Machine Learning 1

Lecture 01 - Overview and Probability Theory

Patrick Forré

1 Overview

2 Probability Theory

Literature

- C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2009.
- mathematical monk's YouTube channel.
- Andrew Ng's Machine Learning on Coursera.
- Wikipedia.
- Other media (internet, learning platforms, etc.).

What is Machine Learning about?

What is? (Machine Learning)

Algorithms for inferring unknowns from knowns. [1]

1. What do i want to know in what kind of circumstances (e.g. a target variable t)?
2. What data can i get and process (data $x_1, \ldots, x_n$)?
3. What algorithms can i use?

[1] taken from mathematicalmonk

Example Applications of Machine Learning

- House price prediction on basis of its location, size etc.
- Adaptive websites adjusted according to user activity.
- Online advertisement.
- Spam filter on basis of word usage and users' labeling.
- Handwriting/digit/speech/image recognition.
- Stock price prediction on basis of historic time series.
- Movie or restaurant recommondation based on other ratings.
- Risk for gettig cancer or diabetes based on clinical data.
- Predict protein structure or function on basis of gene location.
- Credit card fraud detection.
- Search engines.
- Sentiment analysis based on written text.
- Classifying DNA sequences based on similarities.
- Automated Internet Poker Playing.

Different Types of Learning: Supervised Learning

What is? (Supervised Learning)

We talk about Supervised Learning, if for the known data cases $x_1, \ldots, x_n$ we also know the target variables $t_1, \ldots, t_n$. The task now is to make a good prediction for the target variable t for new data x, where t is not known anymore.

This boils down to estimating a function f such that for all known and unknown(!) (x, t) we have $f(x) \approx t$.

Definition (Classification and Regression)

- If t is a discrete variable (i.e. takes values in a countable or finite set like $\{0, 1\}$) this task is called *Classification*.
- If t is a continuous variable (i.e. takes values in $\mathbb{R}$ or $\mathbb{R}^d$) it is called *Regression*.

Example: Supervised Learning

- spam filter: x: email text; t: "spam" or "non-spam" (classification).
- house price prediction: x: size of the house; t: house price (regression).
- digit recognition: x: matrix/vector of pixels; t: $0, 1, \ldots, 9$.

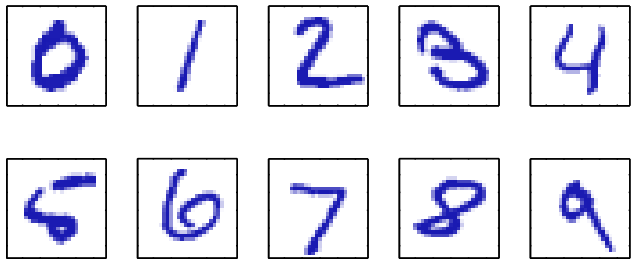


Figure: MNIST data set (Bishop 1.1), classification.

- medical diagnosis: $x = (x_1, \ldots, x_r)$: vector of clinical measurements (e.g. age, blood pressure, bmi), t: risk type for having a stroke: "high", "normal", "low" (classification).

