02450 Introduction to Machine Learning and Data Mining

Week 4: Probability and probability densities

Bjørn Sand Jensen

25 February 2025

DTU Compute, Technical University of Denmark

Today

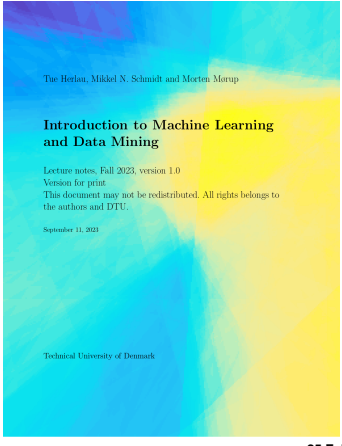
Feedback Groups of the day:

Jakob Utne, Philipp Ruf, Jeppe Klitgaard, Rasmus Lund Hansen, Martin Holme Surlykke, Linus Juni, Astrid Mantzius Jepsen, Peter Møller Naur, Nehal Sharma, Kenneth Plum Toft, Martin Herwin Schaarup, Nikolaj Nguyen, Viola Sofie Demuth Andersen, David Borgbjærg Madsby, Kacper Jan Rokosz, Ella Korre Stenholt, Maria Siembor, Bogdan-Petru Pascut, Alexander Halse Hansen, William Christian Brøchner Søndergaard, Josephine Schwarz, Matilde Marie Grønkjær Matell, Riccardo Nicora, Zografia Lelidou, Gustav Heron Melhus, Alexandru ., Emil Bom, Shengdi Chang, Frida Marie Lund Jeppesen, Bror Sigurd Bruland, Sigurbjorg Katla Valdimarsdottir, Daniel Malik Mapouyat, Mikkel Ole Neldahl Warner, Alberto Magagna, Nicolai Christian Brock, Oliver Lamine Thiam, Alfred Bonde Jacobsen, Alessandro Nardin, Andreas Foss Højgaard Rasmussen, Divya Khurana, Benedikt Mainzer, Viktor Ellehammer Andersen, Joachim Bo Jensen

Reading/homework material:

Chapter 5,6

P5.1, P6.1, P6.2



Lecture Schedule

1. Introduction
   4 February: C1,C2

Data: Feature extraction, and visualization

2. Summary statistics, similarity and visualization
   11 February: C4,C7
3. Computational linear algebra and PCA
   18 February: C3
4. **Probability and probability densities**
   25 February: C5, C6

Supervised learning: Classification and regression

5. Decision trees and linear regression
   4 March: C8, C9 **(Project 1 due 6 March at 17:00)**
6. Overfitting, cross-validation and Nearest Neighbor
   11 March: C10, C12
7. Performance evaluation, Bayes, and Naive Bayes
   18 March: C11, C13
8. Artificial Neural Networks and Bias/Variance
   25 March: C14, C15
9. AUC and ensemble methods
   1 April: C16, C17

Unsupervised learning: Clustering and density estimation

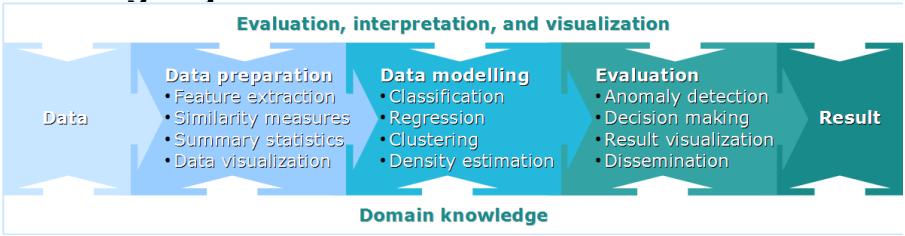
10. K-means and hierarchical clustering
    8 April: C18 **(Project 2 due 10 April at 17:00)**
11. Mixture models and density estimation
    22 April: C19, C20
12. Association mining
    29 April: C21

Recap

13. Recap and discussion of the exam
    6 May: C1-C21

Online help: Piazza
Videos of lectures: `https://panopto.dtu.dk`
Streaming of lectures: Zoom (link on DTU Learn)

Learning Objectives



Learning Objectives

- Understand basics of probability theory and stochastic variables in the discrete setting
- Understand probabilistic concepts such as expectations, independence and the Bernoulli distribution
- Understand the maximum likelihood principle for repeated binary events
- Understand probability densities and related concepts
- Derive cost-functions from likelihood functions using Bayes' theorem"

Plan for today:

- Lecture 4 (13:00 – $\sim$15:00)
  - Discrete random variables and related concepts
  - Continuous random variables and related concepts
  - Machine learning as statistical inference
- Exercises ($\sim$ 15:00–17:00)

Probabilities in machine learning

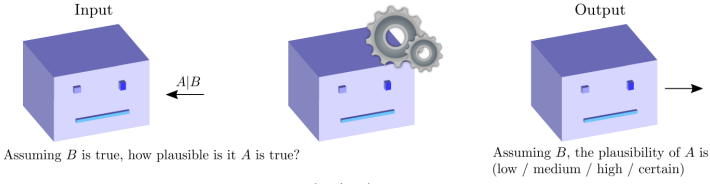
Pragmatically: A big part of AI is dealing with **uncertainty** and **incomplete information**. Probabilities is the formal framework for doing so.

Algorithmically: **If** an image belongs to a particular category is a discrete event. The **probability** it belongs to a category is continuous. Algorithmically, easier to optimize continuous quantities.

Convenience: There are boiler-plate ideas for transforming a **probabilistic** assumption into an algorithm (maximum likelihood).

