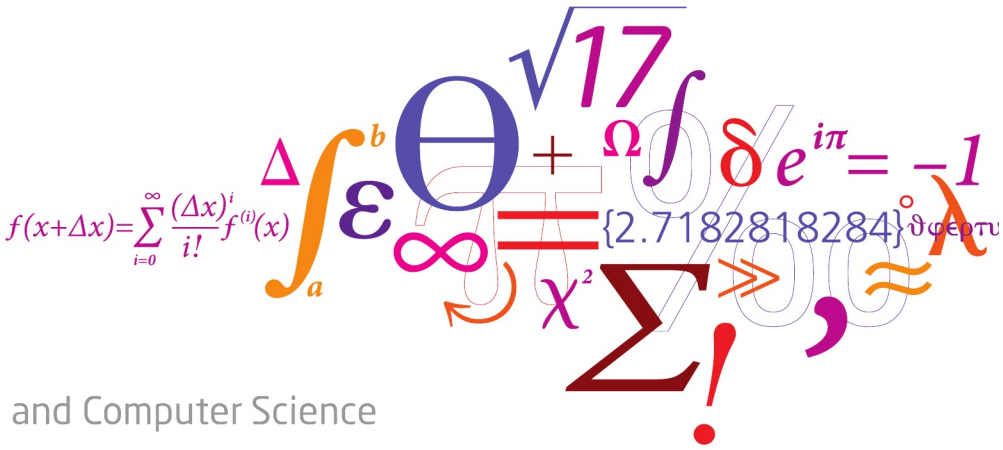
02450: Introduction to Machine Learning and Data Mining

Probability densities and data Visualization

Tue Herlau

DTU Compute, Technical University of Denmark (DTU)



DTU Compute
Department of Applied Mathematics and Computer Science

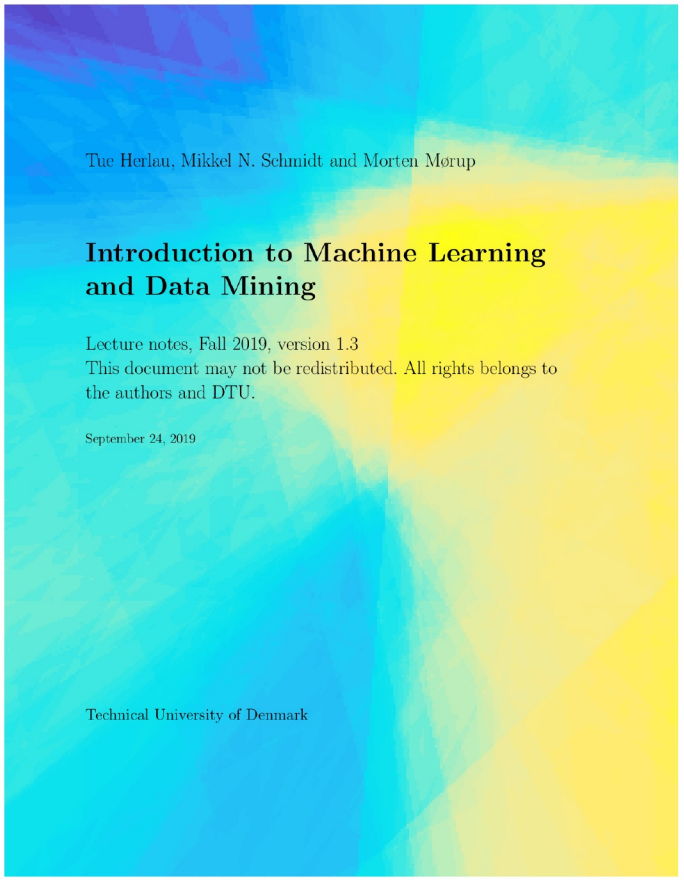
Today

Feedback Groups of the day:

Fabian Felix Bertoldo, Jørgen Taule, Raül Pérez i Gonzalo, Michelle Anker Pihl, Lene Christiansen, Mie Guldhammer Andersen, Nikolaj Geertinger, William Peter Settnes, Wictor Lang Jensen, Hadeel Moustafa, Abdulstar Shihada Kousa, Nojus Mickus, Andreas Goltermann, Jakob Malte Skou Lindstad, Christian Rømer Thulstrup, Niklaes Dino Robbin Jacobsen, Sebastian Stokkebro Sørensen, Kilian Waldemar Conde Frieboes, Thomas Rudolph Sparre Conrad, Asger Conradsen, Lukas Fredrik Cronje, Oskar Steentoft Dahlberg, Matteo D'Andrea, Benjamin Danielsen, Nichlas Davidsen, Seiran Davoudi, Sander de Boer, Matilde Maria de Place, Agathe De Vulpian, Sebastian Nymann Deleuran, Blanca Robledo Diaz, Rebecca Dürmüller, Emma Cathrine Rud Egelund, Pil Skov Eghoff, Lawrence Kofi Asane Egyir, Yosef Khodr El-Fil, Niels Kjær Ersbøll, Mahsa Eskandarzadeh, Kimmo Lilholt Bang Feldbæk, Þórður Páll Fjalarsson, Fredrik Hole Flaa, Maria Mandrup Fogh, Anton Palm Folkmann, Martin Trangbæk Forsingdal, Edvard Foss, Aslak Rønnow Franzen, Andreas Mørk Frendorf, Ditlev Reimer Frickmann, Tina Funck, Annika Lund Gade, Ruairi James Gallagher, Emilie Dybdahl Gandrup, Imogen Garner

Reading material:

Chapter 6, Chapter 7



Lecture Schedule

1. Introduction
   3 September: C1

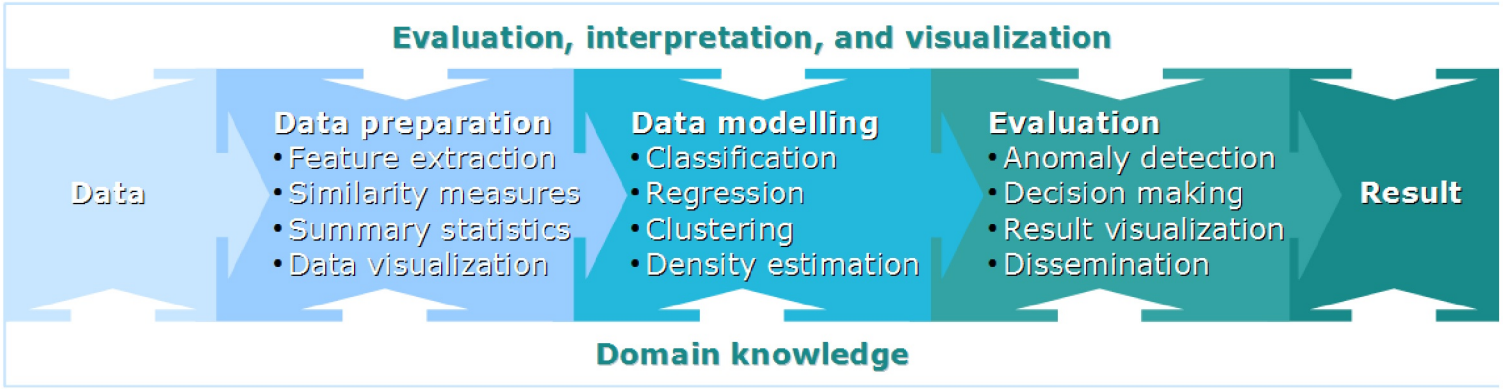
   Data: Feature extraction, and visualization
2. Data, feature extraction and PCA
   10 September: C2, C3
3. Measures of similarity, summary statistics and probabilities
   17 September: C4, C5
4. **Probability densities and data Visualization**
   24 September: C6, C7

   Supervised learning: Classification and regression
5. Decision trees and linear regression
   1 October: C8, C9 **(Project 1 due before 13:00)**
6. Overfitting, cross-validation and Nearest Neighbor
   8 October: C10, C12
7. Performance evaluation, Bayes, and Naive Bayes
   22 October: C11, C13
8. Artificial Neural Networks and Bias/Variance
   29 October: C14, C15
9. AUC and ensemble methods
   5 November: C16, C17

   Unsupervised learning: Clustering and density estimation
10. K-means and hierarchical clustering
    12 November: C18 **(Project 2 due before 13:00)**
11. Mixture models and density estimation
    19 November: C19, C20
12. Association mining
    26 November: C21

    Recap
13. Recap and discussion of the exam
    3 December: C1-C21 **(Project 3 due before 13:00)**

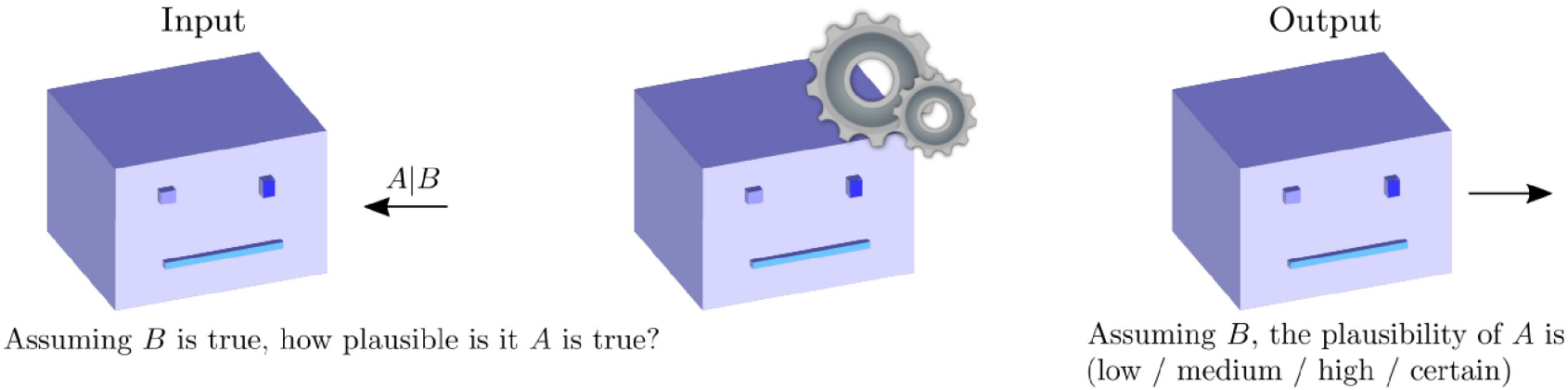
Piazza online help: `https://piazza.com/dtu.dk/fall2019/02450`
Videos/streaming of lectures: `https://video.dtu.dk`



Learning Objectives

- Understand probability densities and related concepts
- Derive cost-functions from likelihood functions using Bayes' theorem
- Be able to understand and apply a wide range of data visualization approaches
- Understand good practice in plotting including Tufte's guidelines

Probabilities recap



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

$$P(A|B) = x \in [0,1]$$

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 recap

- Probabilities $P(A|B)$ obey

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

Product rule $P(AB|C) = P(B|AC)P(A|C)$

$A_i A_j = 0$ $\quad$ $A_1 + A_2 + \cdots + A_n = 1$

- If we consider **mutually exclusive** and **exhaustive** events $A_1, \ldots, A_n$:

Bayes' theorem

$$P(A_i|BC) = \frac{P(B|A_iC)P(A_i|C)}{\sum_j P(B|A_jC)P(A_j|C)}$$

Stochastic variables

Suppose X, Y, etc. are quantities we measure. $X = x$ is a binary event. For instance

- $N = 1, 2, \ldots$: Number of children
- $T = 1, \ldots, 24$: Hour of day
- $C = 1, \ldots, K$ favorite cereal

Bayes theorem is then:

$$P(Y = y|X = x, Z = z) = \frac{P(X = x|Y = y, Z = z)P(Y = y|Z = z)}{\sum_{y'} P(X = x|Y = y', Z = z)P(Y = y'|Z = z)}$$

Using $p(x|y, z) = P(X = x|Y = y, Z = z)$ and ignoring z this becomes:

$$p(y|x, z) = \frac{p(x|y)p(y)}{\sum_{y'} p(x|y')p(y')}$$

Quiz 1, Probabilities (Spring 2014)

Consider the Wholesale 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?