Example: Classification

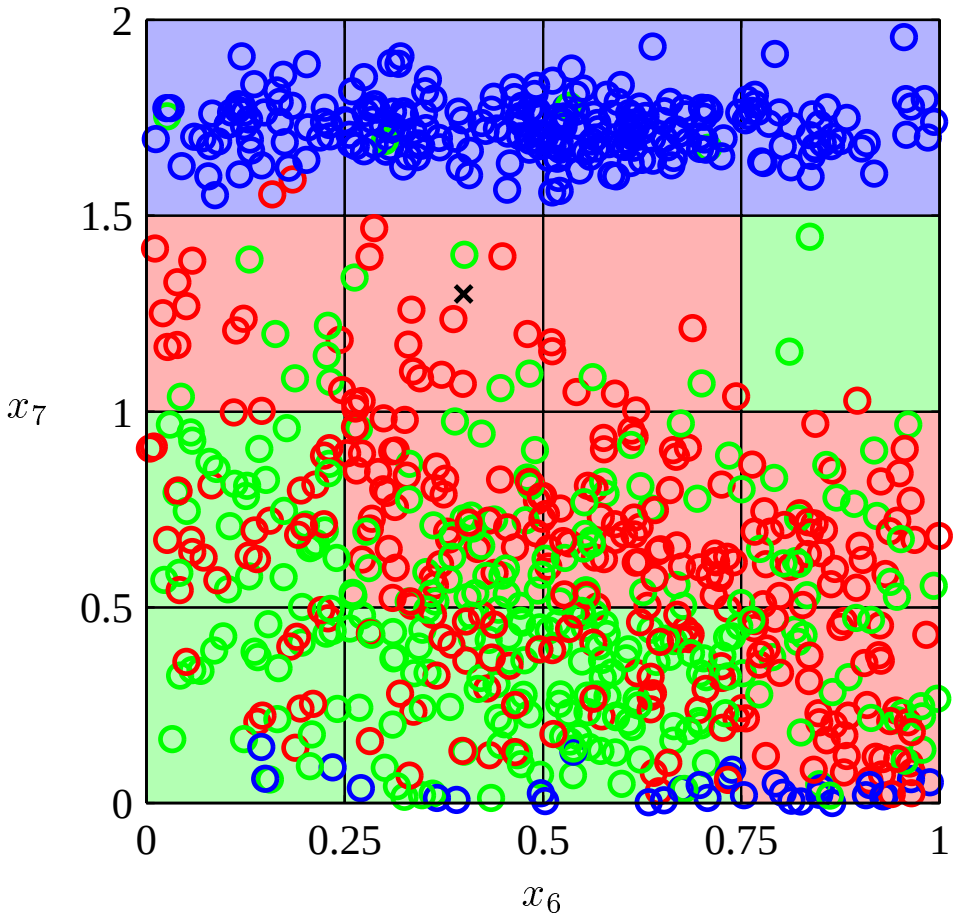
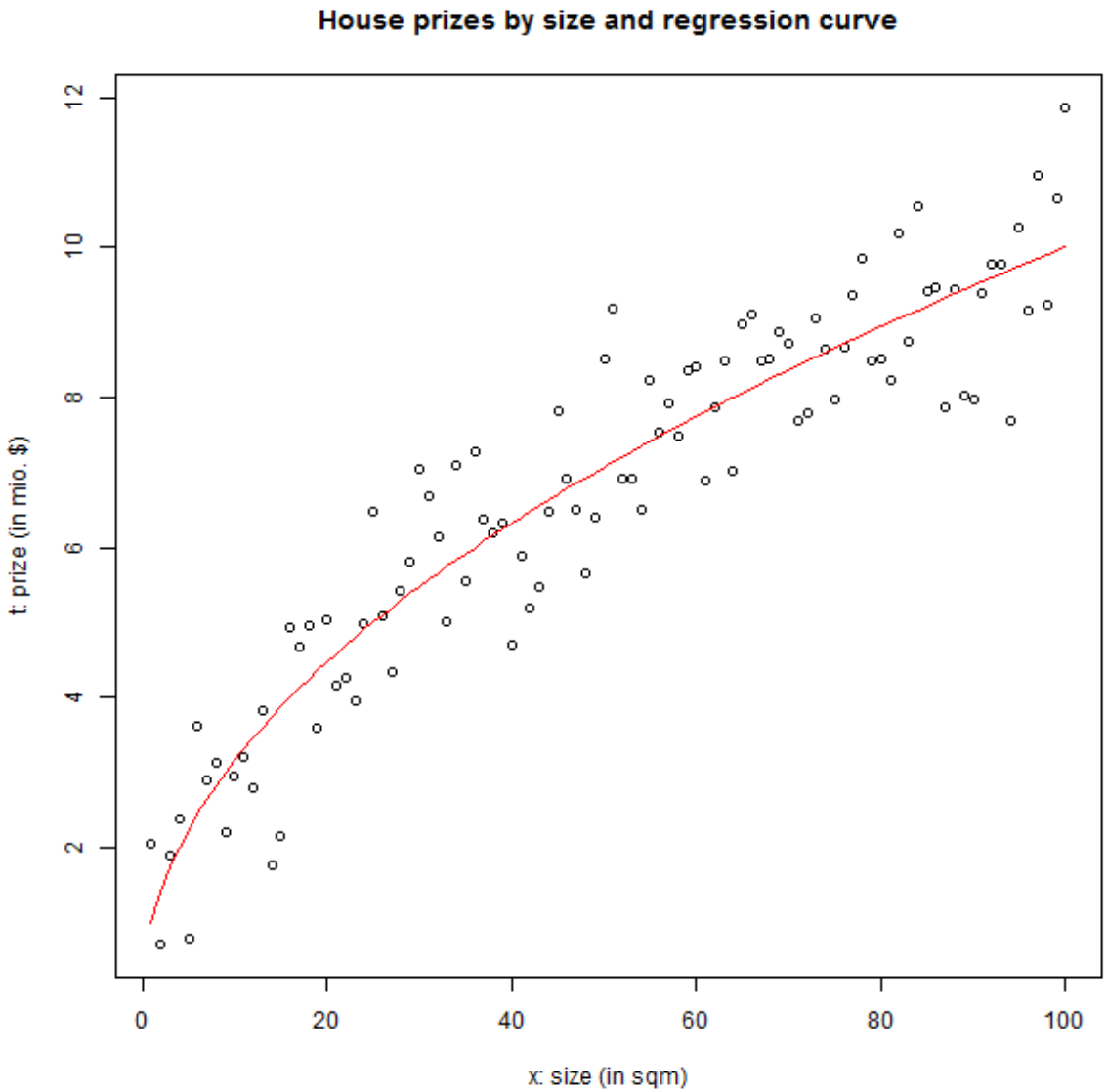


