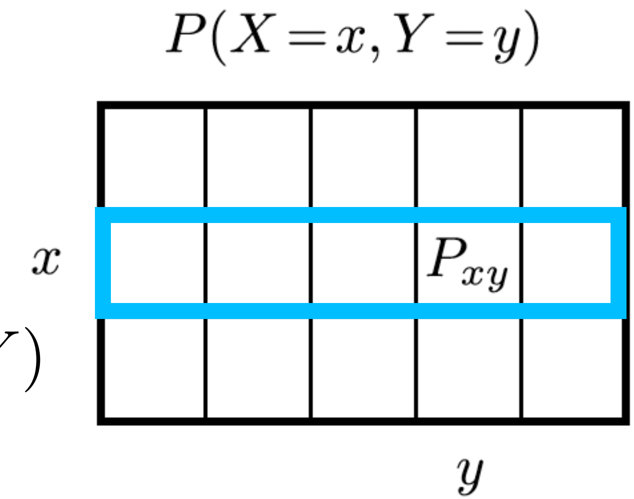
Joint distributions

- Assume we have two random variables X, Y
- Implemented as a matrix of probability values $P_{x,y}$



Joint distributions of two discrete random variable

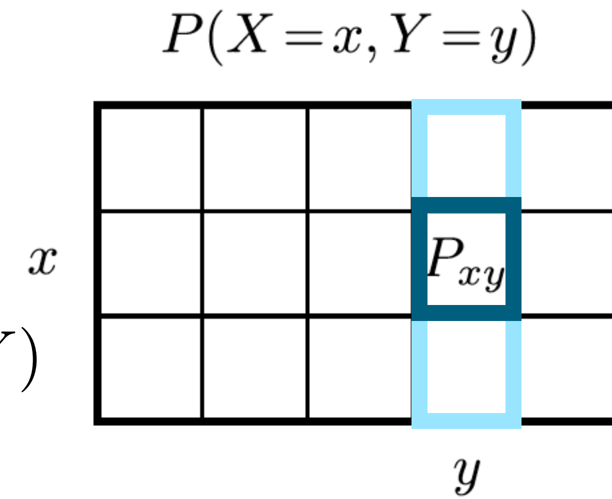
- Assume we have two random variables X, Y



- Definitions:
  - Joint distribution: $P(X, Y)$
  - Marginal distribution: $P(X) = \Sigma_{\boxed{Y}} P(X, Y)$

Joint distributions

- Assume we have two random variables X, Y

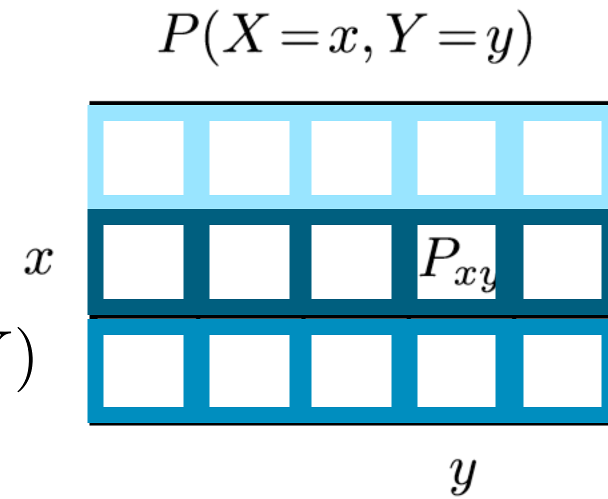


- Definitions:
  - Joint distribution: $P(X, Y)$
  - Marginal distribution: $P(X) = \Sigma_Y P(X, Y)$
  - Conditional: $P(X|Y) = \frac{\boxed{P(X,Y)}}{\boxed{P(Y)}}$

  - The conditional is normalized: $\forall y \in \mathrm{dom}(Y)\colon \sum_X P(X|Y = y) = 1$

Joint distributions

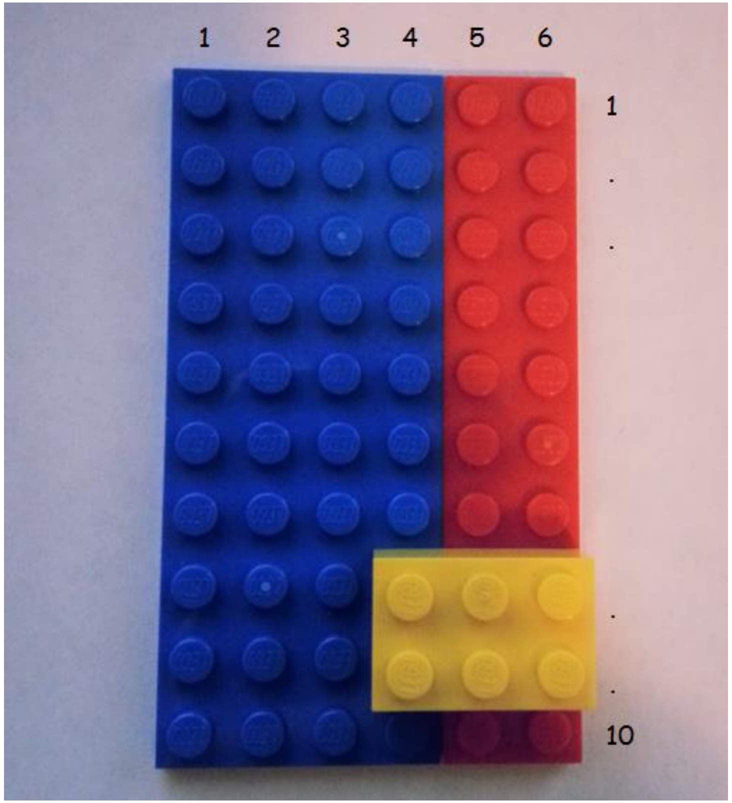
- Assume we have two random variables X, Y



- Definitions:
  - Joint distribution: $P(X, Y)$
  - Marginal distribution: $P(X) = \Sigma_Y P(X, Y)$
  - Conditional: $P(X|Y) = \frac{P(X,Y)}{P(Y)}$
  - The conditional is normalized: $\forall Y : \Sigma_X P(X|Y) = 1$
  - X is independent of Y iff: $P(X|Y) = P(X)$

identical colors stand for identical values

Visualizing specific conditional events



P (yellow | red) = ?

P (red| yellow) = ?

Can we infer something about P (red| yellow) if we know P (yellow | red) ?