Probabilities

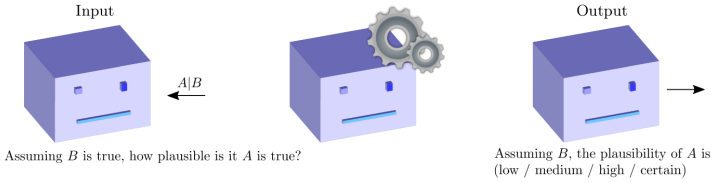
We reason about a proposition A in light of evidence B:



The degree-of-belief that A is true given B is accepted as true is at a level x

- A number between 0 and 1
- A and B are always binary (true/false) propositions
- Represents a *state of knowledge*

Probabilities: Trial example



G : *The accused is guilty*

E_1 : *A car similar to his was seen at the crime scene.*

E_2 : *A large sum of money was found in his posession*

E_3 : *His fingerprints was found at the door of the bank.*

Probabilities express states-of-knowledge

$$E \equiv E_1 \text{ and } E_2 \text{ and } E_3$$

$$P(G|E) > P(G|E_2)$$

Binary propositions

A binary proposition is a statement which is either true or false (we might not know, but someone with complete knowledge would)

A : *In 49 BCE, Caesar crossed the Rubicon*

B : *Acceleration sensor 39 measures more than 0.85*

C : *Patient 901 has high cholesterol*

Propositions can be combined with **and**, **or** and **not**:

$AB \equiv$ True if A and B are both true

$A + B \equiv$ True if either A or B are true

$\overline{A} \equiv$ True if A is false

We define two special propositions which is always **true**/**false**:

1 : *A proposition which is always true*

0 : *A proposition which is always false*

...and the following identities:$A1 = A, \quad A + \overline{A} = 1, \quad \overline{\overline{A}} = A$ and

$$A(B_1 + B_2 + \cdots + B_n) = AB_1 + AB_2 + \cdots + AB_n$$

Quiz 1: Probabilities

Assume we define the following 4 boolean variables.

R_1 : *Handed in report 1*
R_2 : *Handed in report 2*
R_3 : *Handed in report 3*
F : *Student failed 02450*

How would you express the probability of the statement:

If a student hand in report 1, 2 and 3, the chance of passing 02450 is greater than 90%?

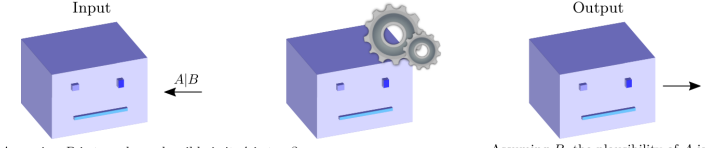
A. $P(R_1R_2R_3|F) > 0.9$

B. $P(\overline{F}|R_1 + R_2 + R_3) > 0.9$

C. $P(\overline{F}|R_1R_2R_3) > 0.9$

D. $P(R_1 + R_2 + R_3|F) > 0.9$

E. Don't know.



Rules of probability

The sum rule: $P(A|C) + P(\overline{A}|C) = 1$

The product rule: $P(AB|C) = P(B|AC)P(A|C)$

Interpretation:

$P(A|B) = 0$ *(interpretation: given B is true, A is certainly false)*

$P(A|B) = 1$ *(interpretation: given B is true, A is certainly true)*

We also use the shorthand:

$$P(A|1) = P(A)$$

$$p(A) + P(\overline{A}) = 1$$

$$p(AB) = P(A|B)P(B)$$

Remarkably, this is the mathematical basis for this course

Marginalization and Bayes' theorem

Sum rule	$P(A\|C) + P(\overline{A}\|C) = 1$
Product rule	$P(AB\|C) = P(B\|AC)P(A\|C)$

$$
\begin{aligned}
P(B|C) &= P(B|C)\left[P(A|BC) + P(\overline{A}|BC)\right] = P(AB|C) + P(\overline{A}B|C) \\
&= P(B|AC)P(A|C) + P(B|\overline{A}C)P(\overline{A}|C).
\end{aligned}
$$

$$
\begin{aligned}
P(A|BC) &= \frac{P(B|AC)P(A|C)}{P(B|C)} \\
&= \frac{P(B|AC)P(A|C)}{P(B|AC)P(A|C) + P(B|\overline{A}C)P(\overline{A}|C)}.
\end{aligned}
$$

Exclusive and exhaustive events

A_1 : *The side* ⚀ *face up.* A_2 : *The side* ⚁ *face up.*

A_3 : *The side* ⚂ *face up.* A_4 : *The side* ⚃ *face up.*

A_5 : *The side* ⚄ *face up.* A_6 : *The side* ⚅ *face up.*

- When no two propositions can be true at the same time, they are said to be **mutually exclusive**: $A_iA_j = 0$ for $i \neq j$
- Consider any two events A and B

$$P(A + B) = P(A) + P(B) - P(AB)$$

- In general, for n mutually exclusive events

$$P(A_1 + A_2 + \cdots + A_n) = \sum_{i=1}^{n} P(A_i)$$

- A set of events is **exhaustive** if one has to be true: $A_1 + \cdots + A_n = 1$. Then:

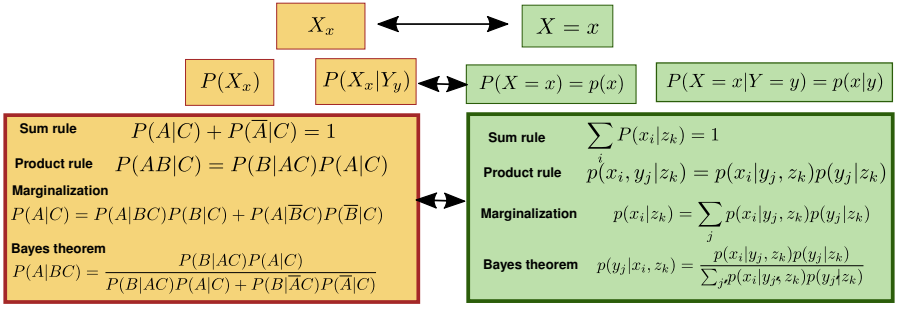
$$\sum_{i=1}^{n} P(A_i) = P(A_1 + A_2 + \cdots + A_n) = 1$$

Stochastic variables

- We often measure numerical quantities (number of children, age of a patient, etc.)
- Suppose a quantity X (number of children) takes a value $x = 3$. We can write this as the binary event X_3 and in general:

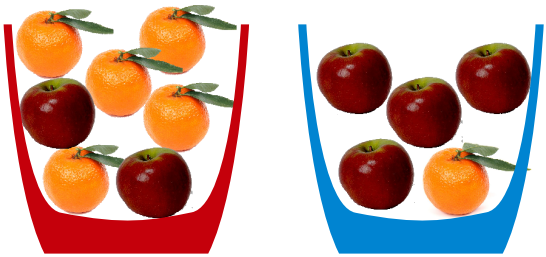
$$X_x : \{\text{The binary event that } X \text{ is equal to the number } x\}$$

- Stochastic variable simplify this notation by the definition:



Example: Everyday probabilities... are easy?

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?



Orange (clementine) taken from: `https://commons.wikimedia.org/wiki/File:Clementine_orange.jpg`

Apple taken from: `https://upload-wikimedia.org/wikipedia/commons/3/32/Dark_apple.png`

Probabilities

- Sum rule

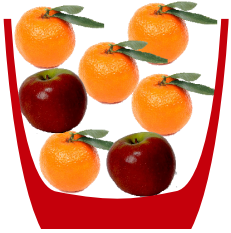
$$p(x) = p(x, y = 0) + p(x, y = 1)$$

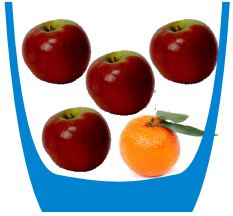
- Product rule

$$\begin{aligned} p(x, y) &= p(x \mid y)p(y) \\ &= p(y \mid x)p(x) \end{aligned}$$

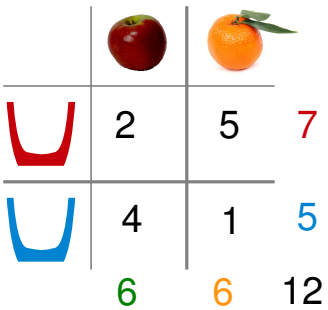
- Bayes rule

$$\begin{aligned} p(x \mid y) &= \frac{p(x, y)p(x)}{p(y)} \\ &= \frac{p(y \mid x)p(x)}{p(y)} \end{aligned}$$





- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?



	2	5	7
	4	1	5
	6	6	12

$$p(o \mid r) = \frac{p(r,o)}{p(r)} = \frac{5/12}{\mathbf{7}/12}$$

$$\begin{aligned} p(r) &= p(r, o=0) + p(r, o=1) \\ &= 2/12 + 5/12 = 7/12 \end{aligned}$$

$$p(r \mid o) = \frac{p(r,o)}{p(o)} = \frac{5/12}{\mathbf{6}/12}$$

$$\begin{aligned} p(r \mid o) &= \frac{p(o \mid r)p(r)}{p(o)} \\ &= \frac{5/\mathbf{7} \times \mathbf{7}/12}{\mathbf{6}/12} \end{aligned}$$

Example: Everyday probabilities... can be tricky!

A medical test for a given disease:

- Correctly identifies the disease 99% of the time (true positives)
- Incorrectly turns out positive 2% of the time (false positives).

You know that

- 1% of the population suffers from the disease.
- As part of a random trial (uniform sampling) you are told to go to the doctor to get tested. The test turns out to be positive.

What is the probability you have the disease?

Hints:
We are looking for $p(Disease|Positive)$

Identify the following probabilities
$p(Positive \mid Disease)$
$p(Positive \mid NoDisease)$
$p(Disease)$
$p(NoDisease)$

Rules of probability:

$$p(y) = \sum_x p(x, y)$$
$$= p(y \mid x)p(x) + p(y \mid \neg x)p(\neg x)$$

$$p(x, y) = p(x \mid y)p(y)$$
$$p(x \mid y) = \frac{p(y|x)p(x)}{p(y)}$$

Quiz 2: Probabilities

Consider a dataset which describe the consumption of delicatessen products in different cities. Each observation in the dataset is a customer, and we record the city the customer is from as well as their consumption of delicatessen. Suppose you are told:

- 17.5 % were from Lisbon, 10.7 % were from Oporto and 71.8 % from the Other region.
- 44.1 % of the costumers from Lisbon spent above the median consumption on delicatessen (DELI).
- 48.9 % of the costumers from Oporto spent above the median consumption on delicatessen (DELI).
- 51.6 % of the costumers from the Other region spent above the median consumption on delicatessen (DELI).

What is the probability based on the wholesale data that a costumer that spent above the median consumption on delicatessen (DELI) come from Lisbon?