Figure: Classification of labeled data points (Bishop 1.20)

Example: Regression



Different Types of Learning: Unsupervised Learning

What is? (Unsupervised Learning)

We talk about Unsupervised Learning if for the known data cases $x_1, \ldots, x_n$ no target variables are given.

The task now is to find an "inner representation" of the known data to make it more accessible and such that new data x can relate to it.

Typical approaches are Clustering, Dimensionality Reduction, Density Estimation.

Which approach to use depends on our application in mind.

Remark

Clustering can in some cases be seen as "unsupervised classification", and dimensionality reduction as a kind of "unsupervised regression".

Example: Unsupervised Learning

- "Classifying" handwritten digits without knowing the labels (clustering: 10 clusters).

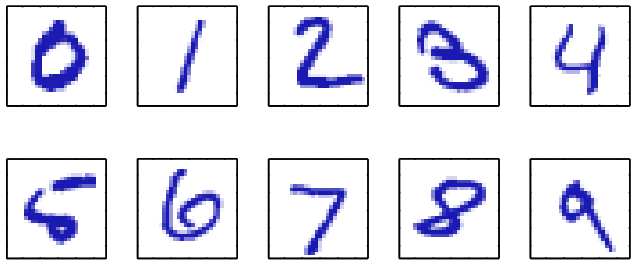
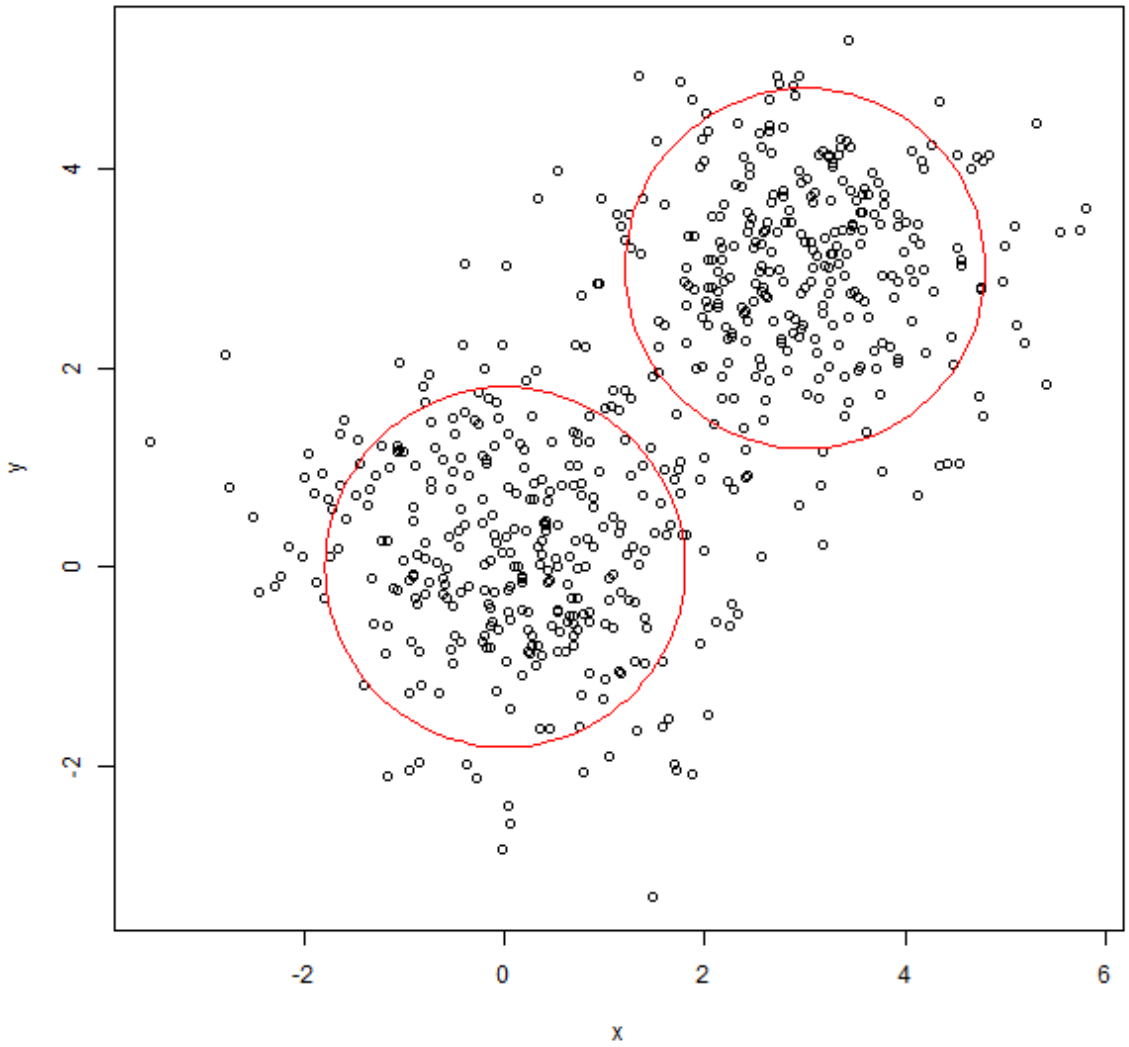