Reason for Bayes' Theorem

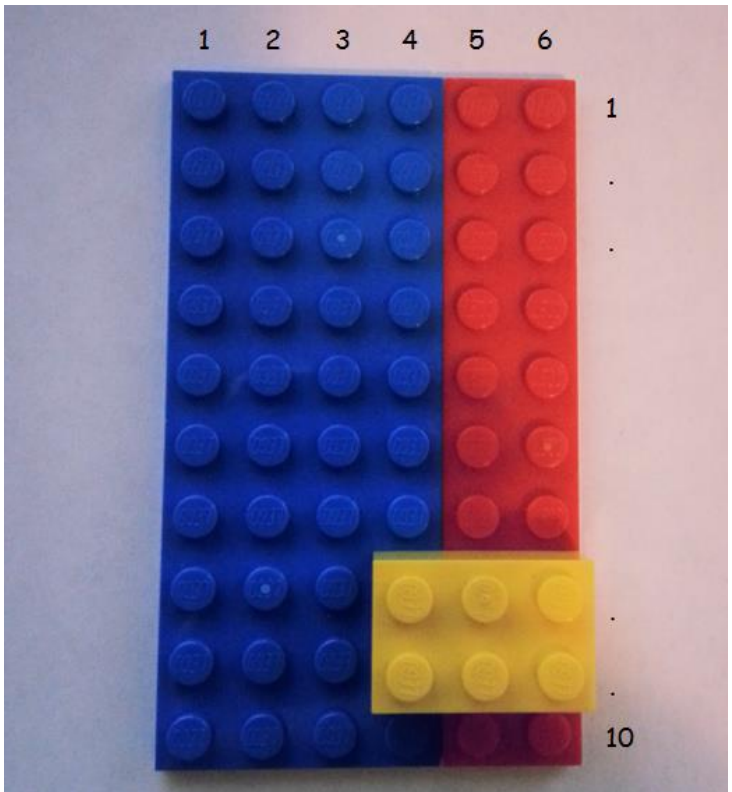
$$P(A|B) = \frac{P(A,B)}{P(B)}$$
$$P(B|A) = \frac{P(A,B)}{P(A)}$$

$$P(A,B) = P(A|B) * P(B)$$
$$P(A,B) = P(B|A) * P(A)$$

$$P(A|B) * P(B) = P(B|A) * P(A)$$

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Can I infer something about the second, by knowing about the first?



P (yellow | red) = 4/20 = 1/5

P (red| yellow) = 4/6 = 2/3

$$P(\text{red}|\text{yellow}) = \frac{P(\text{yellow}|\text{red}) \cdot P(\text{red})}{P(\text{yellow})} =$$

This is practically useful
if the term on the left side is hard to measure,
but the terms on the right side are easy to measure.

Bayes' Theorem

TALKING ABOUT DISTRIBUTIONS

$$P(X|Y) = \frac{P(Y|X) \cdot P(X)}{P(Y)}$$

TALKING ABOUT PROBABILITIES OF (OFTEN NON-ATOMIC) EVENTS

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

$$\text{posterior} = \frac{\text{likelihood}\cdot\text{prior}}{\text{normalization}}$$

Probability theory for dealing with uncertainty

Frequentist probabilities are defined in the limit of an infinite number of trials

- Example: “The probability of a particular coin landing heads up is 0.43”

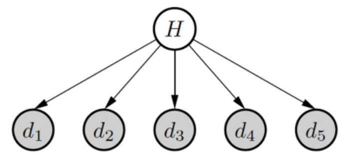
Bayesian (subjective) probabilities quantify degrees of belief

- Example: “The probability of rain tomorrow is 0.3”
  – not possible to repeat “tomorrow”

Fair or manipulated?

- Coin flipping: What process produces these sequences?

HHTHT



HHHHH

- Two hypotheses H=1, H=2
  - Fair coin: $H = 1 \qquad P(d_i = Head \mid H = 1) = \frac{1}{2}$
  - Manipulated: $H = 2 \qquad P(d_i = Head \mid H = 2) = 1$

H1 fair, H2 manipulated

D=HHTHT

$$P(D|H=1) = \left(\frac{1}{2}\right)^5$$

$$P(H=1|D) = \frac{P(D|H=1) \cdot P(H=1)}{P(D)}$$

D=HHTHT

$P(D|H=1) = \left(\frac{1}{2}\right)^5$

$P(H=1|D) = \frac{P(D|H=1)\cdot P(H=1)}{P(D)}$

$P(D|H=2) = 0$

$P(H=2|D) = \frac{P(D|H=2)\cdot P(H=2)}{P(D)}$

We do not know the probability of $P(D)$ but it is the same for either observation, hence we compare:

$$\frac{P(H=1|D)}{P(H=2|D)} = \frac{P(D|H=1)\cdot P(H=1)}{P(D|H=2)\cdot P(H=2)}$$

What are our prior beliefs that the coin is fair or manipulated? Let‘s assume:

$$P(H=1) = 0.999, P(H=2) = 0.001$$

Then

$$\frac{P(H=1|D)}{P(H=2|D)} = \frac{\left(\frac{1}{2}\right)^5}{0} \cdot \frac{0.999}{0.001} = \infty$$

D=HHHHH

$$P(D|H=1) = \left(\frac{1}{2}\right)^5$$

$$P(H=1|D) = \frac{P(D|H=1) \cdot P(H=1)}{P(D)}$$

$$P(D|H=2) = 1$$

$$P(H=2|D) = \frac{P(D|H=2) \cdot P(H=2)}{P(D)}$$

D=HHHHH

$P(D|H=1) = \left(\frac{1}{2}\right)^5$

$P(D|H=2) = 1$

$P(H=1|D) = \frac{P(D|H=1) \cdot P(H=1)}{P(D)}$

$P(H=2|D) = \frac{P(D|H=2) \cdot P(H=2)}{P(D)}$

We do not know the probability of $P(D)$ but it is the same for either observation, hence we compare:

$$\frac{P(H=1|D)}{P(H=2|D)} = \frac{P(D|H=1) \cdot P(H=1)}{P(D|H=2) \cdot P(H=2)}$$

What are our prior beliefs that the coin is fair or manipulated? Let‘s assume:

$P(H=1) = 0.999, P(H=2) = 0.001$

Then

$$\frac{P(H=1|D)}{P(H=2|D)} = \frac{\left(\frac{1}{2}\right)^5}{1} \cdot \frac{0.999}{0.001} \approx 30$$

H1 fair, H2 manipulated

D=HHHHHHHHHH

$$P(D|H=1) = \left(\tfrac{1}{2}\right)^1 0 \qquad P(H=1|D) = \frac{P(D|H=1) \cdot P(H=1)}{P(D)}$$

$$P(D|H=2) = 1 \qquad P(H=2|D) = \frac{P(D|H=2) \cdot P(H=2)}{P(D)}$$

We do not know the probability of $P(D)$ but it is the same for either observation, hence we compare:

$$\frac{P(H=1|D)}{P(H=2|D)} = \frac{P(D|H=1) \cdot P(H=1)}{P(D|H=2) \cdot P(H=2)}$$