A. 7.7 %

B. 15.4 %

C. 44.1 %

D. 59.6 %

E. Don't know.

Independence

Independent: $p(x_i, y_j) = p(x_i)p(y_j)$

Conditionally independent given z_k : $p(x_i, y_j|z_k) = p(x_i|z_k)p(y_j|z_k)$

Expectations

$$\textit{Expectation:} \quad \mathbb{E}[f] = \sum_{i=1}^{N} f(x_i)p(x_i).$$

$$\text{mean: } \mathbb{E}[x] = \sum_{i=1}^{N} x_i p(x_i), \quad \text{Variance: } \operatorname{Var}[x] = \sum_{i=1}^{N} (x_i - \mathbb{E}[x])^2 p(x_i).$$

Example: Uniform probability

$$p(x_i) = \frac{1}{N}$$

$$\mathbb{E}[f] = \frac{1}{N} \sum_{i=1}^{N} f(x_i)$$

$$\mathbb{E}[x] = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$\operatorname{Var}[x] = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mathbb{E}[x])^2$$

Densities and models

- In machine learning, we want to learn a parameter from data
- Models of the data which use parameters are how we do that
- We build models out of simpler building blocks namely distributions and densities (see chapter 5 for descrete variables and chapter 6 for continuous variables). In this course we will u

  - **Bernoulli distribution**
  - The Categorical distribution
  - The Beta density
  - The Multivariate normal density

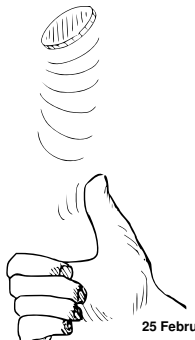




The Bernoulli distribution

- Let $b = 0, 1$ denote a binary event.
- For instance,
  - $b = 0$ corresponds to heads, and $b = 1$ to tails, or
  - $b = 0$ corresponds to a person being ill, and $b = 1$ that a person is well.
- The probability of b is expressed using a parameter θ in the unit interval $[0, 1]$

Bernoulli distribution: $p(b|\theta) = \theta^b(1-\theta)^{1-b}$.



The Bernoulli distribution, repeated events

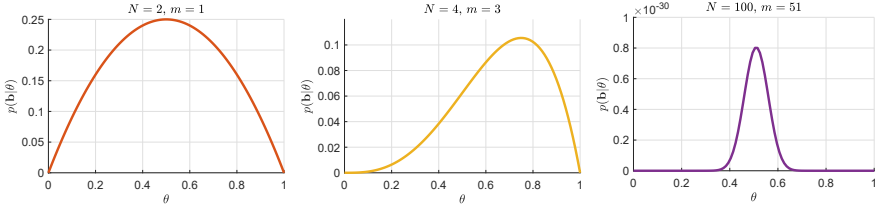
Conditional independence $p(x_i, x_j|z_k) = p(x_i|z_k)p(x_j|z_k)$

- Suppose we observe a sequence $b_1, \ldots, b_N$ of Bernoulli (binary) events.
- For instance, for N patients we record whether person 1 is ill or well ($b_1 = 0$ or $b_1 = 1$) and up to whether patient N is ill or well ($b_N = 0$ or $b_N = 1$)
- When we **know** θ (the chance a person is well or ill), the events are **independent**

Bernoulli distribution: $p(b|\theta) = \theta^b(1-\theta)^{1-b}$.

$$
\begin{aligned}
p(b_1, \cdots, b_N|\theta) &= \prod_{i=1}^{N} p(b_i|\theta) \\
&= \prod_{i=1}^{N} \theta^{b_i}(1-\theta)^{1-b_i} \\
&= \theta^{\sum_{i=1}^{N} b_i}(1-\theta)^{N-\sum_{i=1}^{N} b_i} \\
&= \theta^m(1-\theta)^{N-m}, \quad m = b_1 + b_2 + \cdots + b_N
\end{aligned}
$$

The Bernoulli distribution, maximum likelihood



$$p(b_1, \ldots, b_N|\theta) = \theta^m(1-\theta)^{N-m}$$

An idea for selecting θ^* is **Maximum likelihood**

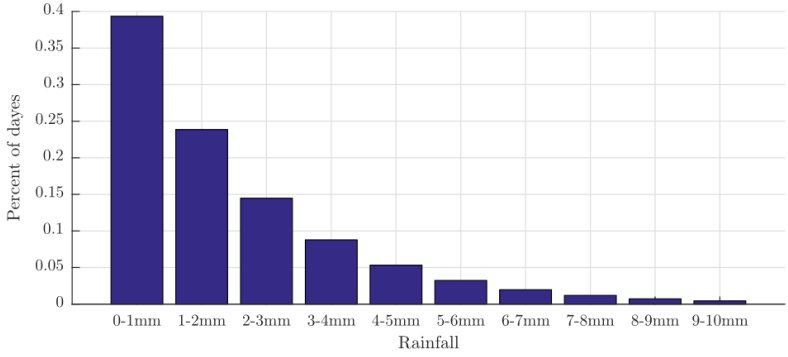
$$\theta^* = \arg\max_{\theta} p(b_1, \ldots, b_N|\theta)$$

The value of θ according to which the data is most plausible

Probability density functions

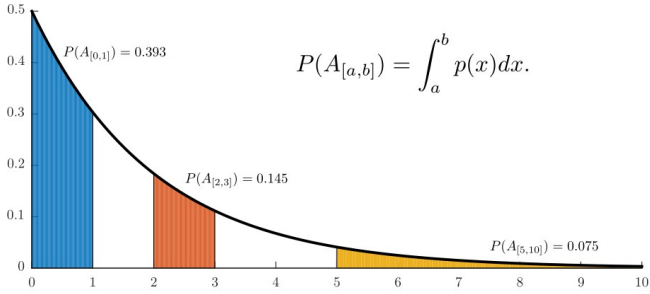
Probability vs. Density

- Suppose we consider the rainfall on an average day r
- **Can not** talk about the probability there will be **exactly** $r = 2.3$ mm of rain, $P(r = 2.3)$ mm
- **Can** talk about the probability there will be **between** 1 and 2 mm of rain