D

A. 7.7 %

B. 15.4 %

C. 44.1 %

D. 59.6 %

E. Don't know.

$R = L, G_p, T$ (1, 2, 3)

$$p(R=L \mid D) = \frac{p(D \mid R=L)\,p(R=L)}{\sum_{r=1}^{3} p(D \mid R=r)\,p(R=r)} = \frac{(44.1\%) \times (17.5\%)}{\text{---}}$$

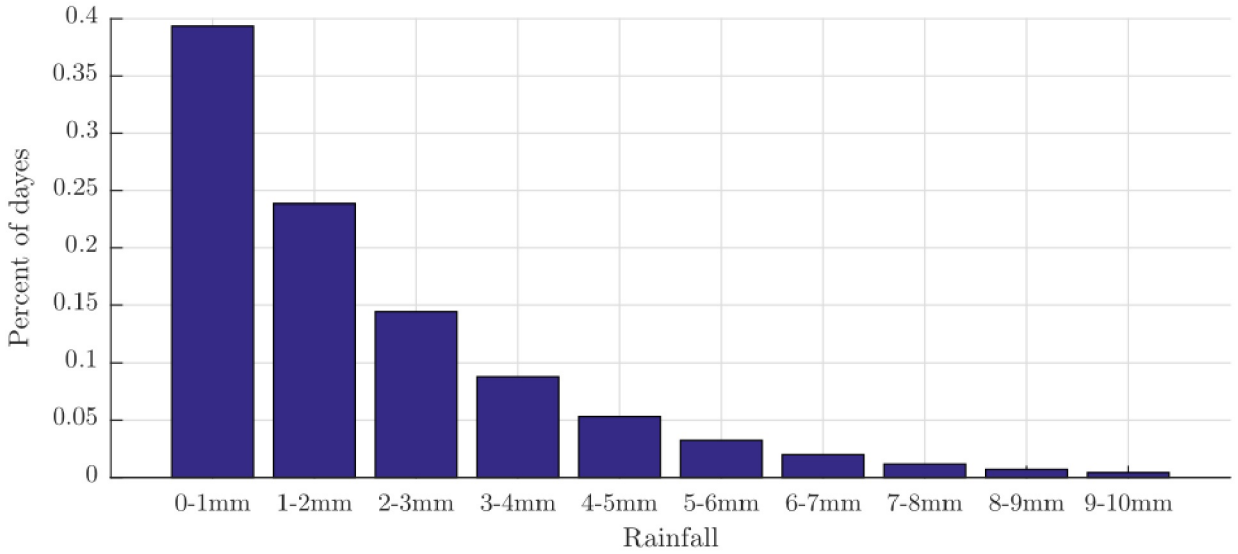
$$= \frac{0.441 \times .175}{0.441 \times .175 + 0.489 \times .107 + 0.516 \times 0.718} = 0.1544$$

We will let $LISBON$, $OPORTO$ and $OTHER$ denote coming from Lisbon, Oporto and the other region respectively. $DELI_H$ will denote above the median value of delicatessen consumption. We now find using Bayes theorem:

$$\begin{aligned} P(LISBON|DELI_H) &= \frac{P(DELI_H|LISBON)P(LISBON)}{P(DELI_H)} \\ &= \frac{P(DELI_H|LISBON)P(LISBON)}{\begin{array}{c}P(DELI_H|LISBON)P(LISBON) \\ +P(DELI_H|OPORTO)P(OPORTO) \\ +P(DELI_H|OTHER)P(OTHER)\end{array}} \\ &= \frac{44.1 \cdot 0.175}{0.441 \cdot 0.175 + 0.489 \cdot 0.107 + 0.516 \cdot 0.718} \\ &= 0.1544 \approx 15.4\% \end{aligned}$$

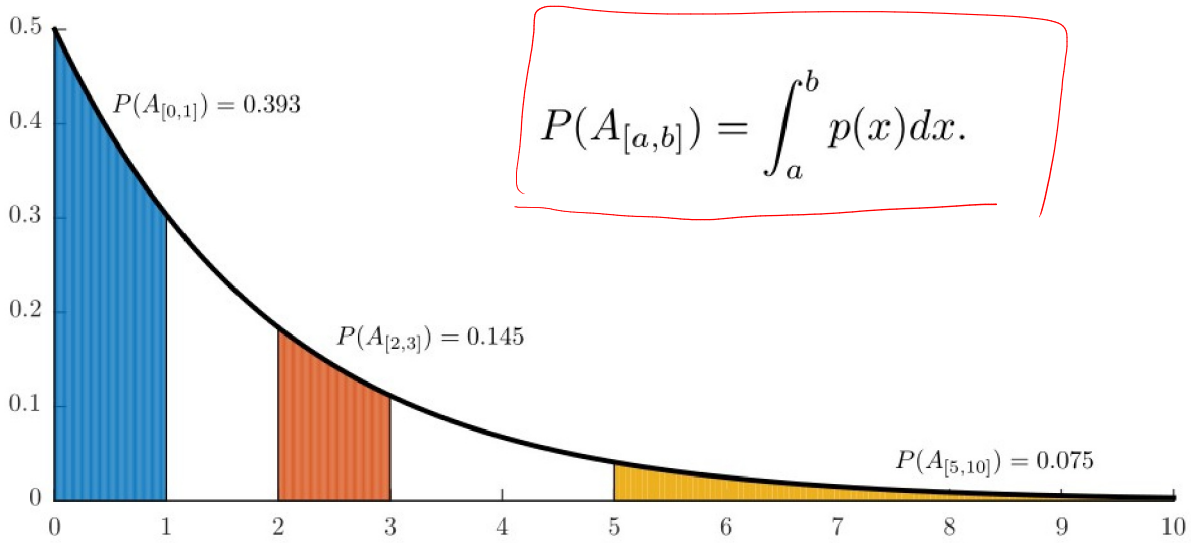
Probability vs. Density

- Suppose we consider the rainfall on an average day r $\in [0, \infty[$
- **Can't** talk about the probability there will be **exactly** r=2.3 mm of rain, P(r=2.3mm)
- **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**

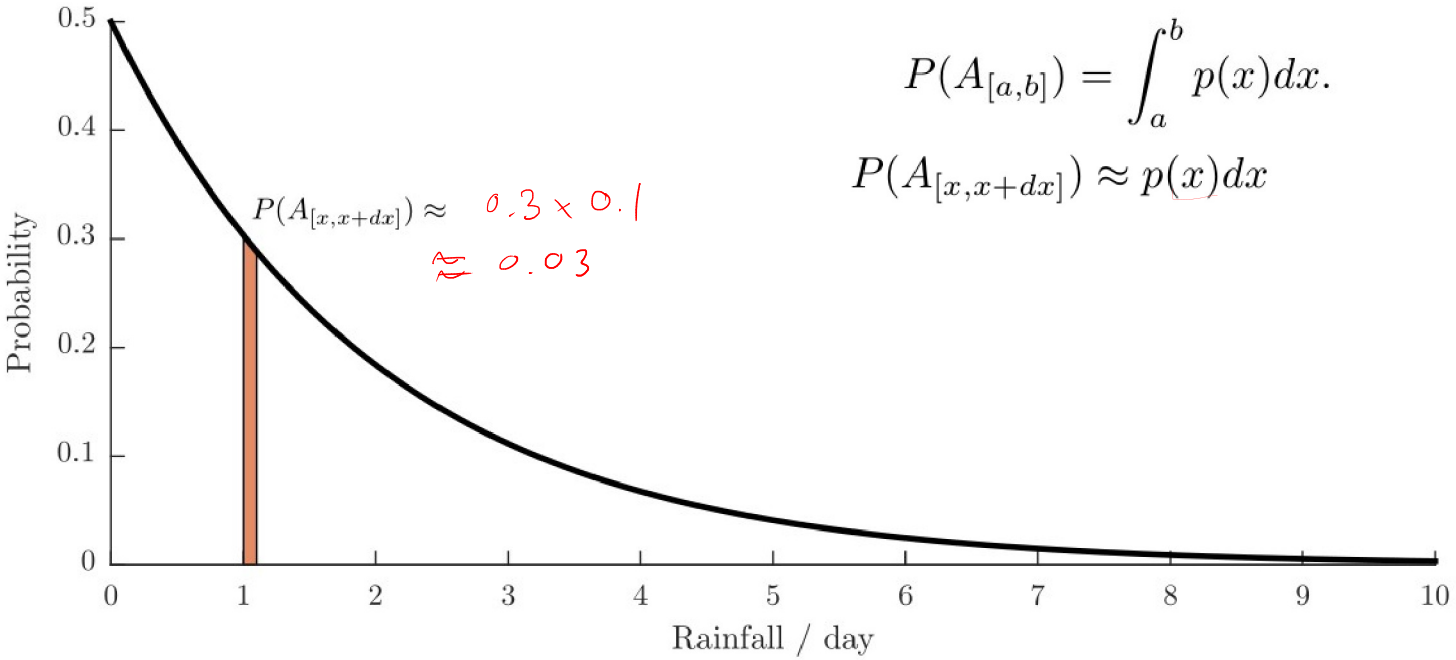


$$P(A_{[a,b]}) = \int_a^b p(x)dx.$$

$A_{[a,b]}$: *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?



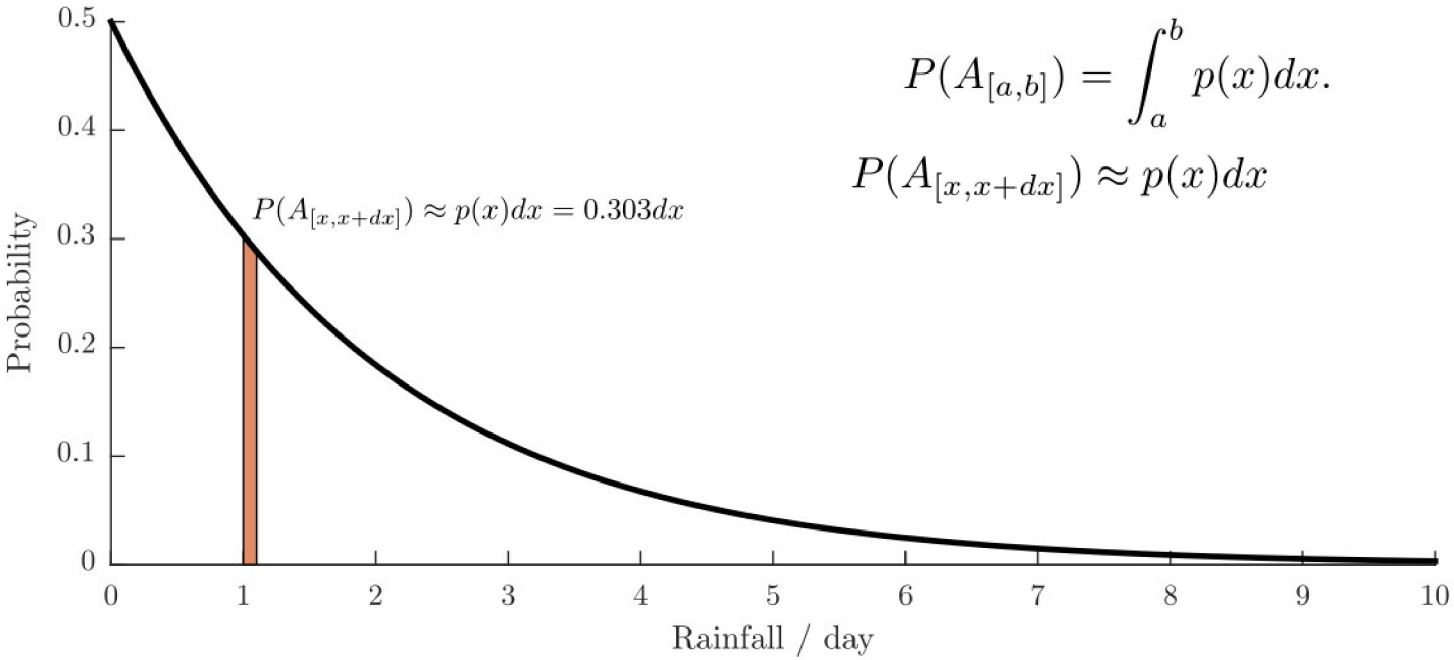
$$P(A_{[a,b]}) = \int_a^b p(x)dx.$$

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

$A_{[a,b]}$: *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

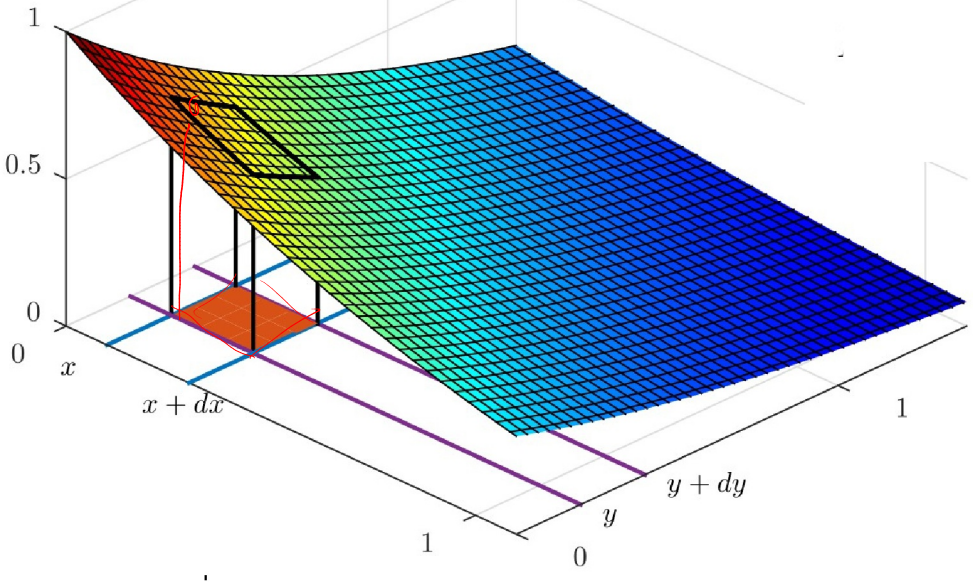
- For two variables x and y, the **probability** is an integral over an **area**

$$P(A_{[x,x+dx]}) = p(x)dx.$$

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

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

$$= (dy \times p(y \mid x))\,(p(x)\,dx)$$



This implies:

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

$$\int p(x,y)\,dy = p(x)$$

Probability vs. Density

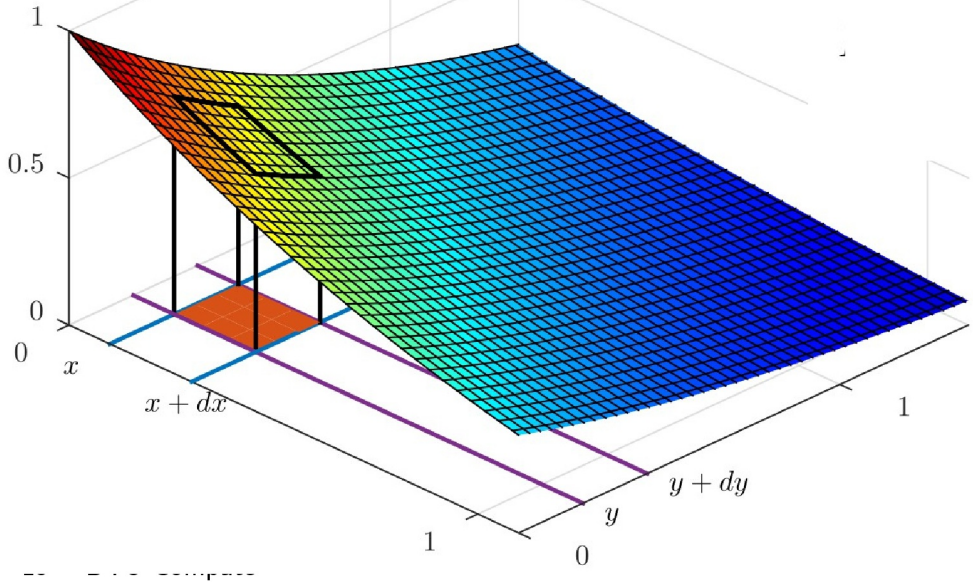
- For two variables x and y, the **probability** is an integral over an **area**

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

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

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

$$p(x,y)dxdy = p(x|y)dxp(y)dy$$



This implies:

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

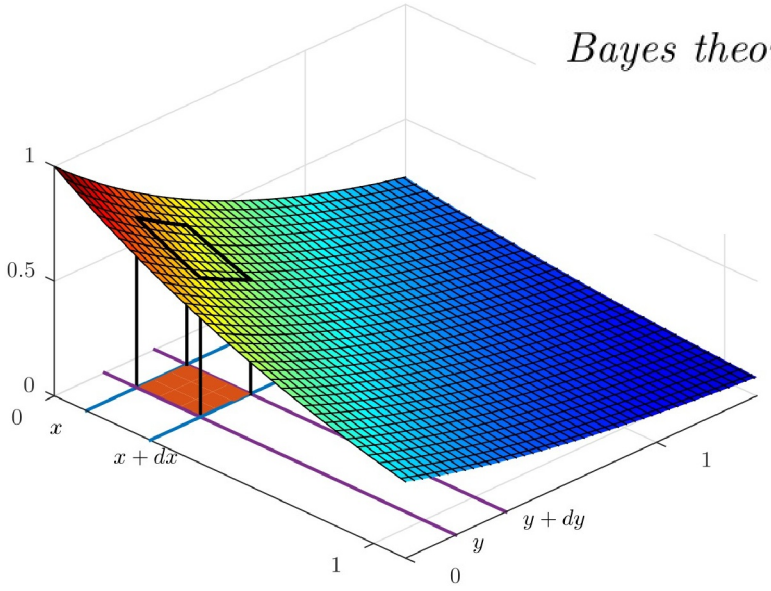
Probability vs. Density

- Thus, we have shown the rules of probability theory also holds for densities

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

$$\textit{The product rule:} \qquad p(x, y|z) = p(y|x, z)p(x|z)$$

$$\textit{Bayes theorem:} \qquad p(x|y, z) = \frac{p(y|x, z)p(x|z)}{p(y|z)} = \frac{p(y|z)p(x|y, z)}{\int p(y|x', z)p(x'|z)dx'}.$$



Collecting all of this we obtain:

- Rules of probability for densities

$$\begin{aligned} \textit{Marginalization:} \quad & \int p(x,y|z)dx = p(y|z) \\ \textit{The product rule:} \quad & p(x,y|z) = p(y|x,z)p(x|z) \\ \textit{Bayes theorem:} \quad & p(x|y,z) = \frac{p(y|x,z)p(x|z)}{\int p(y|x',z)p(x'|z)dx'}. \end{aligned}$$

- Rules of probability for discrete variables

$$\begin{aligned} \textit{Marginalization:} \quad & \sum_c p(x=c,y|z) = p(y|z) \\ \textit{The product rule:} \quad & p(x,y|z) = p(y|x,z)p(x|z) \\ \textit{Bayes theorem:} \quad & p(x|y,z) = \frac{p(y|x,z)p(x|z)}{\sum_c p(y|x=c,z)p(x=c|z)}. \end{aligned}$$

$\int p(x,y|z)dx = p(x|z)$

Expected values

- Discrete random variable

$$\mathbb{E}\left[g\right] = \sum_i g(x_i) P(x_i)$$

- Continuous random variable

$$\mathbb{E}\left[g\right] = \int_{-\infty}^{\infty} g(x) p(x) \mathrm{d}x$$

Statistics

- **Mean**

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

- **Covariance**

$$\mathrm{cov}(x, y) = \mathbb{E}[(x - \bar{x})(y - \bar{y})]$$

- **Variance**

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

- **Standard deviation**

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

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 our of simpler building blocks (densities).
  In this course we will learn four:

Bernouilli density

The Catagorical density

The Beta density

The Multiviate normal density

$p(x) = ?$

The multivariate normal distribution

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

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

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

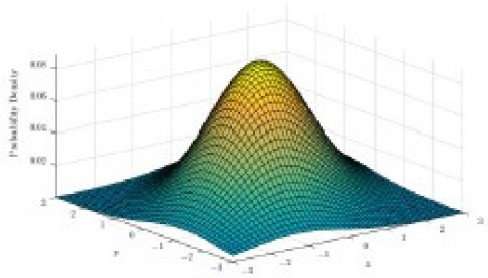
$\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ is the covariance matrix:

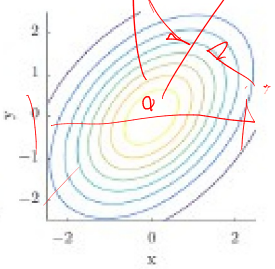
$$\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{x}], \quad \Sigma_{ij} = \mathrm{cov}[x_i, x_j]$$

- Example: 2-dimensional Normal distribution

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

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

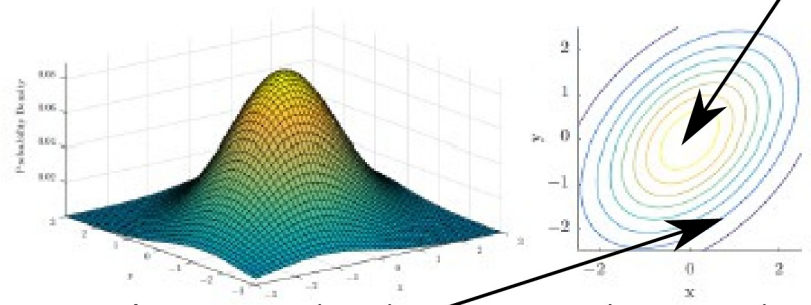




Quiz 2, Covariance

. Match the covariances to the contour plots

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



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

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

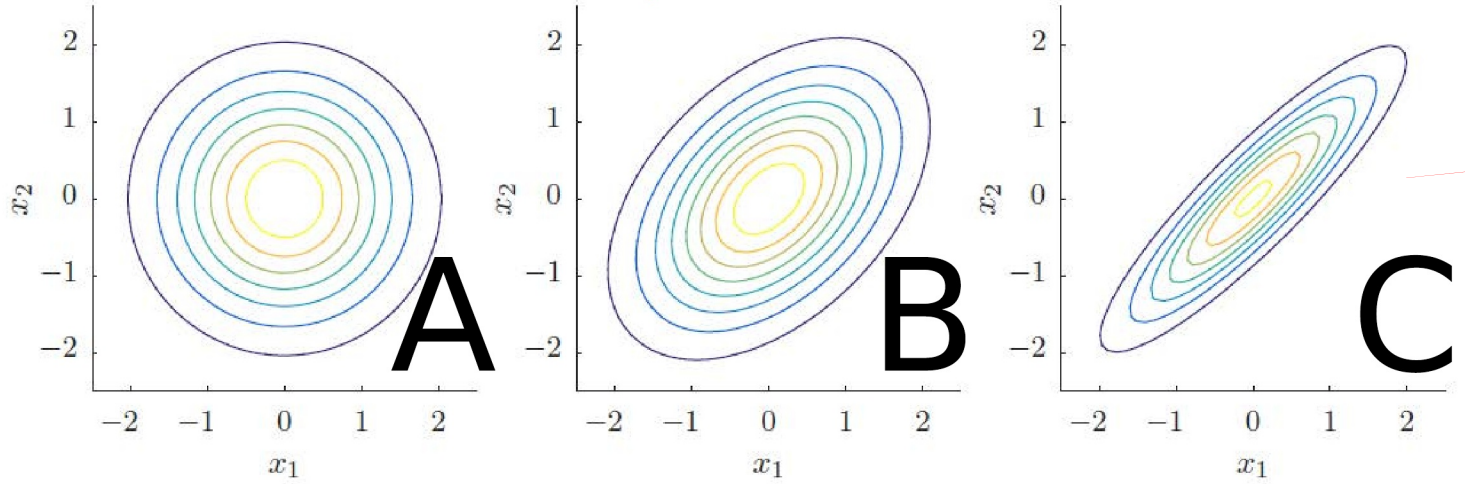
B. $\mathbf{\Sigma}_B = \begin{bmatrix} 1 & -0.45 \\ 0.45 & 1 \end{bmatrix}$, $\mathbf{\Sigma}_C = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$

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.

Check out the online demo http://www2.imm.dtu.dk/courses/02450/DemoNormal.html



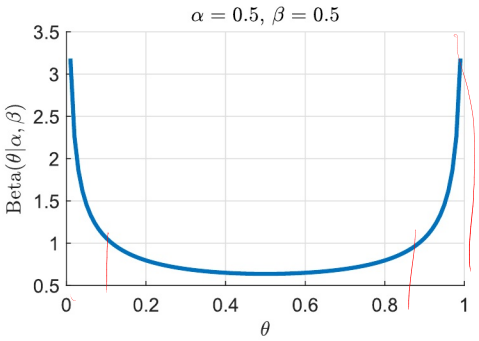
The right answer is D. The covariance has to be positive (because x_1 and x_2 are positively correlated), and the variance is 1 in all cases. Furthermore, since A is axis-aligned, the covariance terms are zero. All in all

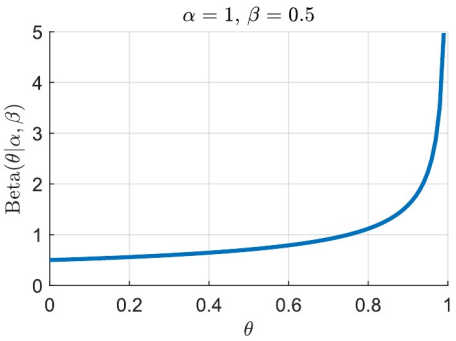
$$\boldsymbol{\Sigma}_A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{\Sigma}_A = \begin{bmatrix} 1 & 0.45 \\ 0.45 & 1 \end{bmatrix}, \quad \boldsymbol{\Sigma}_C = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$$

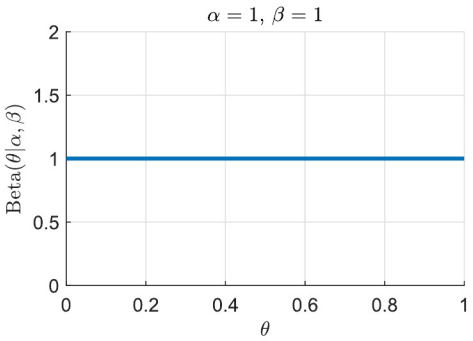
Beta distribution

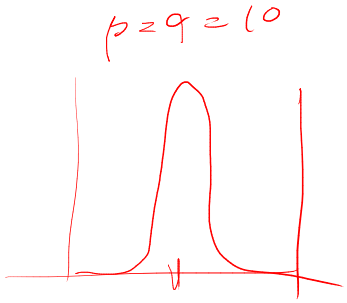
Suppose θ is defined on the unit interval $[0, 1]$

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





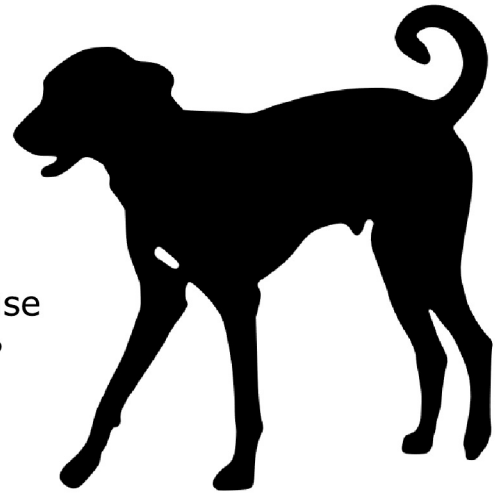




$\alpha, \beta > 0$ are related to the variance and mean

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

Probabilities 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 from last week: 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 $b_1, b_2, \ldots, b_N$ **Independence**

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

- **What is θ?**

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 $b_1, b_2, \ldots, b_N$

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

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

$$p(\theta|\boldsymbol{b}) = \frac{p(\boldsymbol{b}|\theta)\,p(\theta)}{\int d\theta'\; p(\boldsymbol{b}|\theta')\,p(\theta')}$$

- Assume $p(\theta) = \mathrm{Beta}(\theta|\alpha, \beta)$

$$p(\theta|\boldsymbol{b}, \alpha, \beta) = \frac{\theta^m (1-\theta)^{N-m} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int \theta^m (1-\theta)^{N-m} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta}$$

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

wiki

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**!