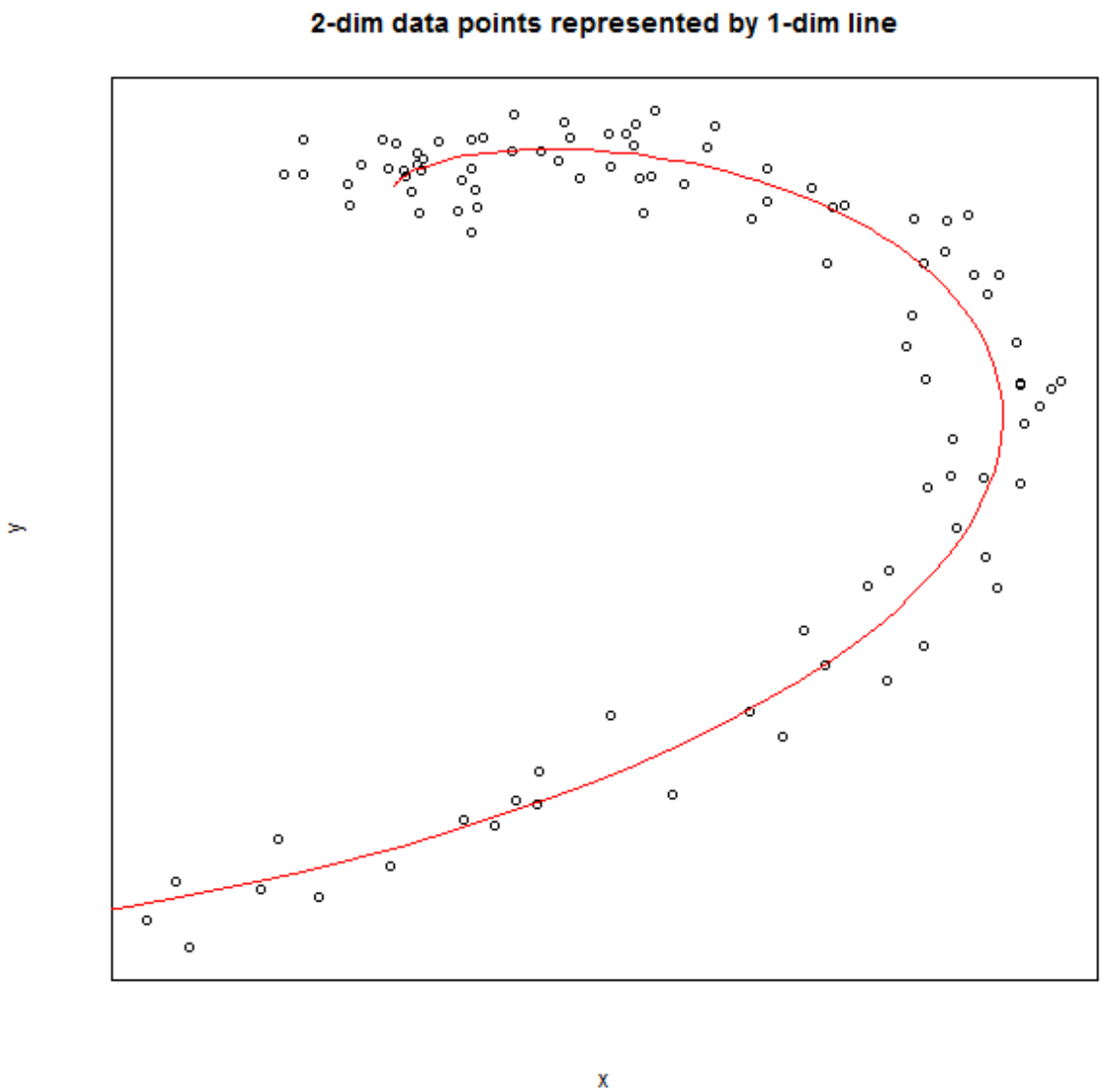
Figure: MNIST data set (Bishop 1.1), classification.