Probability vs. Density

- These probabilities can be **represented** as integrals
- **Events** are **intervals**, the **probability** is the **integral**, the curve is the **density**

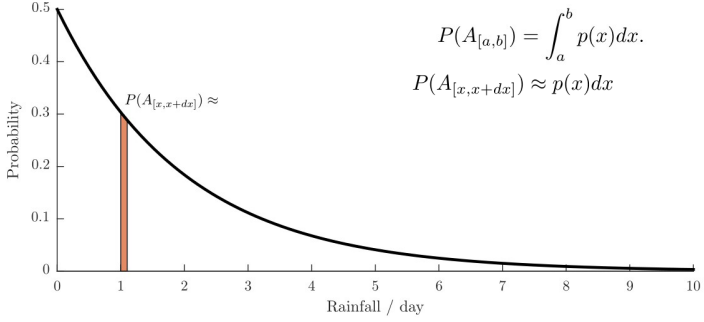


$$P(A_{[a,b]}) = \int_a^b p(x)dx.$$

$A_{[a,b]}$: *There will be between* a *and* b *mm of rain*

Probability vs. Density

- These probabilities can be **represented** as integrals
- **Events** are **intervals**, the **probability** is the **integral**, the curve is the **density**
- What is the probability there will be between 1 and 1.1 mm of rain?



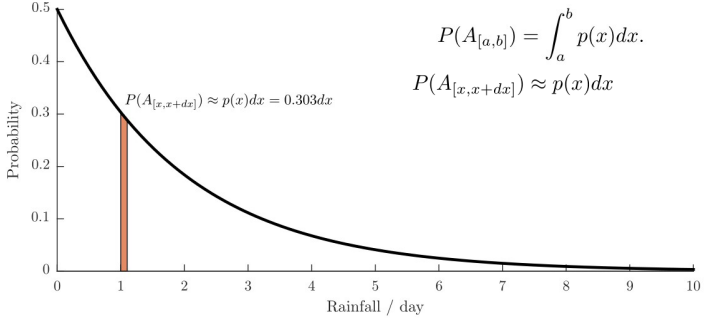
$$P(A_{[a,b]}) = \int_a^b p(x)dx.$$

$$P(A_{[x,x+dx]}) \approx p(x)dx$$

$A_{[a,b]}$: *There will be between* a *and* b *mm of rain*

Probability vs. Density

- These probabilities can be **represented** as integrals
- **Events** are **intervals**, the **probability** is the **integral**, the curve is the **density**
- What is the probability there will be between 1 and 1.1 mm of rain?



$$P(A_{[a,b]}) = \int_a^b p(x)dx.$$

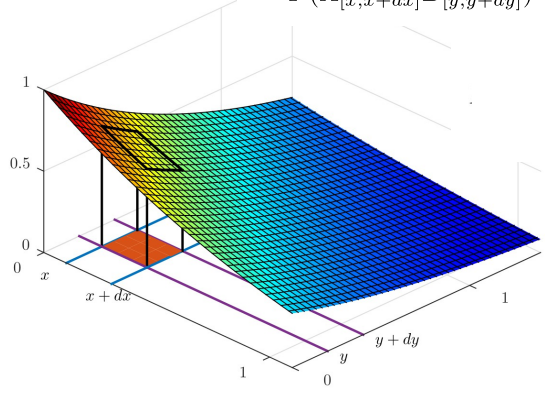
$$P(A_{[x,x+dx]}) \approx p(x)dx$$

$A_{[a,b]}$: *There will be between* a *and* b *mm of rain*

Probability vs. Density

$$P((x,y) \in D) = \int_{(x,y)\in D} p(x,y)dxdy$$

$$P(A_{[x,x+dx]}B_{[y,y+dy]}) =$$



This implies:

$$p(x,y) = p(y|x)p(x)$$

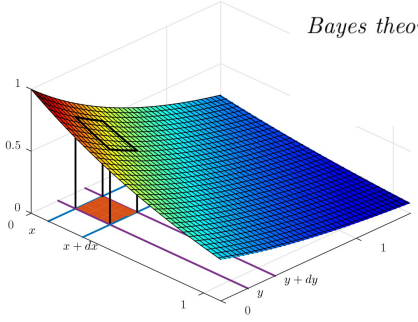
Probability vs. Density

The sum rule: $$\int dx\ p(x|z) = 1$$

The product rule: $$p(x, y|z) = p(y|x, z)p(x|z)$$

Bayes theorem: $$p(x|y, z) = \frac{p(y|x, z)p(x|z)}{p(y|z)}$$

$$= \frac{p(y|z)p(x|y, z)}{\int p(y|x', z)p(x'|z)dx'}.$$



Collecting all of this we obtain:

For **discrete** random variables and probability mass functions:

Marginalization	$\sum_c p(x = c, y \mid z)dx = p(y \mid z)$
Product rule	$p(x, y \mid z) = p(y \mid z, x)p(x \mid z)$
Bayes theorem	$p(x \mid y, z) = \frac{p(y \mid z, x)p(x \mid z)}{\sum_c p(x = c, y \mid z)p(x = c \mid z)}$

For **continuous** random variables and densities

Marginalization	$\int p(x, y \mid z)dx = p(y \mid z)$
Product rule	$p(x, y \mid z) = p(y \mid z, x)p(x \mid z)$
Bayes theorem	$p(x \mid y, z) = \frac{p(y \mid z, x)p(x \mid z)}{\int p(y \mid z, x')p(x' \mid z)dx'}$