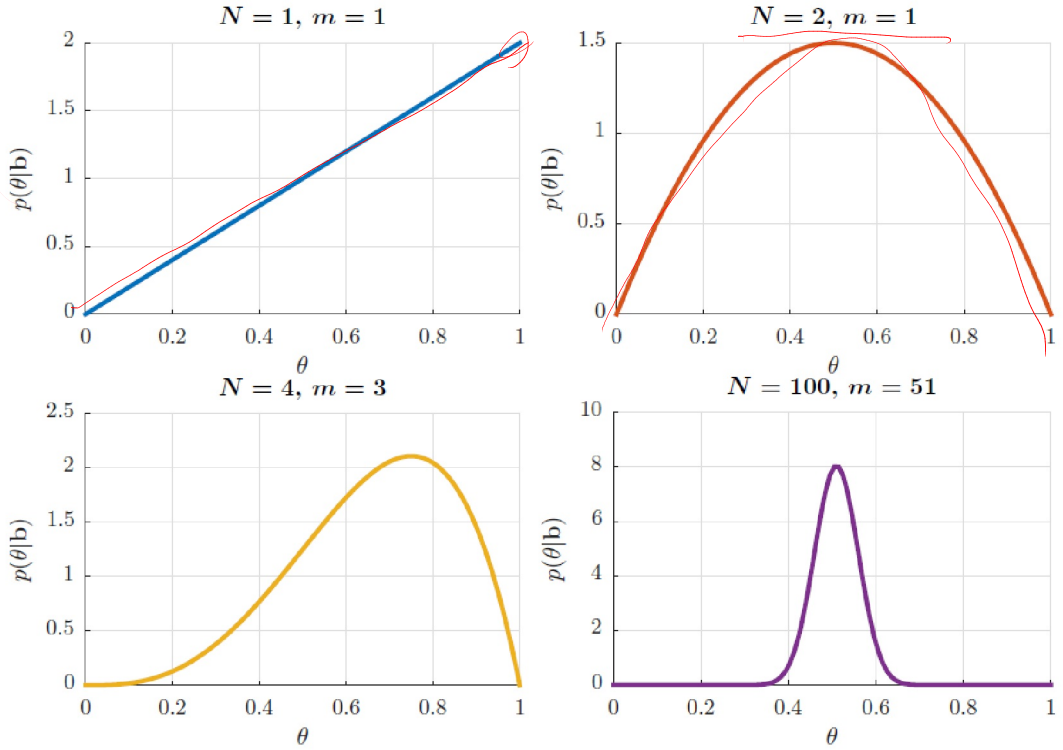
$$p(\theta|\boldsymbol{b}) = \frac{p(\boldsymbol{b}|\theta)p(\theta)}{p(\boldsymbol{b})} = \frac{p(\boldsymbol{b}|\theta)p(\theta)}{\int_0^1 p(\boldsymbol{b}|\theta')p(\theta')d\theta'}$$

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

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

Example: $\alpha = \beta = 1$

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



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?

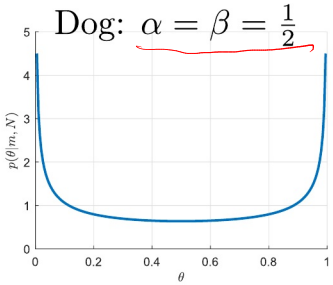
Dogs and coins

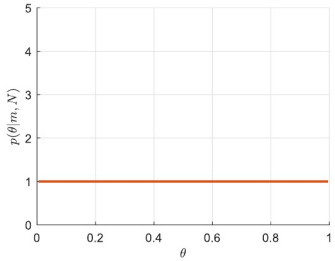
- Your friend just got a dog.
- The dog can either be in the doghouse or outside
- The first four times you come by the dog is in the doghouse
- What is the chance the dog is in the doghouse tomorrow?

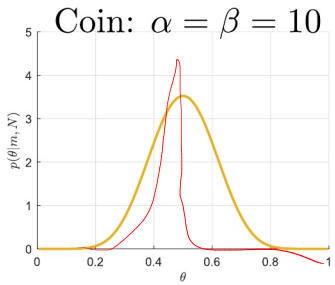
- Your friend buys a coin.
- The coin can either come up heads or tails
- The first four times you flip the coin it comes up tails
- What is the chance the coin comes up tails in the next flip?

Prior

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







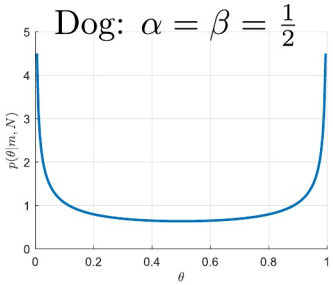
Dogs and coins

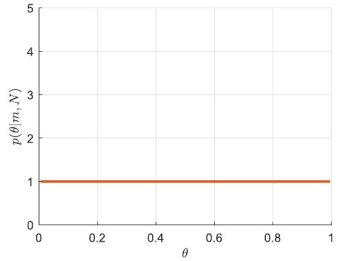
- Your friend just got a dog.
- The dog can either be in the doghouse or outside
- The first four times you come by the dog is in the doghouse
- What is the chance the dog is in the doghouse tomorrow?

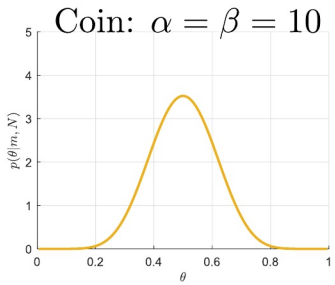
- Your friend buys a coin.
- The coin can either come up heads or tails
- The first four times you flip the coin it comes up tails
- What is the chance the coin comes up tails in the next flip?

Prior

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







Likelihood

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

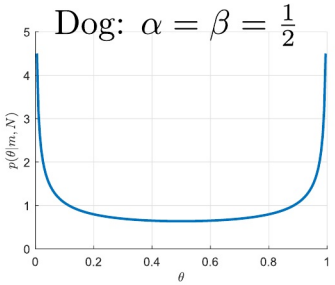
Dogs and coins

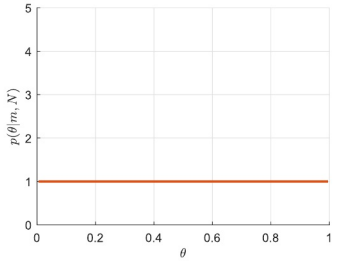
- Your friend just got a dog.
- The dog can either be in the doghouse or outside
- The first four times you come by the dog is in the doghouse
- What is the chance the dog is in the doghouse tomorrow?

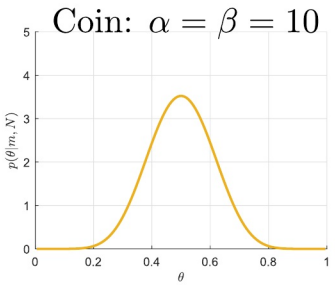
- Your friend buys a coin.
- The coin can either come up heads or tails
- The first four times you flip the coin it comes up tails
- What is the chance the coin comes up tails in the next flip?

Prior

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







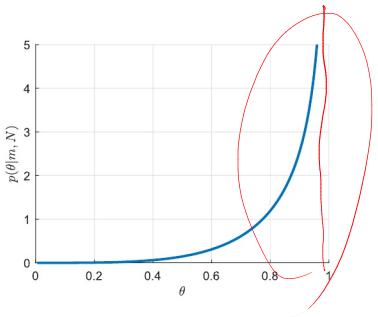
Likelihood

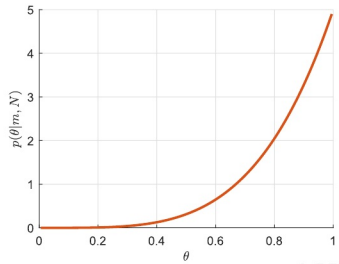
$$p(m=4, N=4|\theta) = \theta^m(1-\theta)^{N-m} = \theta^4$$

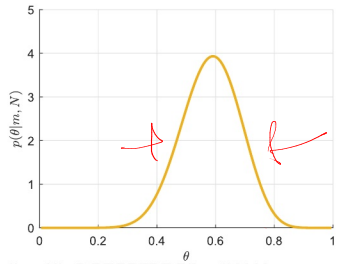
The difference between the two cases is that we have prior knowledge which tell us most coins are fair, and this affects our conclusions.
In most practical situations, we should assume as little as possible and choose $\alpha = \beta = \frac{1}{2}$

Posterior

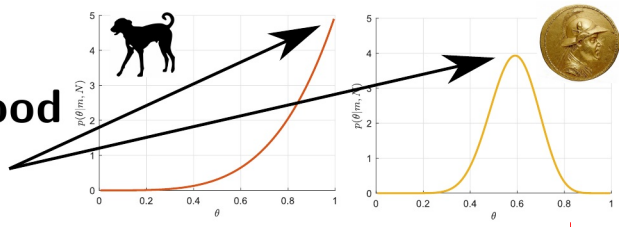
$$p(\theta|m,N) = \frac{p(m,N|\theta)p(\theta|\alpha,\beta)}{p(m,N)} =$$







Learning principle ii: Maximum likelihood



- 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:

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

A learning principle: Maximum likelihood

- 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 $f(x) > 0$:

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

- All in all:

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

Maximum likelihood 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 $\boldsymbol{w}$

Observations are not informative about each other when we know parameters

- **Assume**

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

Without y, we cannot learn the parameters

- Then

$$p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{X})\, p(\boldsymbol{w}|\boldsymbol{X})}{\int d\boldsymbol{w}\; p(\boldsymbol{w}|\boldsymbol{X})\, p(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{X})} = \frac{\left(\prod_{i=1}^{N} p(y_i|x_i, \boldsymbol{w})\right) p(\boldsymbol{w})}{\int \left(\prod_{i=1}^{N} p(y_i|x_i, \boldsymbol{w})\right) p(\boldsymbol{w})\, d\boldsymbol{w}}$$

- The following are equivalent:

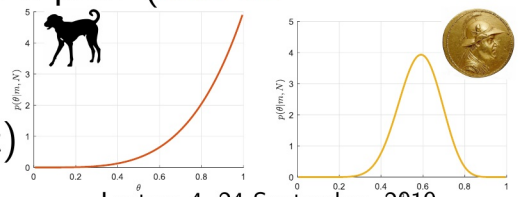
Maximum likelihood principle

Maximize: $\boldsymbol{w}^* = \arg\max_{\boldsymbol{w}} p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y}) = \arg\max_{\boldsymbol{w}} \left(\log \left(\prod_{i=1}^{N} p(y_i|x_i, \boldsymbol{w})\, p(\boldsymbol{w}) \right) \right)$

Minimize: $\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} E(\boldsymbol{w})$

$$E(\boldsymbol{w}) = \left(\sum_{i=1}^{N} -\log p(y_i|x_i, \boldsymbol{w}) \right) - \log p(\boldsymbol{w})$$

- 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)



Maximum likelihood 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**

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

- Then

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

- The following are equivalent:

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

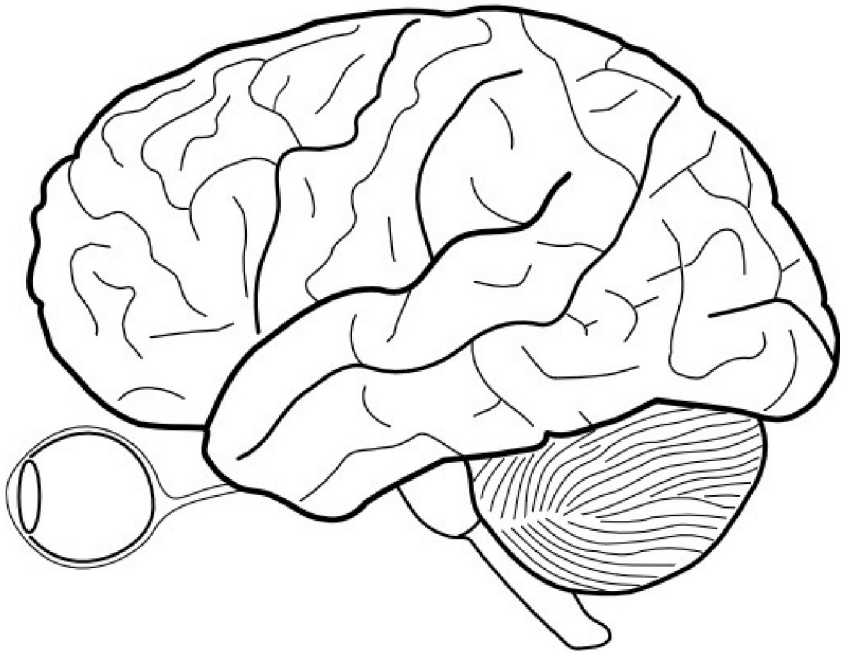
$$\textit{Minimize:} \quad \boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} E(\boldsymbol{w})$$

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

Visualization

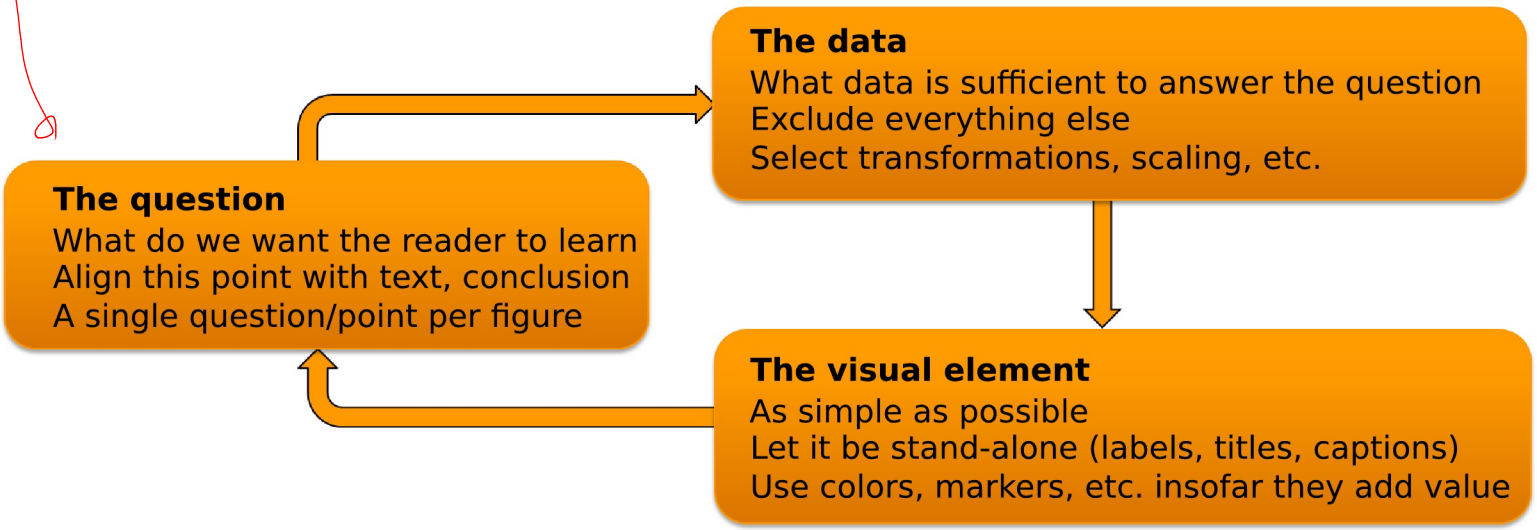
> The drawing shows me at one glance what might be spread over ten pages in a book."
> - Ivan S. Turgenev's novel Fathers and Sons, 1862.
> Use a picture. It's worth a thousand words."
> - Arthur Brisbane to the Syracuse Advertising Men's Club, in March 1911

- A main function of the brain is to process visual information
- We can exploit this capacity using visualization of the data in order to:
  - Detect new patterns, i.e. exploratory data analysis)
  - **Dissiminate results, i.e. visualizations/plots in written work (today)**
- We should take into account how the brains visual system works



Illustrations as technical writing

- The purpose of the text is to communicate an idea *(vs. plots has a purpose)*
- Be grammatically correct *(vs. elementary "rules" of good plotting)*
- Ensure the text is readable *(vs. labels, legends or lines nobody can read)*
- Avoid long/complicated paragraph *(vs. plots that are overly complicated)*
- Dont lie or exagerate. *(vs. distort data in a plot)*



Important choices for visualizations

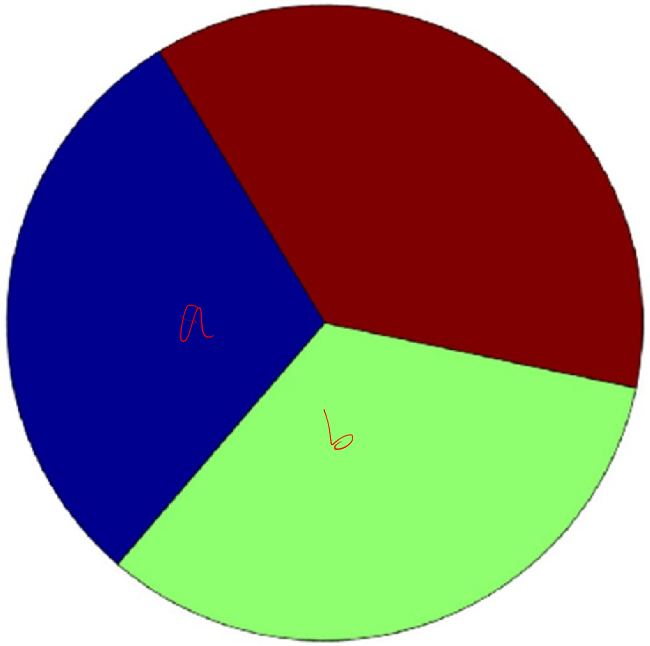
- **Representation:** How will you map objects, attributes, and relations to visual elements?
  - Positions, lengths, colors, areas, orientation

- **Arrangement:** How will you display the visual elements?
  - Viewpoint, transparency, separation, grouping

- **Selection:** How will you handle a large number of attributes and data objects?
  - Display a subset, focus on a region of interest, show summaries

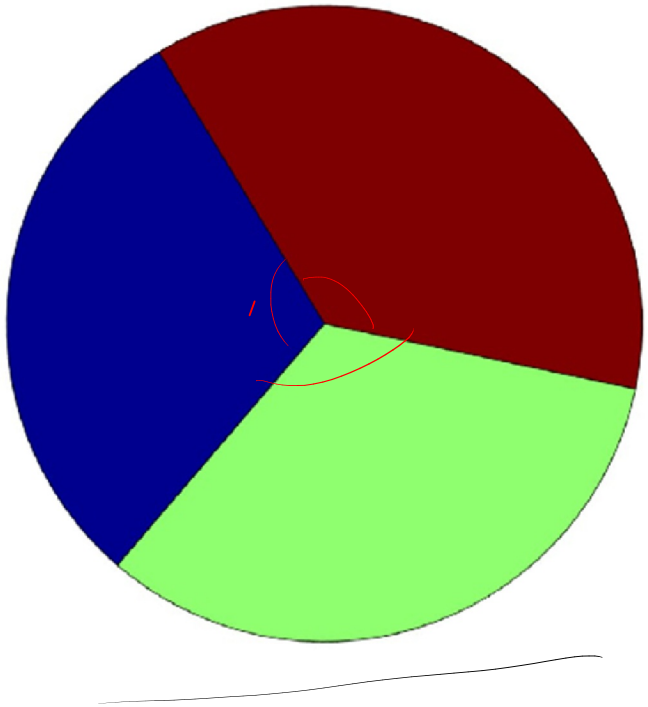
Representation

- **Area represents proportion**
  - Which is smallest, middle, and largest?
  - What are the proportions approximately?

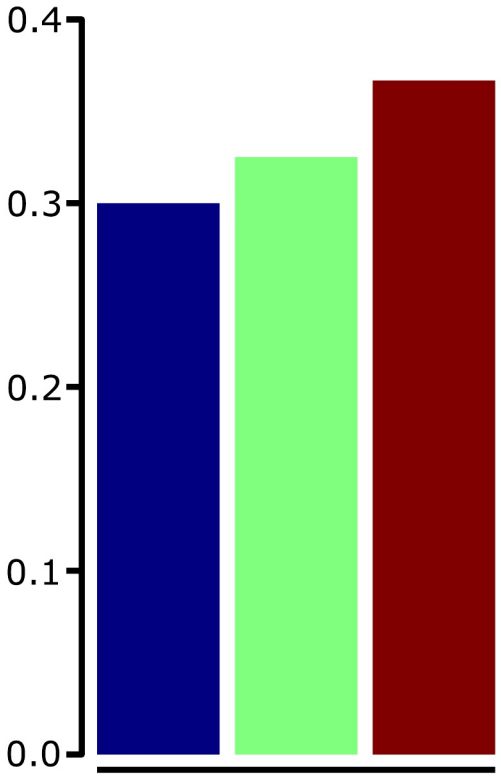


Representation

- **Area represents proportion**
  - Which is smallest, middle, and largest?
  - What are the proportions approximately?



- **Height represents proportion**



Selection

- Elimination or de-emphasis of certain objects or attributes

- A subset of **attributes**
  - **Why?** A graph can only show so many attributes – focus on the relevant
  - **How?**
    - Dimensionality reduction
    - Plot pairs of attributes

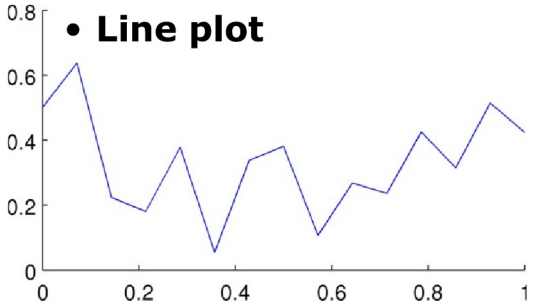
- A subset of **objects**
  - **Why?** A graph can only show so many objects – focus on the relevant
  - **How?**
    - Random sampling
    - Display of region of interest
    - Use density estimation

Types of plots

- **Distribution of a single attribute**
  - Histogram
  - Empirical cumulative distribution
  - Percentile plots
  - Box plot
- **Relation between attributes**
  - 2D histogram
  - Heat maps and contour plots
  - Scatter plots
- **Visualization of high-dimensional objects**
  - Matrix plots
  - Parallel coordinates
  - Star plots

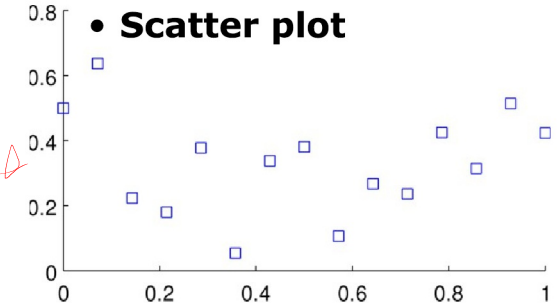
Basic plots

- **Line plot**



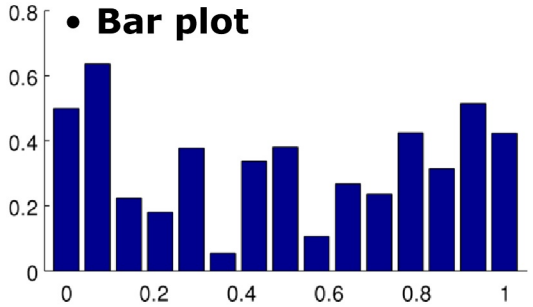

```
plot(x,y);
```

- **Scatter plot**




```
plot(x,y,'s');
scatter(x,y,'s')
```

- **Bar plot**




```
bar(x,y);
```

The iris data set

- **Three flowers**
  - 50 instances of each class, 150 in total

- **Attributes**
  - Sepal (outermost leaves)
    - length in cm
    - width in cm
  - Petal (innermost leaves)
    - length in cm
    - width in cm
  - Class of flower
    - Iris Setosa
    - Iris Versicolour
    - Iris Virginica