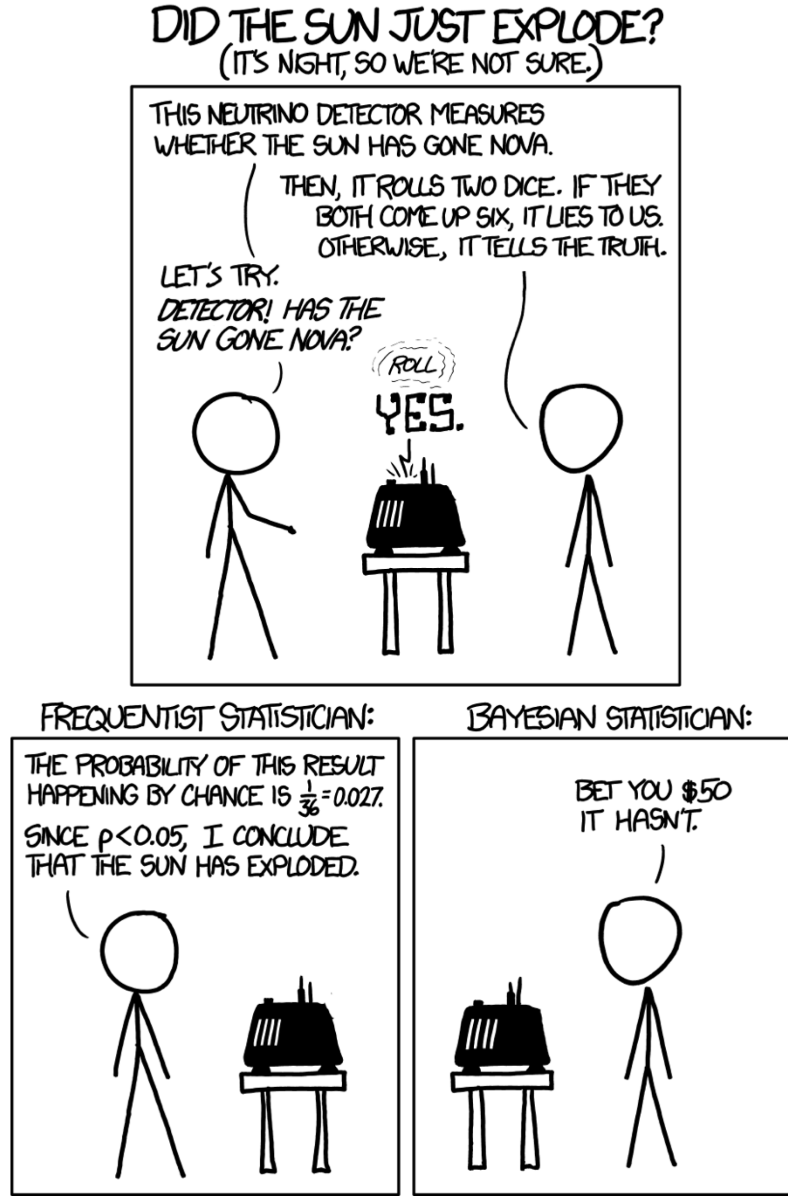
What are our prior beliefs that the coin is fair or manipulated? Let‘s assume:

$$P(H=1) = 0.999, P(H=2) = 0.001$$

Then

$$\frac{P(H=1|D)}{P(H=2|D)} = \frac{\frac{1}{1024}}{1} \cdot \frac{0.999}{0.001} \approx 1$$

https://xkcd.com/1132/



Joint Distributions for Three Random Variables

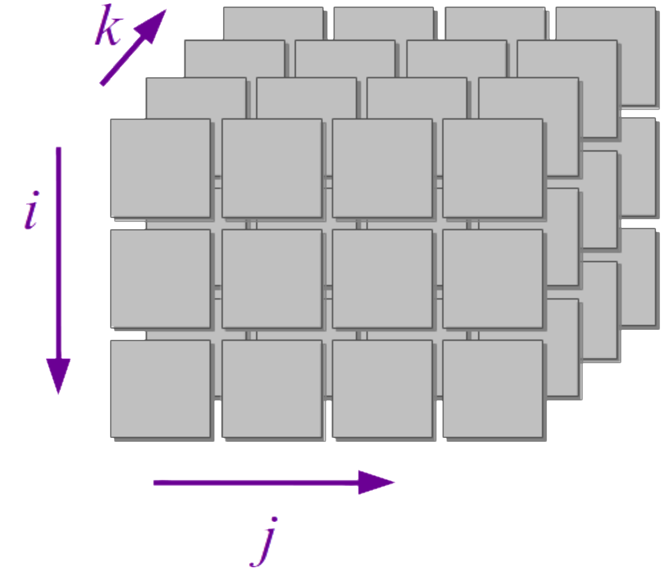
3 random variables X, Y, Z, stored as rank 3 tensor

Joint distribution: $P(X, Y, Z)$

Marginal distribution: $P(X) = \Sigma_{Y,Z} P(X, Y, Z)$

Conditional distribution: $P(X|Y, Z) = \frac{P(X,Y,Z)}{P(Y,Z)}$

X is conditionally independent of Y given Z iff: $P(X|Y, Z) = P(X|Z)$



https://www.programming-idioms.org/idiom/27/create-a-3-dimensional-array/452/rust

Product rule and Bayes‘ theorem:

$$P(X, Z, Y) = P(X|Y, Z)\ P(Y|Z)\ P(Z)$$

$$P(X|Y, Z) = \frac{P(Y|X,Z)\ P(X|Z)}{P(Y|Z)}$$

$$P(X, Y|Z) = \frac{P(X,Z|Y)\ P(Y)}{P(Z)}$$

n Random Variables

Analogously for n random variables $X_{1:n}$ (stored as a rank n tensor)

Joint: $P(X_{1:n})$

Marginal: $P(X_1) = \sum_{X_{2:n}} P(X_{1:n})$,

Conditional: $P(X_1|X_{2:n}) = \frac{P(X_{1:n})}{P(X_{2:n})}$

Mapping joint atomic events to probabilities.

Storing probabilities
in an n-dimensional array (tensor)

$$p[\underbrace{\ldots,\ldots,\ldots,\quad,\ldots}_{n \text{ entries}}]$$

Product rule and Bayes‘ theorem:

$$P(X_{1:n}) = \prod_{i=1}^{n} P(X_i|X_{i+1:n})$$

$$P(X_1|X_{2:n}) = \frac{P(X_2|X_1,X_{3:n})\ P(X_1|X_{3:n})}{P(X_2|X_{3:n})}$$

https://www.programming-idioms.org/idiom/27/create-a-3-dimensional-array/452/rust

Naïve Bayes

Bayes' Classifier

- Probabilistic classification
  - Estimate (soft classification) $\hat{\boldsymbol{f}}_{\boldsymbol{NBsoft}}(\boldsymbol{o}) = \boldsymbol{P}(\boldsymbol{c}|\boldsymbol{o})$
  - Classification (hard) $\hat{\boldsymbol{f}}_{\boldsymbol{NB}}(\boldsymbol{o}) = \underset{\boldsymbol{c}\in\boldsymbol{C}}{\mathbf{argmax}}\,\boldsymbol{P}(\boldsymbol{c}|\boldsymbol{o})$

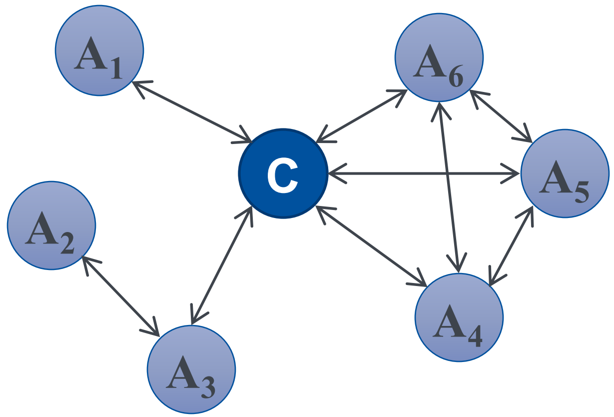
- Bayes Theorem:

$$P(c \mid o) = \frac{P(c) \cdot P(o \mid c)}{P(o)}$$

  - $P(o)$: Probability of the object – constant for all categories
  - $P(c)$: ***Prior*** probability of a category
  - $P(o|c)$: Probability to observe o in c ***(Likelihood)***
  - $P(c|o)$: Probability that observation o should be classified as c ***(Posterior)***

Estimate from training data

Non-naïve Bayes



Attributes and class label are (interdependent) random variables

Resolving the Probabilities with all Dependencies

Non-naive!!!

$$P(c \mid o) = \frac{P(c) \cdot P(o \mid c)}{P(o)}$$

From objects
to attributes

$$P(c|x_1, \ldots, x_m) = \frac{P(x_1,\ldots,x_m|c) \cdot P(c)}{P(x_1,\ldots,x_m)} =$$

$$= \frac{P(x_1,\ldots,x_{m-1}|x_m,c) \cdot P(x_m|c) \cdot P(c)}{P(x_1,\ldots,x_m)} =$$

$$= \frac{P(x_1,\ldots,x_{m-2}|x_{m-1},x_m,c) \cdot P(x_{m-1}|x_m,c) \cdot P(x_m|c) \cdot P(c)}{P(x_1,\ldots,x_m)} =$$