Expected value

- **Discrete** random variables

$$\mathbb{E}[g] = \sum_i g(x_i)p(x_i)$$

- **Countinous** random variables

$$\mathbb{E}[g] = \int g(x)p(x)\,dx$$

Expected value

- Mean

$$\bar{x} = \mathbb{E}[x] = \int x\, p(x)\, dx$$

- Covarinace

$$\begin{aligned} cov(x, y) &= \mathbb{E}[(x - \bar{x})(y - \bar{y})] \\ &= \mathbb{E}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])] \end{aligned}$$

- Variance

$$var(x) = cov(x, x) = \mathbb{E}[(x - \bar{x})^2]$$

- Standard deviation

$$std(x) = \sqrt{var(x)}$$

Densities and models

- In machine learning, we want to learn a parameter from data
- Models of the data which use parameters are how we do that
- We build models out of simpler building blocks namely distributions and densities (see chapter 5 for descrete variables and chapter 6 for continuous variables). In this course we will u

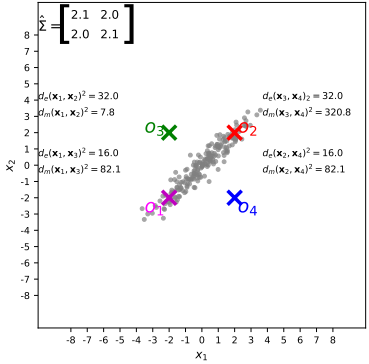
  - Bernoulli distribution
  - The Categorical distribution
  - **The Beta density**
  - **The Multivariate normal density**

Recall: Mahalanobis distance

Define a (squared) distance that takes into account variance and covariance such that the distance between the kth and lth data point is:

$$d_{euclidian}(\boldsymbol{x}_k, \boldsymbol{x}_l)^2 = (\boldsymbol{x}_k - \boldsymbol{x}_l)^\top \boldsymbol{I}^{-1} (\boldsymbol{x}_k - \boldsymbol{x}_l)$$

$$d_{mahanalobis}(\boldsymbol{x}_k, \boldsymbol{x}_l)^2 = (\boldsymbol{x}_k - \boldsymbol{x}_l)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_k - \boldsymbol{x}_l)$$



The multivariate normal distribution

A distribution for M-dimensional vectors $\boldsymbol{x}$:

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{M}{2}}\sqrt{|\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

$$M = 1: \quad \mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ is the covariance matrix:

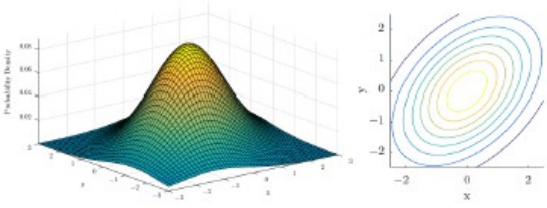
$$\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{x}], \quad \Sigma_{ij} = \mathrm{cov}[x_i, x_j]$$

where Σ_{ij} is the covariance between attribute (a random variable) i and attribute j.

Example: The 2D normal distribution

$$\boldsymbol{\mu} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathrm{var}(x_1) & \mathrm{cov}[x_1, x_2] \\ \mathrm{cov}[x_2, x_1] & \mathrm{var}(x_2) \end{bmatrix}$$

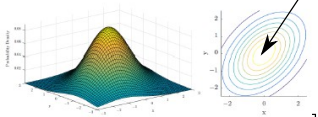


The multivariate normal distribution: Covariance

Quiz 2: Covariance

. Match the covariances to the contour plots

$$\mu = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix}$$



$$\Sigma = \begin{bmatrix} \text{var}(x_1) & \text{cov}(x_1, x_2) \\ \text{cov}(x_2, x_1) & \text{var}(x_2) \end{bmatrix}$$

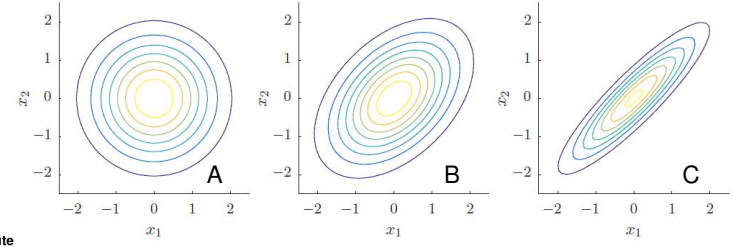
A. Covariance of A is $\mathbf{\Sigma}_A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$, $\mathbf{\Sigma}_C = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

B. $\mathbf{\Sigma}_B = \begin{bmatrix} 1 & -0.45 \\ 0.45 & 1 \end{bmatrix}$, $\mathbf{\Sigma}_C = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$

C. $\mathbf{\Sigma}_B = \begin{bmatrix} 10 & 4.5 \\ 4.5 & 10 \end{bmatrix}$, $\mathbf{\Sigma}_C = \begin{bmatrix} 10 & 9 \\ 9 & 10 \end{bmatrix}$

D. $\mathbf{\Sigma}_A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $\mathbf{\Sigma}_C = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$

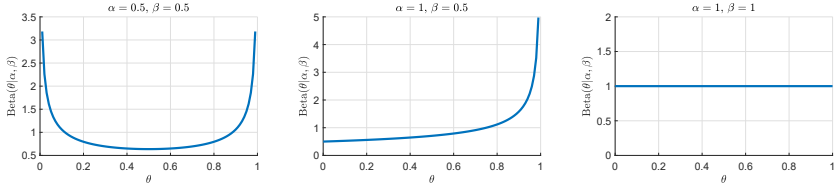
E. Don't know.