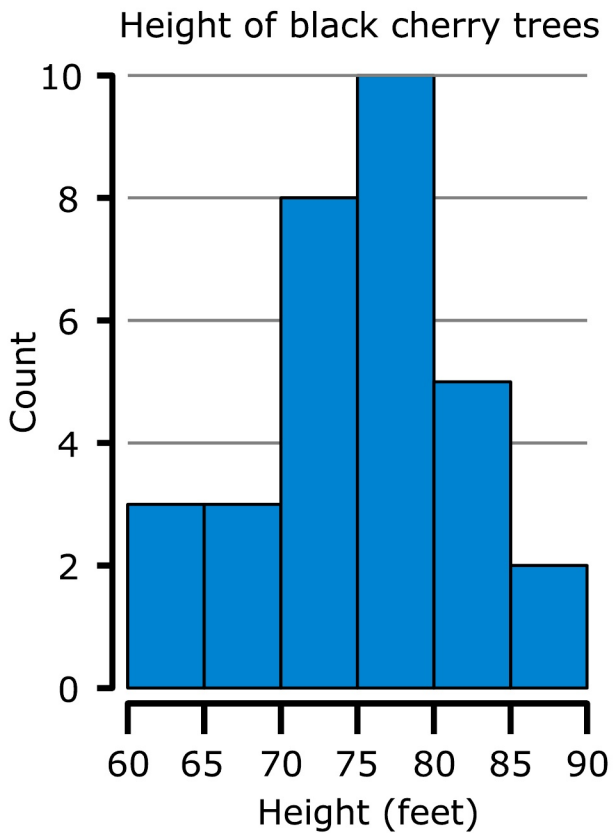
Fower ID	Attribute			
	Sepal Length	Sepal Width	Petal Length	Petal Width
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
.	.	.	.	.
.	.	.	.	.
150	5.9	3.0	5.1	1.8

X Observation x Attribute

Distribution of a single attribute

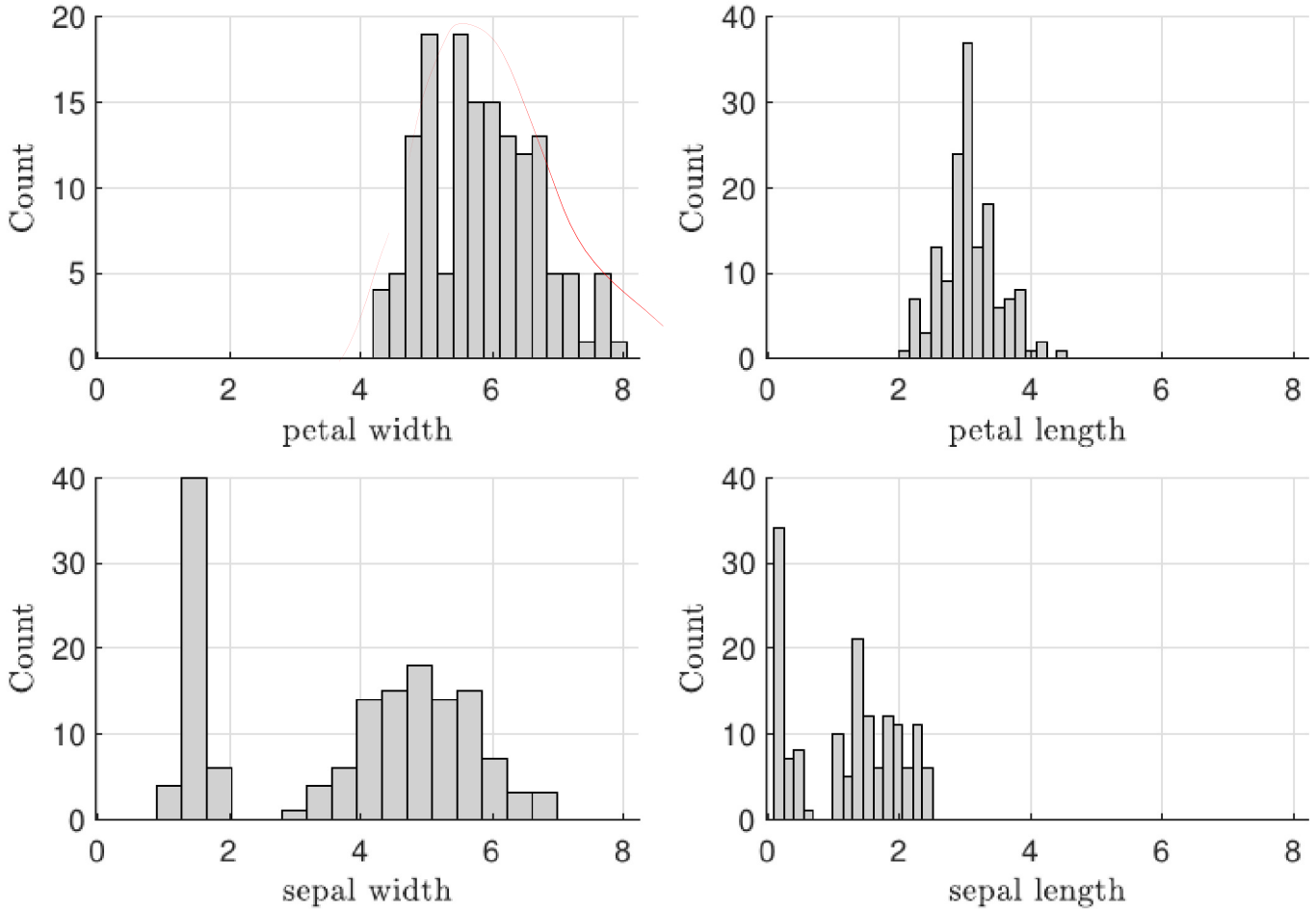
Histograms

- **Shows distribution of a single variable**
  - Divide the values into bins
  - Bar plot of the number of values in bin
  - Height indicates count of values
  - Shape determined by
    - Distribution of data
    - Number of bins / bin width



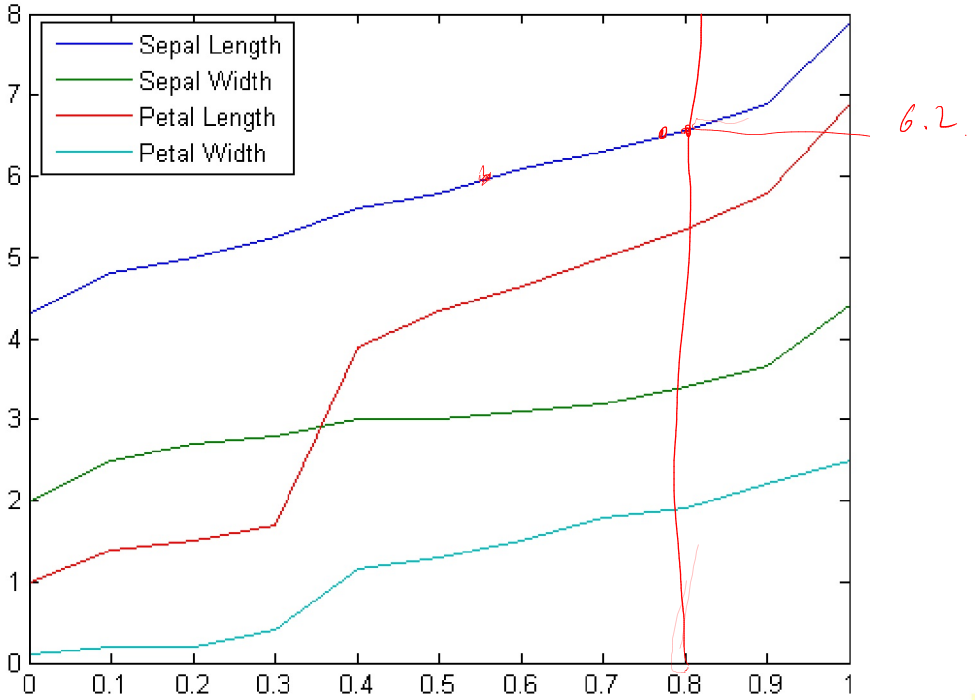
H = {60, 64, 64, 66, 67, 69, 71, 72, 72, 72, 72, 73, 74, 74, 75, 75, 76, 76, 76, 77, 77, 78, 78, 79, 80, 80, 81, 82, 84, 85, 85, 89}

Histograms of the Iris data attributes



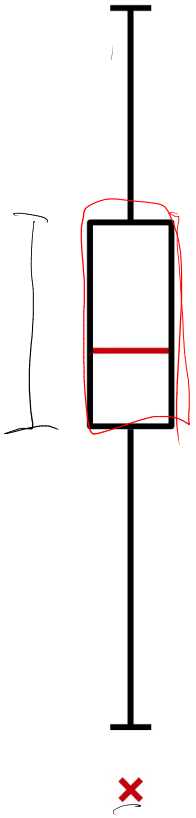
Percentile plots

Percentiles: Given an ordinal or continuous attribute **x** and a number **p** between 0 and 100, the **p**th percentile is a value $\mathbf{x}_p$ of **x** such that **p** percent of the observed values of **x** are less than $\mathbf{x}_p$.




```
prctile = 0:0.1:1;
Y = quantile(X,prctile);
plot(prctile,Y);
legend(attributeNames);
```

Box plots



Min(max data, $q_{75}+1.5(q_{75}-q_{25})$)

75th percentile (q_{75})

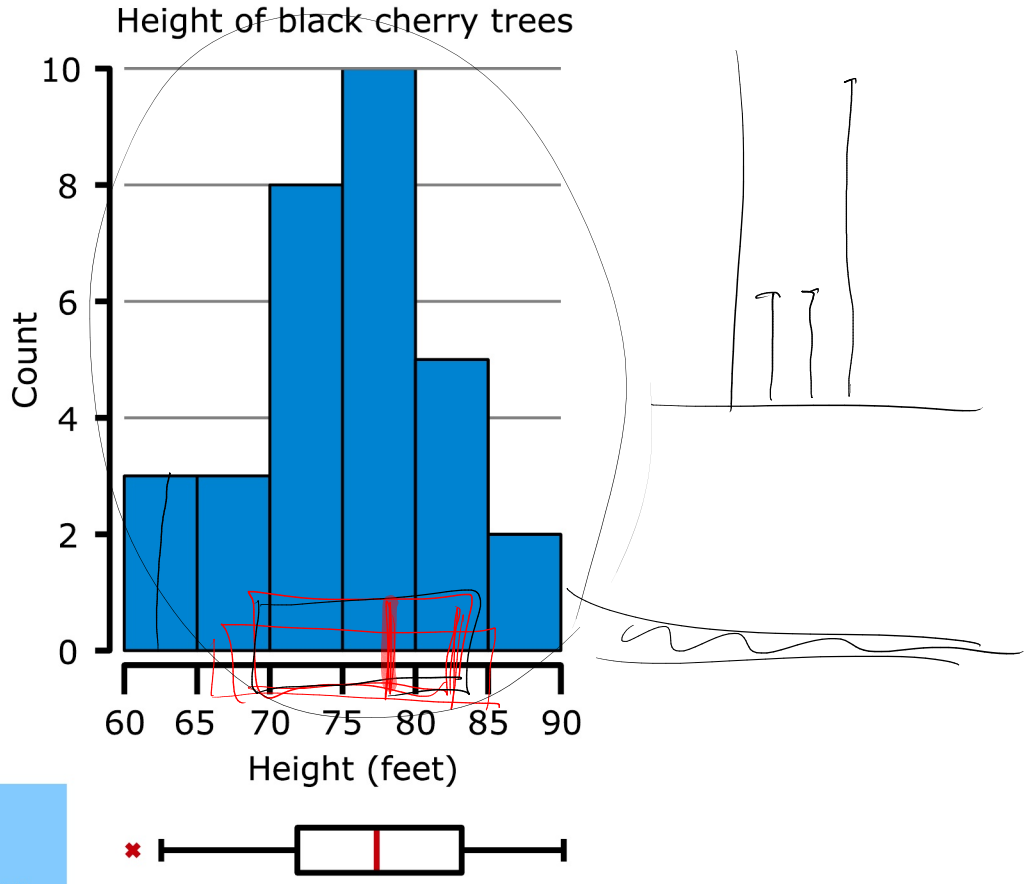
Median (q_{50})

25th percentile (q_{25})