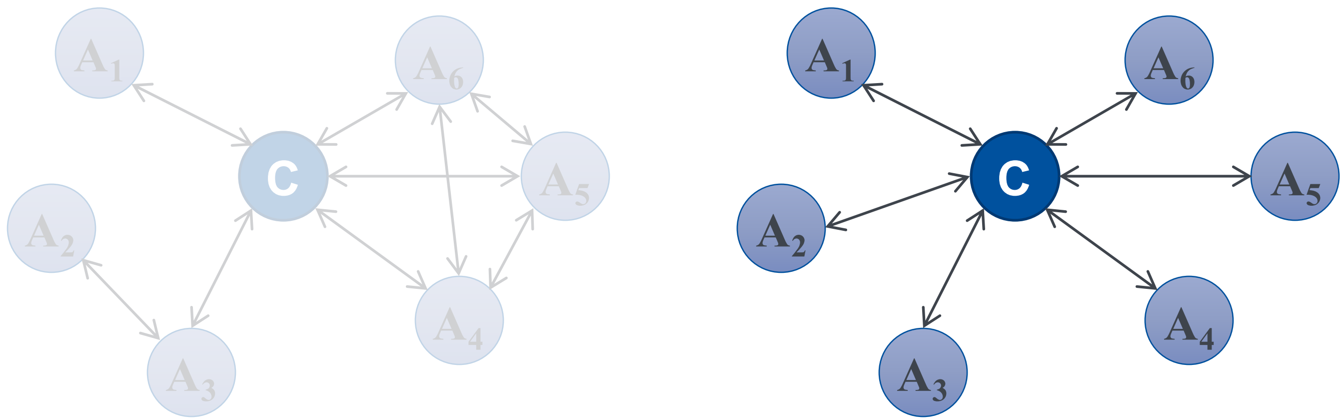
$$= \frac{P(x_1,\ldots,x_{m-3}|x_{m-2},x_{m-1},x_m,c) \cdot \underbrace{P(x_{m-2}|x_{m-1},x_m,c)}}{} \cdot$$

$$\cdot \frac{\overbrace{P(x_{m-1}|x_m,c)} \cdot P(x_m|c) \cdot P(c)}{P(x_1,\ldots,x_m)} = \ldots$$

impossible to measure,

getting worse and worse with increasing number of attributes

Naïve Bayes Assumption



- (Naive) assumption:
  - The class label only depends bi-laterally on the attribute values:

$$P(o|c) = P(x_1, \dots, x_m|c) =$$
$$= P(x_1, \dots, x_{m-1}|x_m, c)P(x_m|c) = \ \dots$$
$$= P(x_1|x_2, \dots, x_m, c)P(x_2|x_3, \dots, x_m, c) \dots P(x_{m-1}|x_m, c)P(x_m|c) \approx$$
$$\approx P(x_1|c)P(x_2|c) \dots P(x_{m-1}|c)P(x_m|c)$$

Simplifying Probabilities with Naïve Bayes Assumption

From objects
to attributes,
from non-naive
to naive Bayes

$$P(c \mid o) = \frac{P(c) \cdot P(o \mid c)}{P(o)}$$

$$P(c \mid x_1, \cdots, x_m) \approx \frac{P(c)}{P(x_1, \cdots, x_m)} \cdot \prod_{j=1}^{m} P(x_j \mid c)$$

$$P(c \mid x_1, \cdots, x_m) \propto P(c) \cdot \prod_{j=1}^{m} P(x_j \mid c)$$

To be estimated from training data: which fraction in a class has a certain attribute value?

Example of Naive Bayes

Wine

Color	Age	Class
Red	Old	Good
White	Old	Good
White	Young	Good
Red	Old	Bad
White	Young	bad

To classify: A new instance wine(red, young)
Calculate value of the decision rule for both classes:

Class „good“:
P (good|red, young) = P (good) · P(red|good) · P(young|good) =

Class „bad“:
P (bad|red, young) = P (bad) · P(red| bad) · P(young| bad) =

⇒ Higher probability for class „bad“
⇒ Naive Bayes classifies new instance as „bad“

Estimating probabilities

- In general:
  - Use count information, i.e. frequencies

- Category priors:

$$P(c_i) = \frac{\left|\left\{(x_1, \cdots, x_m, c) \in O : c = c_i\right\}\right|}{|O|}$$

- Example: Mushrooms

$$P(\text{edible}) = \frac{79}{100} = 0.79$$

$$P(\text{poisonous}) = \frac{21}{100} = 0.21$$

„If you find a mushroom and do not know anything about it, there is a 21% chance it is poisonous"

Estimating probabilities

- Categorial features:
  - Use count frequencies:

$$P(x_j = v \mid c_i) = \frac{\left|\{(x_1,\cdots,x_m,c) \in O : x_j = v \wedge c = c_i\}\right|}{\left|\{(x_1,\cdots,x_m,c) \in O : c = c_i\}\right|}$$

- Example:

$$P(\text{gill}-\text{color} = \text{p} \mid \text{poisonous}) = \frac{3}{21}$$

$$P(\text{bruises} = \text{f} \mid \text{edible}) = \frac{13}{79}$$

$$P(\text{habitat} = \text{d} \mid \text{poisonous}) = \frac{0}{21}$$

Zero Probabilities!

Feature		Edible	Poisonous
cap-shape	bell=b	29	0
	convex=x	34	20
	flat=f	13	1
	sunken=s	3	0
cap-surface	fibrous=f	14	0
	scaly=y	36	13
	smooth=s	29	8
cap-color	brown=n	12	9
	gray=g	7	0
	white=w	19	12
	yellow=y	41	0
bruises	bruises=t	66	21
	no=f	13	0
gill-spacing	close=c	65	21
	crowded=w	14	0
gill-size	broad=b	64	0
	narrow=n	15	21
gill-color	black=k	20	8
	brown=n	22	6
	gray=g	10	0
	pink=p	10	3
	white=w	17	4
habitat	grasses=g	28	8
	meadows=m	28	0
	paths=p	8	0
	urban=u	7	13
	woods=d	8	0

Smoothing for Naive Bayes

$$P(c \mid x_1, \cdots, x_m) \propto P(c) \cdot \prod_{j=1}^{m} P(x_j \mid c)$$

One entry zero: overall product zero !!!