Beta distribution

Suppose θ is defined on the unit interval $\theta \in [0, 1]$

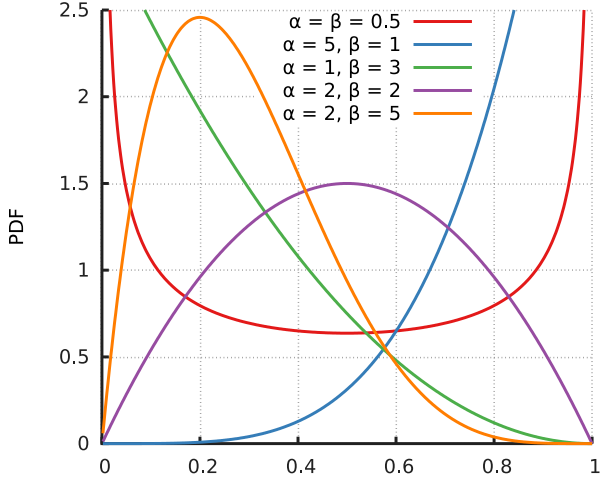
$$\text{Beta density: } p(\theta \mid \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$



$\alpha, \beta > 0$ are releated to the mean and variance

$$\mathbb{E}_{p(\theta|\alpha,\beta)}[\theta] = \frac{\alpha}{\alpha + \beta} \qquad \text{Var}_{p(\theta|\alpha,\beta)}[\theta] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

Beta distribution



Picture from Wikipedia

Learning principles

Machine learning as statistical inference

Probabilities, densities and learning

- Your friend just got a dog.
- The dog can either be in the doghouse or outside
- The first four times you come by the dog is in the doghouse
- What is the chance the dog is in the doghouse tomorrow?

- Your friend buys a coin.
- The coin can either come up heads or tails
- The first four times you flip the coin it comes up tails
- What is the chance the coin comes up tails in the next flip?

Intuition tells us the answers are different, but the situation seems similar...

Recall: The Bernoulli distribution

- Suppose a coin come up heads with probability θ
  - Suppose $b = 1$ is the event the coin land heads
  - $b = 0$ is the event the coin land tails
- The density is given by the **Bernoulli distribution**

$$p(b|\theta) = \theta^b(1-\theta)^{1-b}$$

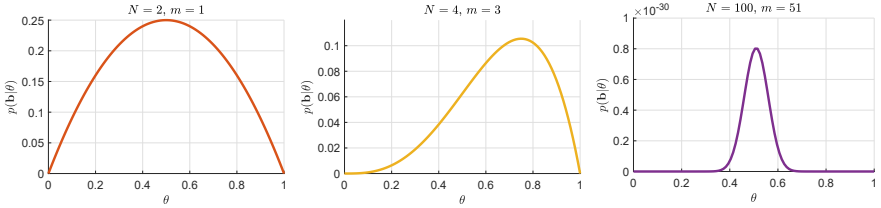
- For a sequence of N flips $b_1, b_2, \ldots, b_N$

$$\begin{aligned} p(b_1, \ldots, b_N|\theta) &= \prod_{i=1}^{N} p(b_i|\theta) \\ &= \theta^{\sum_{i=1}^N b_i}(1-\theta)^{N-\sum_{i=1}^N b_i} \\ &= \theta^m(1-\theta)^{N-m}, \quad m = b_1 + b_2 + \cdots + b_N \end{aligned}$$

Conditional Independence

- **What is θ?**

Recall: The Bernoulli distribution, maximum likelihood



$$p(b_1, \ldots, b_N|\theta) = \theta^m(1-\theta)^{N-m}$$

An idea for selecting θ^* is **Maximum likelihood**

$$\theta^* = \arg\max_\theta p(b_1, \ldots, b_N|\theta)$$

The value of θ according to which the data is most plausible

The Bernoulli distribution

- A magic coin is a coin that comes up heads with probability θ
  - Suppose $b = 1$ is the event the coin land heads
  - $b = 0$ is the event the coin land tails
- For a sequence of N flips $b_1, b_2, \ldots, b_N$

$$p(b_1, \ldots, b_N) = \theta^m (1-\theta)^{N-m}, \quad m = b_1 + b_2 + \cdots + b_N$$

- **What is θ?** Answer: **Use Bayes' Theorem**!

$$p(\theta|\boldsymbol{b}) =$$

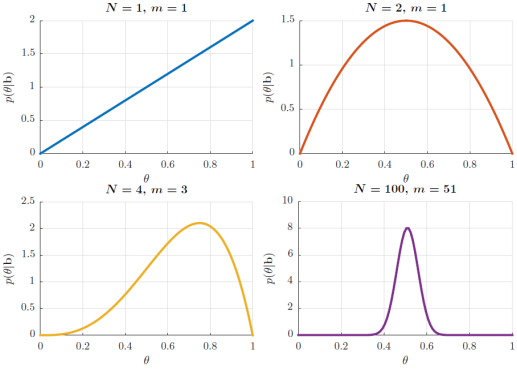
- Assume $p(\theta) =$

$$p(\theta|\boldsymbol{b}, \alpha, \beta) =$$

$$= \frac{\Gamma(\alpha+\beta+N)}{\Gamma(\alpha+m)\Gamma(\beta+N-m)} \theta^{\alpha+m-1}(1-\theta)^{\beta+N-m-1}$$

Example: $\alpha = \beta = 1$

$$
\begin{aligned}
p(\theta|\boldsymbol{b}, \alpha, \beta) &= \frac{\Gamma(\alpha+\beta+N)}{\Gamma(\alpha+m)\Gamma(\beta+N-m)}\theta^{\alpha+m-1}(1-\theta)^{\beta+N-m-1} \\
&= \frac{(N+1)!}{m!(N-m)!}\theta^{m}(1-\theta)^{N-m}
\end{aligned}
$$



Dogs and coins

- Your friend just got a dog.
- The dog can either be in the doghouse or outside
- The first four times you come by the dog is in the doghouse
- What is the chance the dog is in the doghouse tomorrow?