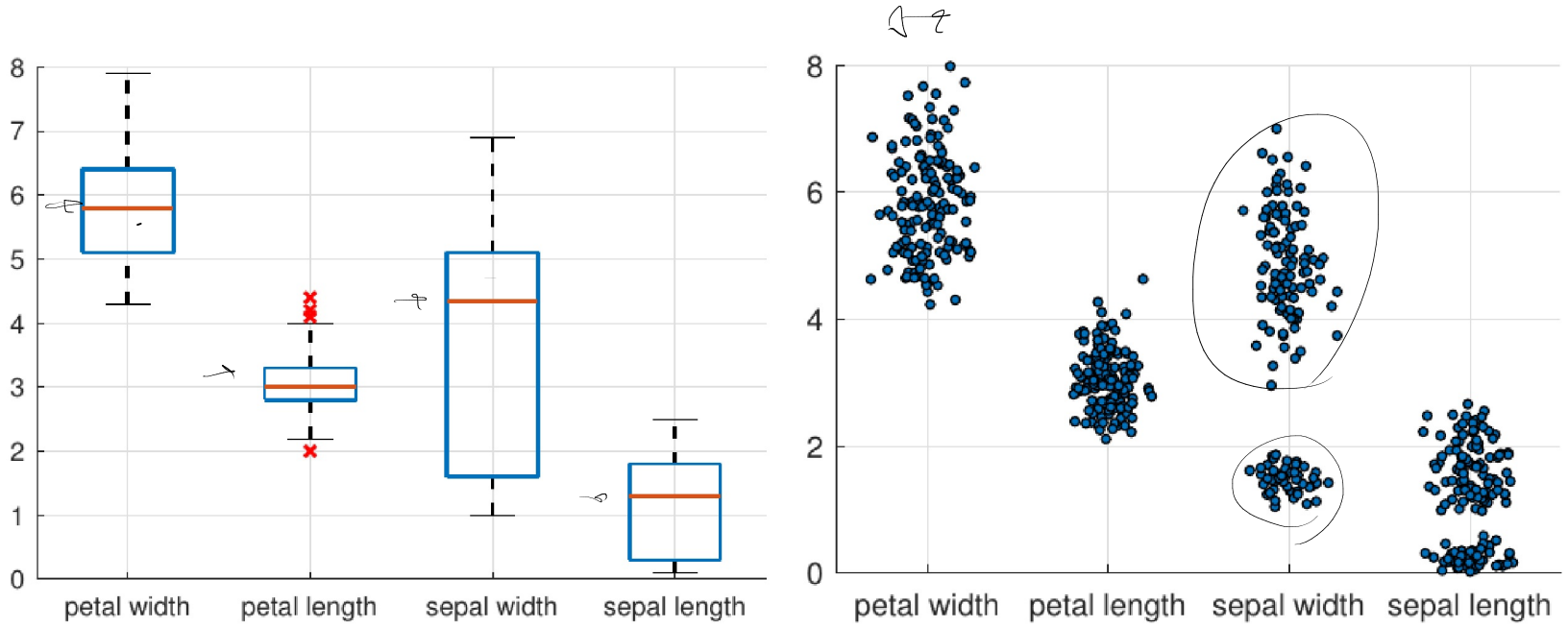
Max(min data, $q_{25}-1.5(q_{75}-q_{25})$)

Outlier

The plotted whisker extends to the adjacent value, which is the most extreme data value that is not an outlier.



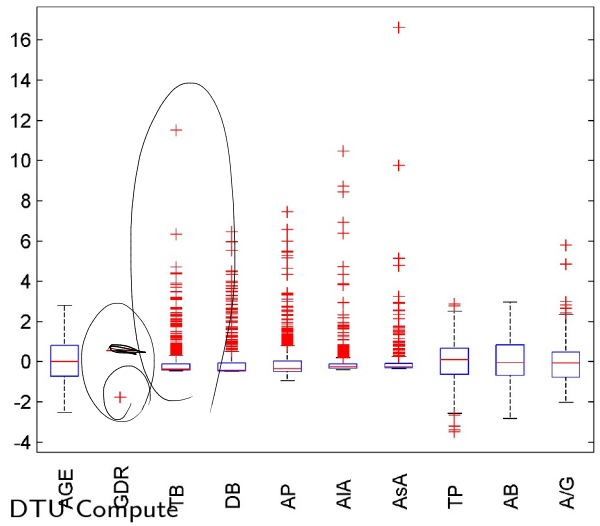
Box plots


8
7
6
5
4
3
2
1
0
petal width
petal length
sepal width
sepal length
8
6
4
2
0
petal width
petal length
sepal width
sepal length

Quiz 3, Boxplots (Fall 2012)

No.	Attribute description	Abbrev.
x_1	Age (in years)	AGE
x_2	Gender (Female=0, Male=1)	GDR
x_3	Total Bilirubin	TB
x_4	Direct Bilirubin	DB
x_5	Alkaline Phosphotase	AP
x_6	Alamine Aminotransferase	AlA
x_7	Aspartate Aminotransferase	AsA
x_8	Total Protiens	TP
x_9	Albumin	AB
x_{10}	Albumin to Globulin ratio	A/G
y	0=No liver disease, 1=Liver disease	LD

Table 1: Liver disease dataset.



The attributes x_1–x_{10} are standardized (i.e., the mean has been subtracted each attribute and the attributes divided by their standard deviations). The figure shows a boxplot fo the standardized data. Which of the following statements is *correct*?

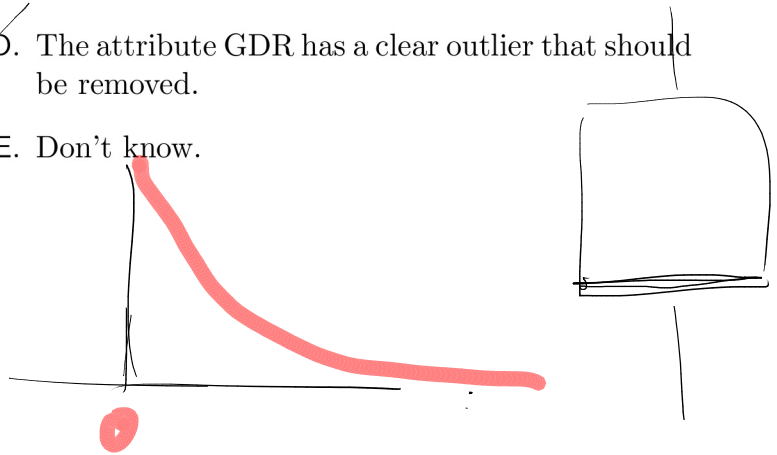
A. The value of the 50th and 75th percentiles of the attribute DB coincides.

B. Even though the distribution of AlA and AsA may have a similar shape this does not imply that the two attributes are correlated.

C. The attribute TB is likely to be normal distributed.

D. The attribute GDR has a clear outlier that should be removed.

E. Don't know.

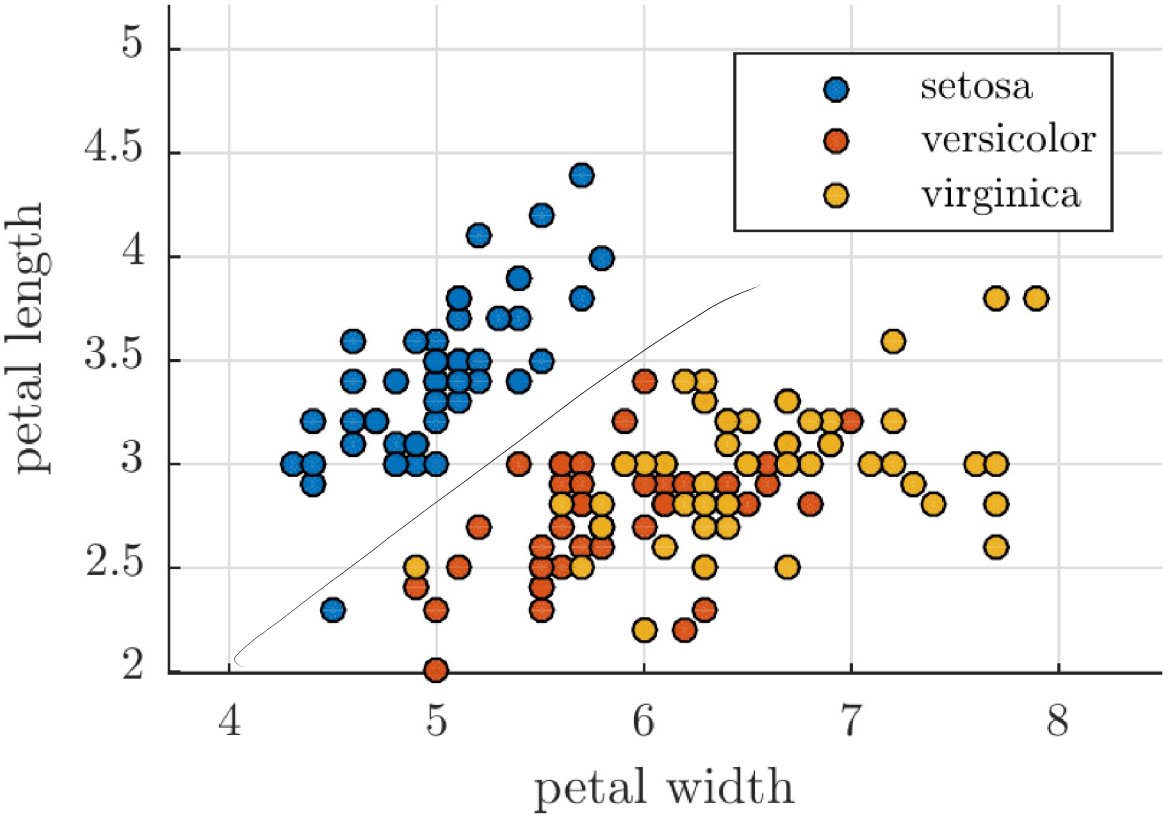


The 25th and 50th percentile but not the 50th and 75th percentiles of the attribute DB coincides. AlA and AsA will not necessarily be highly correlated even though their distributions may have a similar shape (hence, this is correct). For attributes to be correlated it is important they take on high or low values systematically, however, this can not be inspected in a boxplot. TB is not likely to be normal distribution as this attribute does not have a symmetric but highly right skewed distribution. The attribute GDR does not have a clear outlier, in fact the outlier corresponds to the females in the dataset and all we can deduce from the plot is that more than 75 % of the observations are males.

Relation between attributes

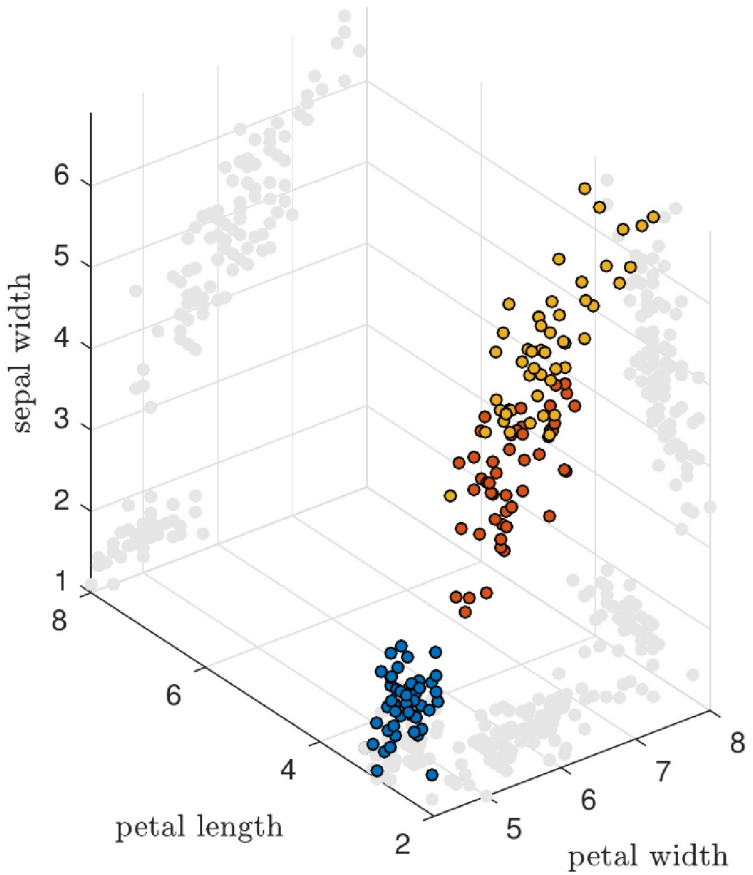
Scatter plots

- Shows **relation** between attributes
  - Assess dependence between attributes
  - Used with classes to assess separability