- Laplace smoothing

$$P_{\text{Laplace}}(x_j = v \mid c_i) = \frac{\left|\{(x_1, \cdots, x_m, c) \in O : x_j = v \wedge c = c_i\}\right| + 1}{\left|\{(x_1, \cdots, x_m, c) \in O : c = c_i\}\right| + \left|V_j\right|}$$

  - Where V_j is the set of values for attribute x_j

- Generalized additive smoothing (Lidstone):

$$P_{\text{Lidstone}}(x_j = v \mid c_i) = \frac{\left|\{(x_1, \cdots, x_m, c) \in O : x_j = v \wedge c = c_i\}\right| + \lambda}{\left|\{(x_1, \cdots, x_m, c) \in O : c = c_i\}\right| + \left|V_j\right| \cdot \lambda}$$

  - with Parameter λ

Example

Estimating probabilities using smoothing

- Example:

$$P(\text{gill}-\text{color}=\text{p}\,|\,\text{poisonous})=\frac{4}{26}$$

$$P(\text{bruises}=\text{f}\,|\,\text{edible})=\frac{14}{81}$$

$$P(\text{habitat}=\text{d}\,|\,\text{poisonous})=\frac{1}{26}$$

Feature		Edible	Poisonous
cap-shape	bell=b	29 **+1**	0 **+1**
	convex=x	34 **+1**	20 **+1**
	flat=f	13 **+1**	1 **+1**
	sunken=s	3 **+1**	0 **+1**
cap-surface	fibrous=f	14 **+1**	0 **+1**
	scaly=y	36 **+1**	13 **+1**
	smooth=s	29 **+1**	8 **+1**
cap-color	brown=n	12 **+1**	9 **+1**
	gray=g	7 **+1**	0 **+1**
	white=w	19 **+1**	12 **+1**
	yellow=y	41 **+1**	0 **+1**
bruises	bruises=t	66 **+1**	21 **+1**
	no=f	13 **+1**	0 **+1**
gill-spacing	close=c	65 **+1**	21 **+1**
	crowded=w	14 **+1**	0 **+1**
gill-size	broad=b	64 **+1**	0 **+1**
	narrow=n	15 **+1**	21 **+1**
gill-color	black=k	20 **+1**	8 **+1**
	brown=n	22 **+1**	6 **+1**
	gray=g	10 **+1**	0 **+1**
	pink=p	10 **+1**	3 **+1**
	white=w	17 **+1**	4 **+1**
habitat	grasses=g	28 **+1**	8 **+1**
	meadows=m	28 **+1**	0 **+1**
	paths=p	8 **+1**	0 **+1**
	urban=u	7 **+1**	13 **+1**
	woods=d	8 **+1**	0 **+1**

Example

Example: Classification

- Mushroom: $o = (f, y, w, t, c, n, p, g)$
- Category: poisonous:

$$P(\text{poisonous}) = \frac{21}{100}$$

$$P(\text{cap}-\text{shape} = \text{f} \mid \text{poisonous}) = \frac{2}{25}$$

$$P(\text{cap}-\text{surface} = \text{y} \mid \text{poisonous}) = \frac{14}{24}$$

$$P(\text{cap}-\text{color} = \text{w} \mid \text{poisonous}) = \frac{13}{25}$$

$$P(\text{bruises} = \text{t} \mid \text{poisonous}) = \frac{22}{23}$$

$$P(\text{gill}-\text{spacing} = \text{c} \mid \text{poisonous}) = \frac{22}{23}$$

$$P(\text{gill}-\text{size} = \text{n} \mid \text{poisonous}) = \frac{22}{23}$$

$$P(\text{gill}-\text{color} = \text{p} \mid \text{poisonous}) = \frac{4}{26}$$

$$P(\text{habitat} = \text{g} \mid \text{poisonous}) = \frac{9}{26}$$

P(poisonous|o) = 0.000247

Feature		Edible	Poisonous
cap-shape	bell=b	29 +1	0 +1
	convex=x	34 +1	20 +1
	flat=f	13 +1	1 +1
	sunken=s	3 +1	0 +1
cap-surface	fibrous=f	14 +1	0 +1
	scaly=y	36 +1	13 +1
	smooth=s	29 +1	8 +1
cap-color	brown=n	12 +1	9 +1
	gray=g	7 +1	0 +1
	white=w	19 +1	12 +1
	yellow=y	41 +1	0 +1
bruises	bruises=t	66 +1	21 +1
	no=f	13 +1	0 +1
gill-spacing	close=c	65 +1	21 +1
	crowded=w	14 +1	0 +1
gill-size	broad=b	64 +1	0 +1
	narrow=n	15 +1	21 +1
gill-color	black=k	20 +1	8 +1
	brown=n	22 +1	6 +1
	gray=g	10 +1	0 +1
	pink=p	10 +1	3 +1
	white=w	17 +1	4 +1
habitat	grasses=g	28 +1	8 +1
	meadows=m	28 +1	0 +1
	paths=p	8 +1	0 +1
	urban=u	7 +1	13 +1
	woods=d	8 +1	0 +1

Example

Example: Classification

- Mushroom: $o = (f,y,w,t,c,n,p,g)$
- Category: edible:

$$P(\text{edible}) = \frac{79}{100}$$

$$P(\text{cap-shape} = f \mid \text{edible}) = \frac{14}{83}$$

$$P(\text{cap-surface} = y \mid \text{edible}) = \frac{37}{82}$$

$$P(\text{cap-color} = w \mid \text{edible}) = \frac{20}{83}$$

$$P(\text{bruises} = t \mid \text{edible}) = \frac{67}{81}$$

$$P(\text{gill-spacing} = c \mid \text{edible}) = \frac{66}{81}$$

$$P(\text{gill-size} = n \mid \text{edible}) = \frac{16}{81}$$

$$P(\text{gill-color} = p \mid \text{edible}) = \frac{11}{84}$$

$$P(\text{habitat} = g \mid \text{edible}) = \frac{29}{84}$$

Poisonous!

P(edible|o) = 0.000087

	Feature	Edible	Poisonous
cap-shape	bell=b	29 +1	0 +1
	convex=x	34 +1	20 +1
	flat=f	13 +1	1 +1
	sunken=s	3 +1	0 +1
cap-surface	fibrous=f	14 +1	0 +1
	scaly=y	36 +1	13 +1
	smooth=s	29 +1	8 +1
cap-color	brown=n	12 +1	9 +1
	gray=g	7 +1	0 +1
	white=w	19 +1	12 +1
	yellow=y	41 +1	0 +1
bruises	bruises=t	66 +1	21 +1
	no=f	13 +1	0 +1
gill-spacing	close=c	65 +1	21 +1
	crowded=w	14 +1	0 +1
gill-size	broad=b	64 +1	0 +1
	narrow=n	15 +1	21 +1
gill-color	black=k	20 +1	8 +1
	brown=n	22 +1	6 +1
	gray=g	10 +1	0 +1
	pink=p	10 +1	3 +1
	white=w	17 +1	4 +1
habitat	grasses=g	28 +1	8 +1
	meadows=m	28 +1	0 +1
	paths=p	8 +1	0 +1
	urban=u	7 +1	13 +1
	woods=d	8 +1	0 +1

Implementation Detail

$$P(c \mid x_1, \cdots, x_m) \propto P(c) \cdot \prod_{j=1}^{m} P(x_j \mid c)$$

- High number of attributes and attribute values:
  - Many multiplications
  - Very small values
- Use logarithm:

Risk of issues with accurate representation

$$\log P(c \mid x_1, \cdots, x_m) \propto \log\left(P(c) \cdot \prod_{j=1}^{m} P(x_j \mid c) \right)$$

$$= \log(P(c)) + \sum_{j=1}^{m} \log(P(x_j \mid c))$$

Naive Bayes for Text Classification

Intermezzo:

Continuous Probabilistic Reasoning

Numeric values

- Single observations are too specific
  - $P(\text{Height} = 83\text{ cm})$?
  - $P(\text{Height} = 83.81\text{ cm})$?