- Credit card fraud detection: $x = (x_1, \ldots, x_n)$: vector of credit card usage, e.g. time, location, amount (density estimation: probability estimation to detect unlikely events).
- Representation and preprocessing of data (dimensionality reduction).

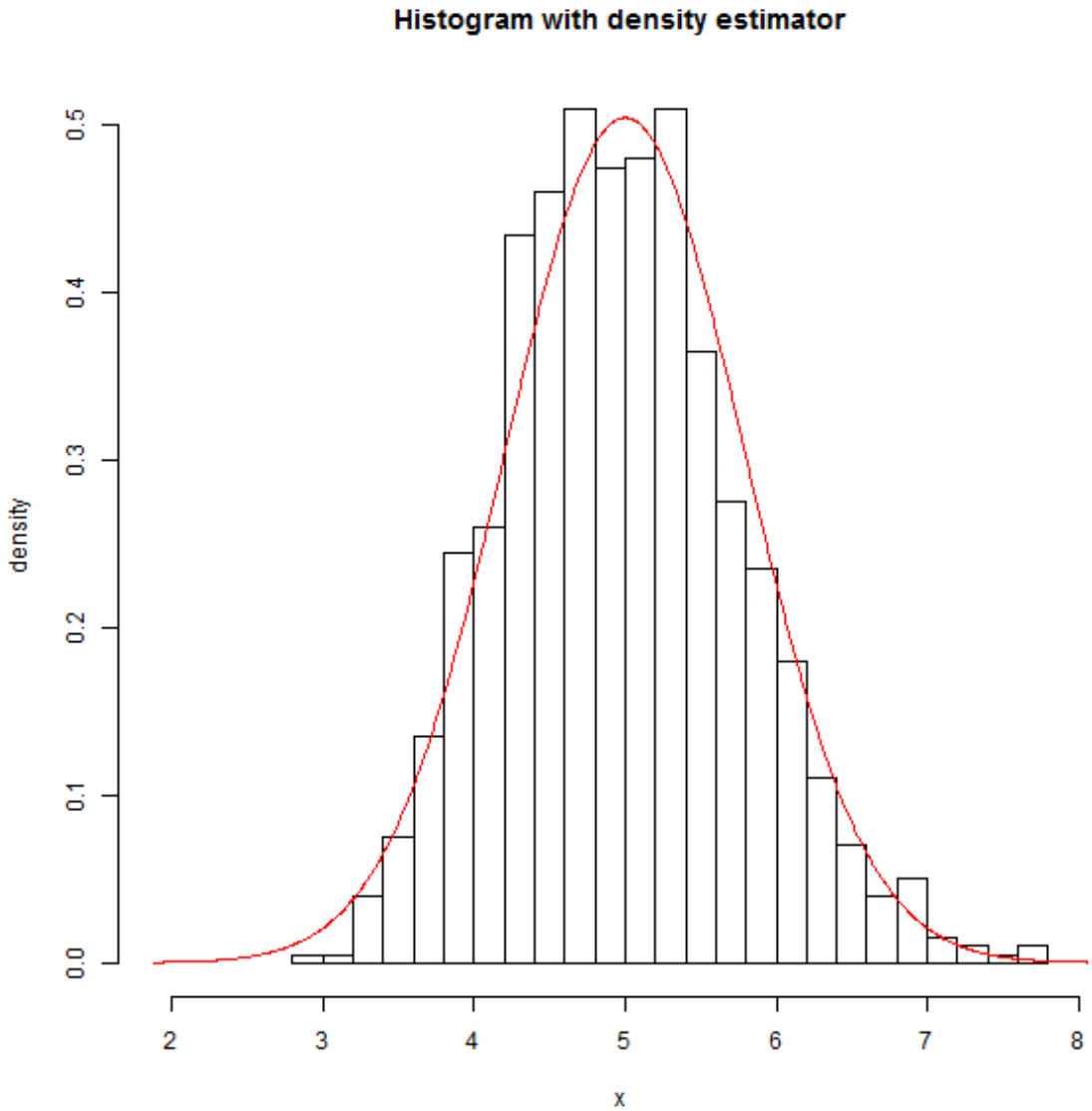
Example: Clustering



Example: Dimensionality Reduction



Example: Density Estimation



Different Types of Learning: Other types

What is? (Semi-Supervised Learning)

We talk about Semi-Supervised Learning if for a few of the data points $x_1, \ldots, x_n$ we also know correct target variables $t_1, \ldots, t_k$ (for a $k < n$).

What is? (Reinforcement Learning)

We talk about Reinforcement Learning if we are in a dynamical environment and we want the machine to evaluate "actions" given the situation and to make good decisions. The machine learns to become better by repetition and adapting to reward and punishment. I.a.W. we want the machine to learn how to play "games".

1 Overview

2 Probability Theory

Probability Theory

What is? (Probability Theory)

Probability theory is a consistent framework for quantifying uncertainty and of rules for manipulating it.

Probability theory is often used to model the influences of unknown parameters.

In a frequentist interpretation the probability of an event is the fraction of times that event occurs in the experiment (if repeated infinitely often), i.e. the relative frequency of that event.

In the Bayesian framework probability is often interpreted as a quantification of plausability or the strength of the belief for an event.