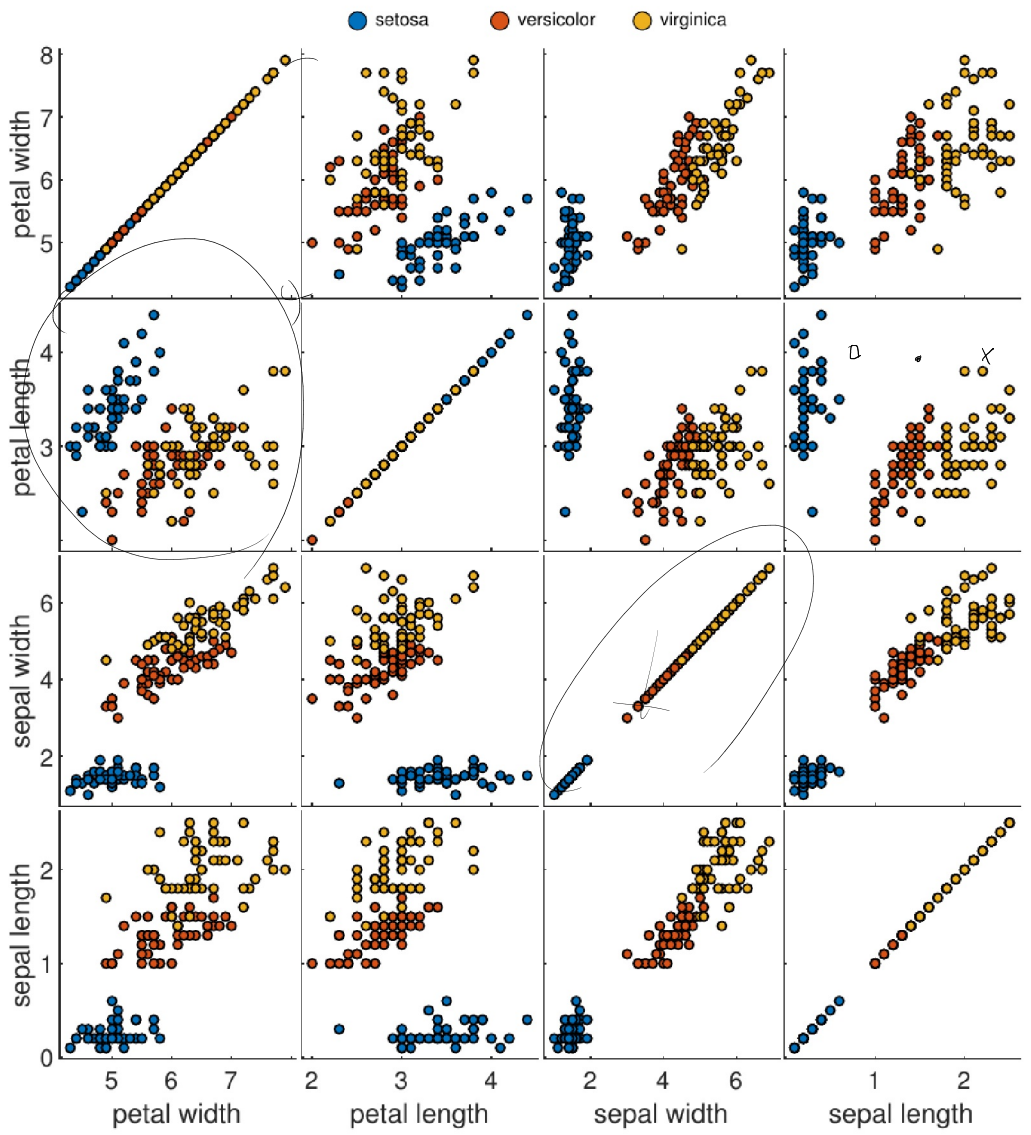
Scatter plots

- Shows **relation** between attributes
  - Assess dependence between attributes
  - Used with classes to assess separability
  - 3D plots are often confusing;
    avoid if possible



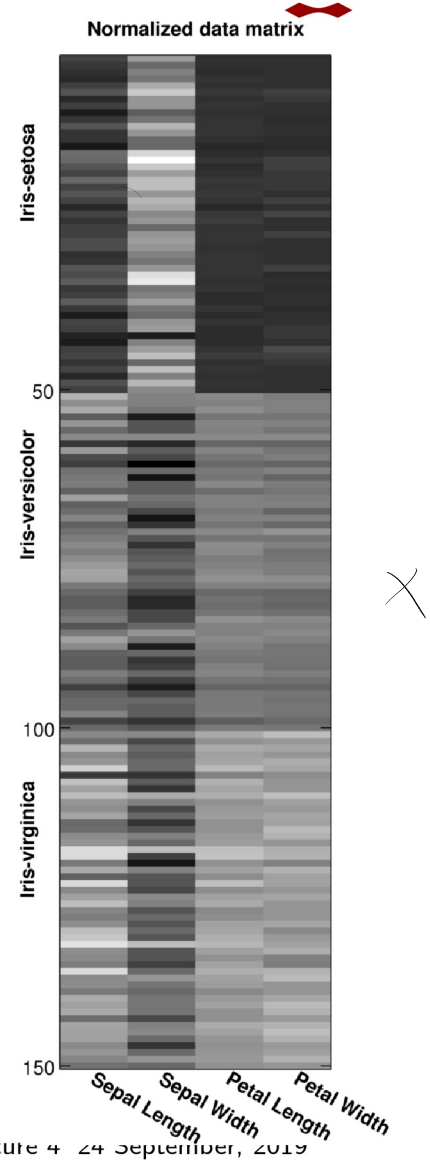
Scatter plots

- Scatter plot matrix
  - All pairs of attributes

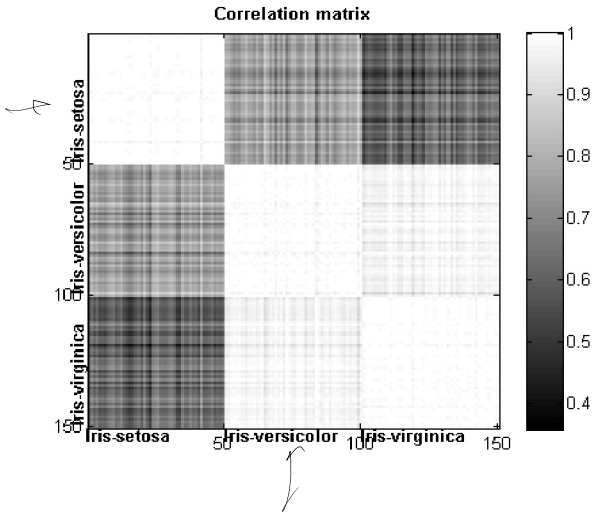


Matrix plots

- **Plot of raw data matrix**
  - Useful when objects are sorted according to class
  - Typically, attributes are normalized

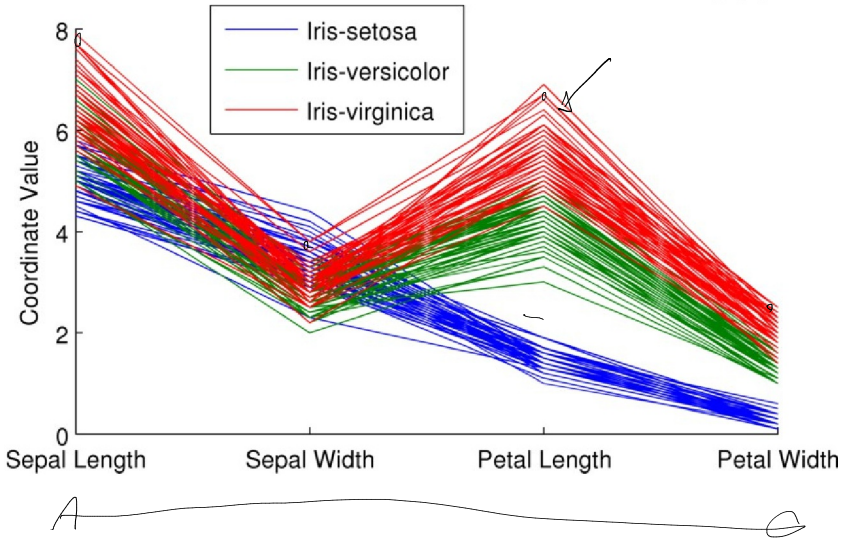


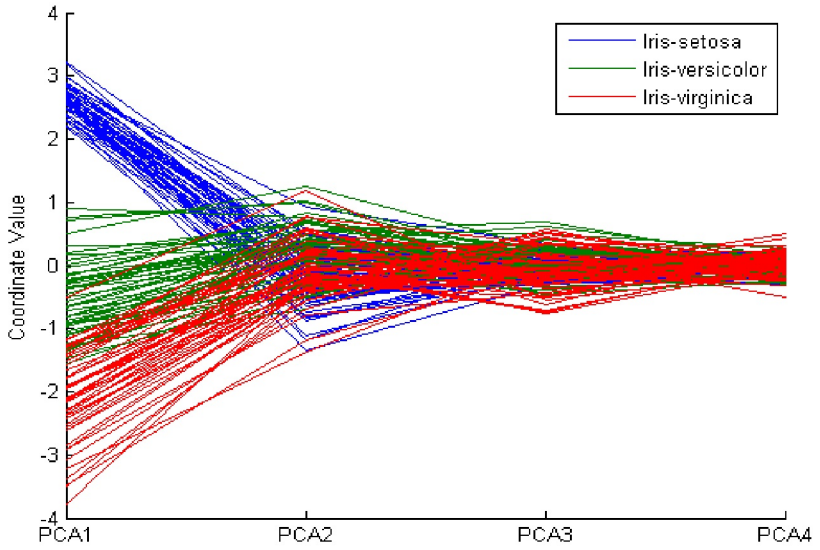
- **Plots of similarity matrices**
  - Useful for visualizing the relation between objects



Parallel coordinates

- **Plot high-dimensional data**
- Instead of perpendicular axes
  - Use parallel axes
- Attribute values are plotted as a point
  - and the points are connected by a line
- Each object is represented as a line
- Lines representing a group of objects
  - Are similar in some sense
  - Ordering of attributes is important in seeing such groupings





ACCENT

- **Apprehension**
  - Is it easy to see what is important in the graph?
- **Clarity**
  - Are the most important elements visually most prominent?
- **Consistency**
  - Have you used the same colors, shapes, etc. as in other graphs?
- **Efficiency**
  - Does it convey its information in the most simple and efficient way?
- **Necessity**
  - Are all elements of the graph necessary to represent data?
- **Truthfulness**
  - Does the graph represent the data correctly?

Tufte's guidelines

- **Graphical excellence**
  - Well-designed presentation of interesting data – a matter of
    - substance, statistics, and design
  - Complex ideas communicated with
    - clarity, precision, and efficiency
  - Gives the viewer
    - the greatest number of ideas
    - in the shortest time
    - with the least ink
    - in the smallest place.
  - Nearly always multivariate
  - Requires telling the truth about the data
  - ~~Maximise~~ Data-ink ratio:

$$\text{Data-ink ratio} = \frac{\text{Data-ink}}{\text{Total ink used}}$$

Edward Tufte

https://commons.wikimedia.org/wiki/File:Edward_Tufte_-_cropped.jpg
Made available by Keegan Peterzell

Making good data visualizations is an art

For some interesting data visualizations see also
http://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization.html
http://www.informationisbeautiful.net/
http://www.junkcharts.typepad.com/

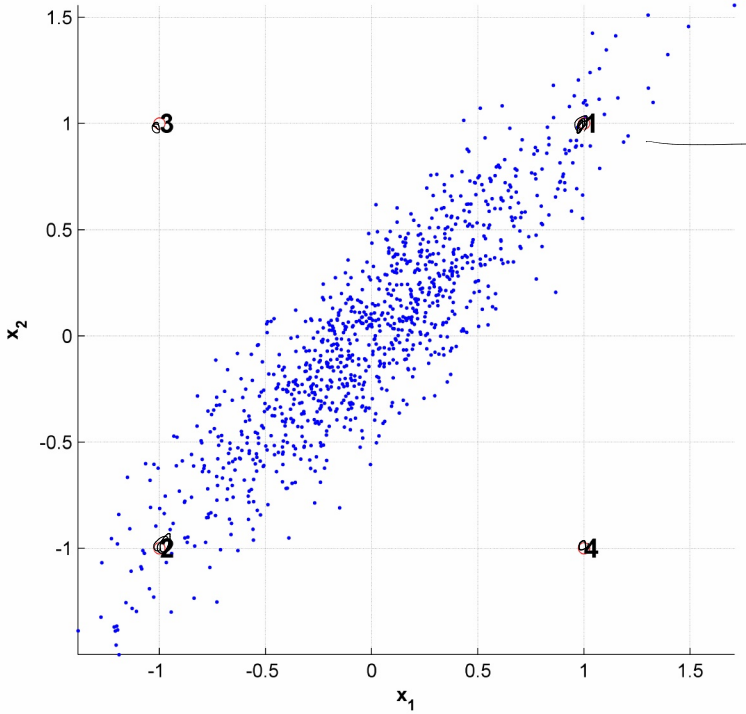
The Mahalanobis distance

How far are x_1 and x_2 apart?

- mahalanobis(x_1,x_2) = 4.2
- $d_{\text{Euclidean}}(x_1,x_2)^2$ = 8.0

How far are x_3 and x_4 apart?

- mahalanobis(x_3,x_4) = 80
- $d_{\text{Euclidean}}(x_3,x_4)^2$ = 8.0



$$\text{mahalanobis}(x,y)^2 = (x-y)^\top \Sigma^{-1}(x-y)$$

$$d_{\text{euclidian}}(x,y)^2 = (x-y)^\top I^{-1}(x-y)$$

Resources

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)

https://junkcharts.typepad.com Excellent resource on creating good visualizations (https://junkcharts.typepad.com/junk_charts/)

http://www2.imm.dtu.dk Our demo of the multivariate normal distribution which illustrates the effect of the covariance matrix

(http://www2.imm.dtu.dk/courses/02450/DemoNormal.html)