Gender	weight	height
F	52,33	66,40
F	64,39	73,64
F	51,07	65,64
F	66,27	74,76
F	69,68	76,81
F	51,03	65,62
F	63,48	73,09
M	79,17	83,81
M	62,40	74,23
M	62,94	74,54
M	67,22	76,98
M	59,27	72,44
M	59,04	72,31
M	62,81	74,46
M	56,14	70,65
M	78,44	83,40
M	74,27	81,01
M	69,61	78,35

Distributions over continuous domains

- Let X be a continuous random variable
- The **probability density function** (pdf) $p(x) \in [0, \infty)$ defines the probability

$$P(a \leq x \leq b) = \int_a^b p(x)dx \in [0,1]$$

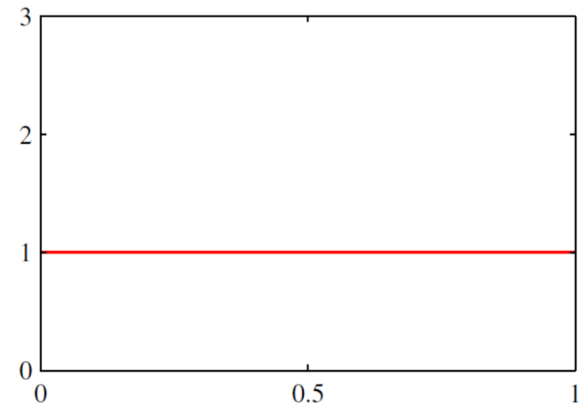
- The **cumulative probability distribution**

$$F(y) = P(x \leq y) = \int_{-\infty}^{y} p(x)dx \in [0,1]$$

is the cumulative integral with $\lim_{y \to \infty} F(y) = 1$

Uniform continuous distribution over $[0, 1]$

- e.g. Python



`random.random()`

Return the next random floating point number in the range [0.0, 1.0).

`random.uniform(a, b)`

Return a random floating point number *N* such that `a <= N <= b` for `a <= b` and `b <= N <= a` for `b < a`.

The end-point value `b` may or may not be included in the range depending on floating-point rounding in the equation `a + (b-a) * random()`.

Numeric values

- Single observations are too specific
  - $P(\text{Height} = 83\text{ cm})$?
  - $P(\text{Height} = 83.81\text{ cm})$?

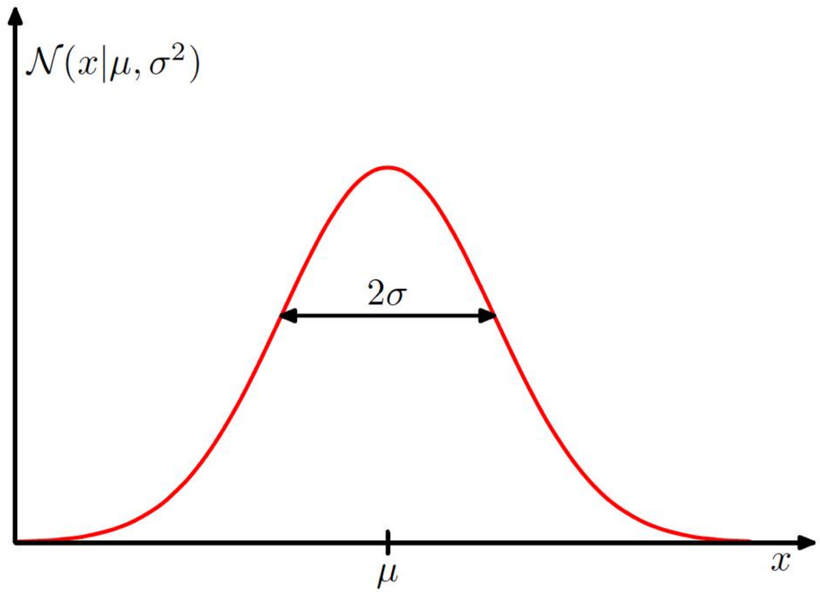
Gender	weight	height
F	52,33	66,40
F	64,39	73,64
F	51,07	65,64
F	66,27	74,76
F	69,68	76,81
F	51,03	65,62
F	63,48	73,09
M	79,17	83,81
M	62,40	74,23
M	62,94	74,54
M	67,22	76,98
M	59,27	72,44
M	59,04	72,31
M	62,81	74,46
M	56,14	70,65
M	78,44	83,40
M	74,27	81,01
M	69,61	78,35

1-dim Gaussian distribution (also „Normal distribution")

- 1-dimensional:

$$\mathcal{N}(x|\mu,\sigma^2) = p(x|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- mean μ
- variance σ^2
- standard deviation σ
- „standard normal",
  iff $\mu = 0, \sigma = 1$



n-dim Gaussian distribution (also „Normal distribution“)

- n-dim Gaussian in normal form:

$$p(\boldsymbol{x}) = \mathcal{N}(\,\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\,)$$

$$= \frac{1}{\sqrt{(2\pi)^n \, |\Sigma|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

with **mean** $\boldsymbol{\mu}$ and **covariance matrix** $\boldsymbol{\Sigma}$

Numeric Values

Numeric values

- Single observations are too specific
  - $P(\text{Height} = 83 \text{ cm})$?
  - $P(\text{Height} = 83.81 \text{ cm})$?
- Assumption:
  - Attribute values follow a normal distribution (in each class).

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Maximum-likelihood estimate of parameters:

$$\mu_i = \frac{1}{n} \sum_o x_i$$

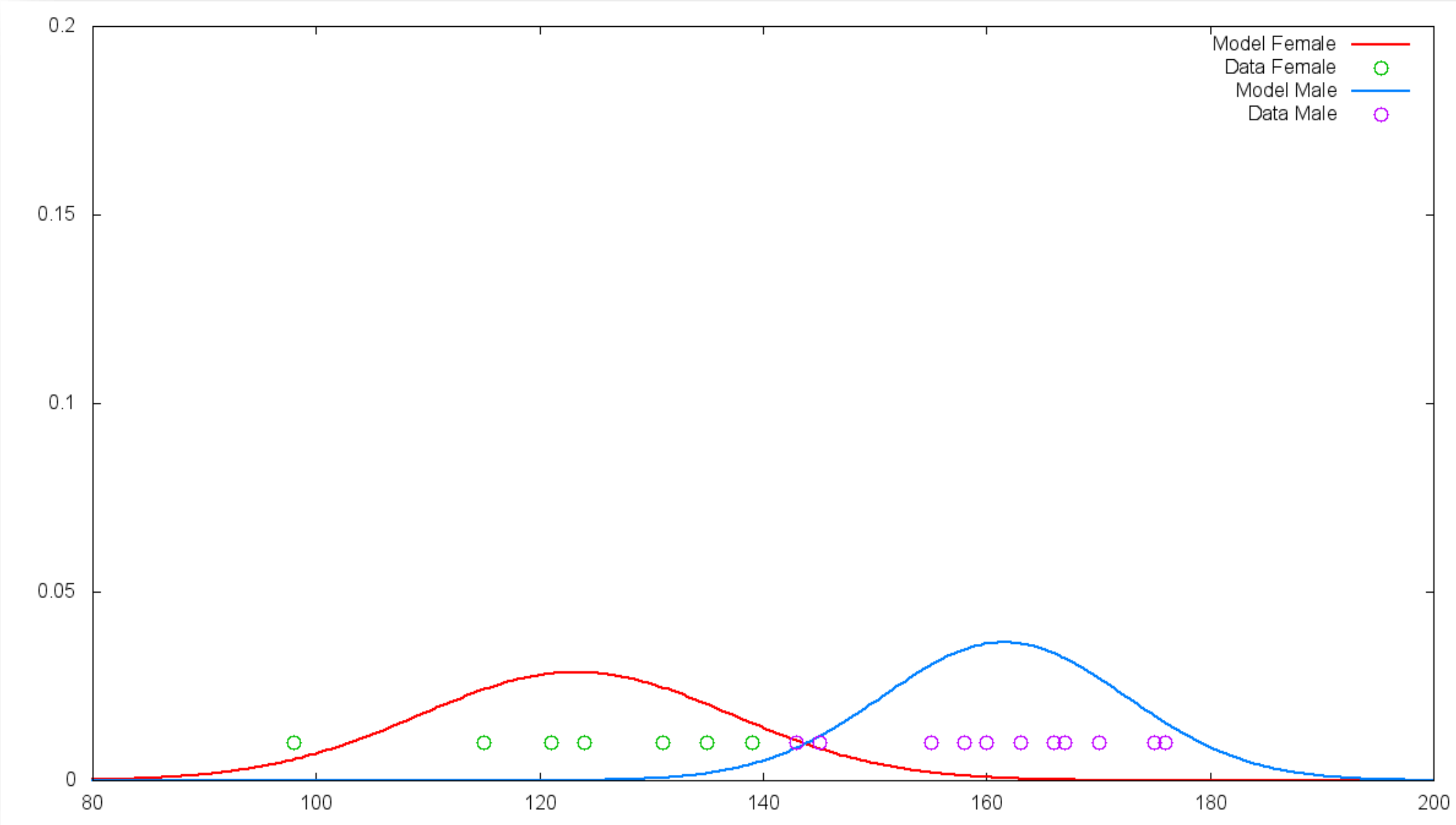
$$\text{var}_i = \frac{1}{n-1} \sum_o (x_i - \mu_i)^2$$