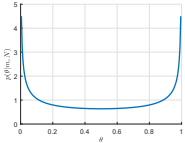
- Your friend buys a coin.
- The coin can either come up heads or tails
- The first four times you flip the coin it comes up tails
- What is the chance the coin comes up tails in the next flip?

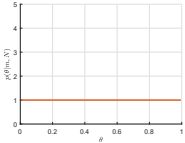
Dog: $\alpha = \beta = \frac{1}{2}$

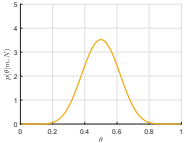
Coin: $\alpha = \beta = 10$

Prior

$p(\theta \mid \alpha, \beta) =$





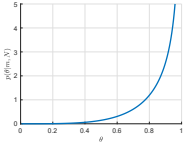


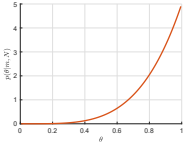
Likelihood

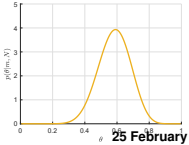
$$p(m = 4, N = 4 \mid \theta) = \theta^m (1-\theta)^{N-m} = \theta^4 (1-\theta)^{4-4} = \theta^4$$

Posterior

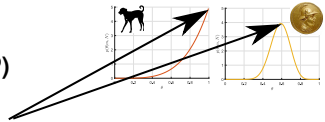
$$p(\theta \mid m = 4, N = 4) = \frac{p(N, m \mid \theta) p(\theta \mid \alpha, \beta)}{p(N, m)}$$







Learning principle: Maximum a posteriori (MAP)



- **Another idea**: Select θ which is "most probably"

$$\theta^* = \arg\max_\theta p(\theta|M,N) = \arg\max_\theta \left[\frac{p(m,N|\theta)p(\theta|\alpha,\beta)}{p(M,N)}\right]$$

- Use that $\arg\max_x f(x) = \arg\min_x [-\log f(x)]$ if $f(x) > 0$:

$$\theta^* = \arg\min_\theta \left[-\log \frac{p(m,N|\theta)p(\theta|\alpha,\beta)}{p(m,N)}\right]$$

(likelihood)
$p(m,N|\theta) = \theta^m(1-\theta)^{N-m}$

(prior)
$p(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$

- All in all:

$$\theta^* = \arg\min E(\theta), \quad E(\theta) = -\log p(m,N|\theta) - \log p(\theta|\alpha,\beta)$$

Maximum a posteriori (MAP) learning

- Consider some data $\boldsymbol{X} = \boldsymbol{x}_1, \ldots, x_N$ and $\boldsymbol{y} = y_1, \ldots, y_N$
- Suppose we think $\boldsymbol{x}_i$ relates to y_i by some parameters $\boldsymbol{\theta}$
- **Assume**

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) = \prod_{i=1}^{N} p(y_i|\boldsymbol{x}_i, \boldsymbol{w}), \quad p(\boldsymbol{w}|\boldsymbol{X}) = p(\boldsymbol{w})$$

Observations are not informative about each other when we know parameters

Without y, we cannot learn the parameters

- Then

$$p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{X})}{p(\boldsymbol{y}|\boldsymbol{X})}$$

- The following are equivalent:

Maximize: $\boldsymbol{w}^* = \arg\max_{\boldsymbol{w}} p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y})$

Minimize: $\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} E(\boldsymbol{w}), \quad E(\boldsymbol{w}) = \left[\frac{1}{N}\sum_{i=1}^{N} -\log p(y_i|\boldsymbol{x}_i, \boldsymbol{w})\right] - \frac{1}{N}\log p(\boldsymbol{w})$

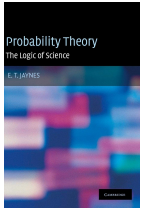
- All we need is a likelihood (**usually pretty simple**) and a prior (**can be omitted**) and we have a machine-learning method.
  - **Pro**: Easy, conceptually simple, efficient
  - **Con**: Can sometimes give spurious results (overfit)

Summary

Summary

- Discrete random variables and related distributions
  - Bernoulli distribution
  - Later: Categorical distribution
- Continuous random variables and related distributions
  - Multivariate normal distribution (akk multivariate Gaussians)
  - Beta distribution
  - Later: Mixture of multivaite normal distributions (aka Gaussian mixture model)
- Learning principles via statistical inference
  - Maximum likelihood
  - Bayesian inference (full posterior distribution)
  - Maximum a posteriori

Resources



Probability: The logic of science Classical textbook which treats probabilities as states-of-knowledge and discuss many practical and philosophical issues

(https://bayes.wustl.edu/etj/prob/book.pdf)

https://www.khanacademy.org An excellent introduction to probability theory which we recommend as a go-to resource

(https://www.khanacademy.org/math/statistics-probability/probability-library)

http://citeseerx.ist.psu.edu A more in-depth discussion of Bayes in the court room" (http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=EFE0328140036A4C668BB5B9FC76C9BE?doi=10.1.1.599.8675&rep=rep1&type=pdf)