$$\sigma_i = \sqrt{\text{var}_i}$$

Gender	weight	height
F	52,33	66,40
F	64,39	73,64
F	51,07	65,64
F	66,27	74,76
F	69,68	76,81
F	51,03	65,62
F	63,48	73,09
M	79,17	83,81
M	62,40	74,23
M	62,94	74,54
M	67,22	76,98
M	59,27	72,44
M	59,04	72,31
M	62,81	74,46
M	56,14	70,65
M	78,44	83,40
M	74,27	81,01
M	69,61	78,35

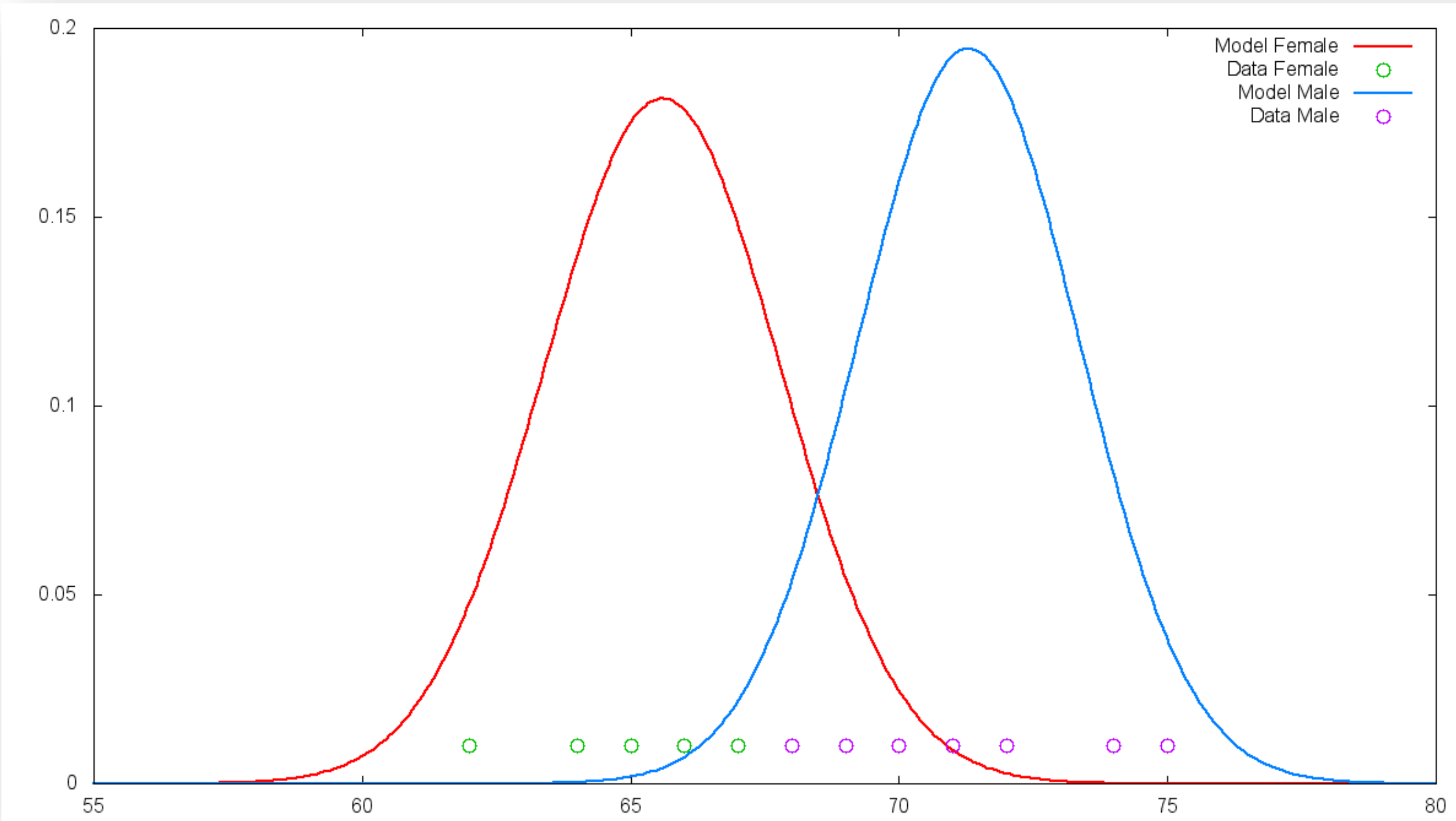
Modelling Distribution for Numeric Data

- Learning the parameters for height



Modelling Distribution for Numeric Data

- Learning the parameters for weight

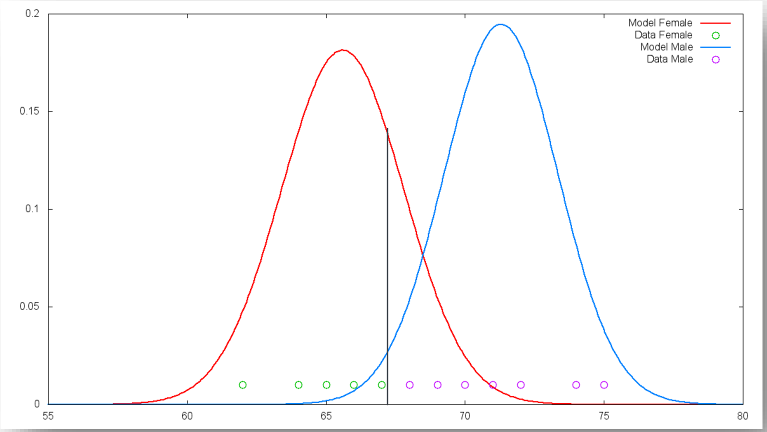


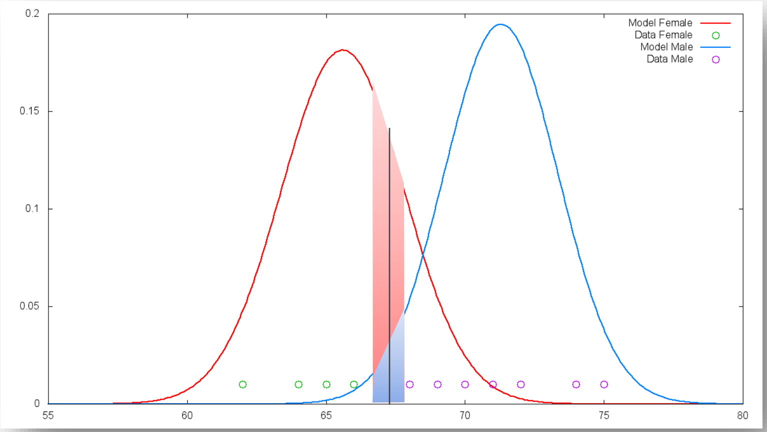
Making Use of the Distributions

- Still:
  - $P(\text{weight} = 67 \mid c = \text{f}) = 0$

$$P(x_i < t) = \int_{-\infty}^{t} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

- Solution:
  - Assume a weight „around“ the actual value.
  - $P(\text{weight} = 67 \pm \varepsilon \mid c = \text{f})$
- Implementation:
  - Use value of density function





Example

- New data point:
  - Weight: 67, Height: 155
- Parameters
  - Weight:
    - $\mu_{wF} = 65.57$ $\sigma_{wF} = 2.22$
    - $\mu_{wM} = 71.57$ $\sigma_{wM} = 2.02$
  - Height:
    - $\mu_{hF} = 123.29$ $\sigma_{hF} = 13.89$
    - $\mu_{hM} = 161.64$ $\sigma_{hM} = 10.90$
- Priors:

$$P(\text{female}) = \frac{7}{18}$$

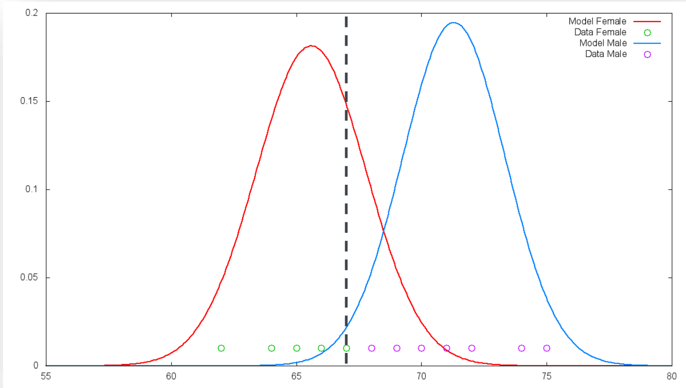
$$P(\text{male}) = \frac{11}{18}$$

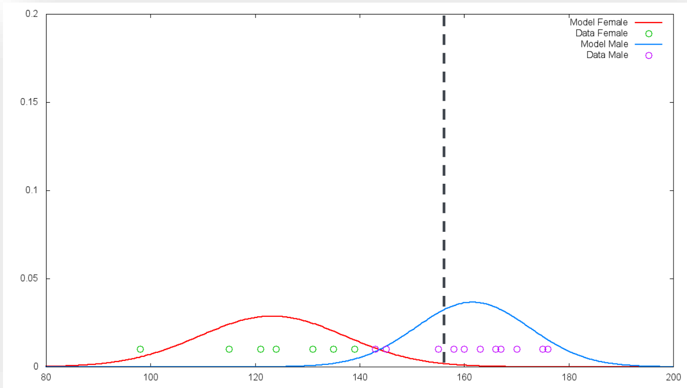
Do the priors make sense?

Gender	weight	height
F	66	124
F	66	115
F	64	121
F	69	139
F	62	98
F	67	135
F	65	131
M	74	170
M	68	166
M	70	155
M	72	167
M	71	158
M	72	175
M	69	143
M	72	163
M	75	160
M	70	145
M	71	176

Example

- Class „Female“:
  - $P(\text{height} = 155 \mid c = \text{f}\,) = 0.0021$
  - $P(\text{weight} = 67 \mid c = \text{f}\,) = 0.1459$
  - With prior:
    - $P(\text{f} \mid o) = 0.00012$
  - Without prior:
    - $P(\text{f} \mid o) = 0.00031$
- Class „Male“:
  - $P(\text{height} = 155 \mid c = \text{m}\,) = 0.0304$
  - $P(\text{weight} = 67 \mid c = \text{m}\,) = 0.0223$
  - With prior:
    - $P(\text{m} \mid o) = 0.00041$
  - Without prior:
    - $P(\text{m} \mid o) = 0.00068$





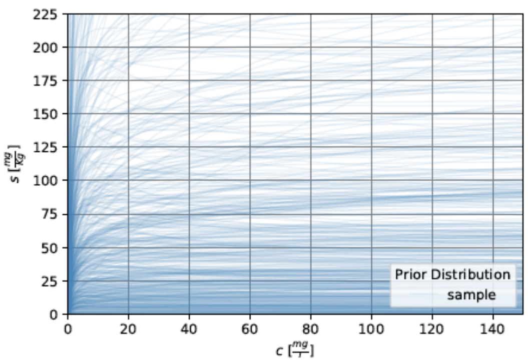
Conclusion: `Graphical Models´

- You have now seen the first, specific `graphical model´: probabilistic reasoning with Naive Bayes
- `graphical models‘ are probabilistic models with multiple random variables and dependencies
- `graphical models´ are a general framework for modelling problems:
  - regression and classification, decision making
  - unsupervised learning
  - reinforcement learning
  - language modeling

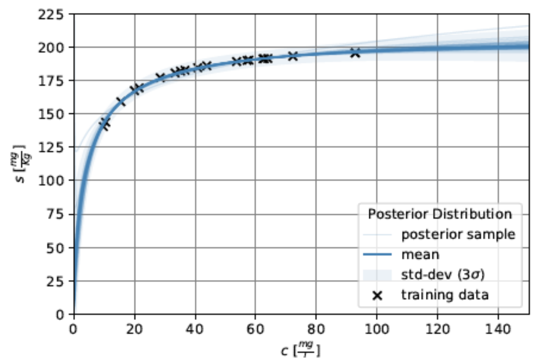
Probabilistic Machine Learning –
Lecture in winter terms

Current research at Analytic Computing: Predicting functions with Bayesian methods

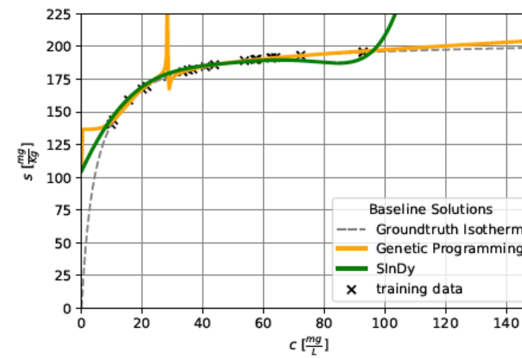
Diagrams by Tim Schneider



(a) Prior Distribution



(b) Posterior Distribution



(c) Baseline Solutions

$$s = s_T \sum_{i=1}^{n} f_i \prod_{j=1}^{m_i} \left(\frac{q_{ij} c^{\alpha_{ij}}}{1 + p_{ij} c^{\beta_{ij}}} \right)^{\gamma_{ij}},$$

Universität Stuttgart
IPVS

Thank you!

Steffen Staab

E-Mail Steffen.staab@ipvs.uni-stuttgart.de

Telefon +49 (0) 711 685-To be defined

www. ipvs.uni-stuttgart.de/departments/ac/

Universität Stuttgart
Analytic Computing, IPVS
Universitätsstraße 32, 50569 Stuttgart