Random Variables

What is? (Random Variable)

A random variable X is a stochastic variable taken from a given set of possible values/outcomes $\mathcal{X}$.

We will write $X \in \mathcal{X}$, but we need to distinguish between the variable X and an instance/outcome x.

It might represent an stochastic "experiment" or the gathering of data.

Every random variable $X \in \mathcal{X}$ can be assigned a probability distribution $\mathbb{P}$ (reflecting the relative frequency or plausability of its outcomes).

For an event $A \subset \mathcal{X}$ we write $\mathbb{P}(X \in A)$ for saying:

"The probability that the value of X will lie in A."

If $\mathcal{X}$ is a countable or finite set we call X a discrete, and if $\mathcal{X}$ is $\mathbb{R}$ or $\mathbb{R}^d$ we call X a continuous random variable.

Example: Random Variables

- Throwing a dice: $X \in \mathcal{X} = \{1, \ldots, 6\}$: $\mathbb{P}(X \in \{2, 5\})$ is then the probability that the dice will show 2 or 5. If the dice is fair then we have $\mathbb{P}(X \in \{2, 5\}) = \frac{2}{6}$.
- Flipping two fair and independent coins:
  $X = (X_1, X_2) \in \mathcal{X} = \{(H, H), (H, T), (T, H), (T, T)\}$.
  $\mathbb{P}(X = (H, T))$ is then the probability that the first coin will show heads and the second tails. It is $\mathbb{P}(X = (H, T)) = \frac{1}{4}$.
- Measuring the gravitational constant g: $X \in \mathcal{X} = \mathbb{R}$. Then $\mathbb{P}(X \in [a, b])$ will approximately be $\int_a^b \mathcal{N}(x|g, \sigma^2)dx$, i.e. Gaussian distributed with mean g and variance σ^2 depending on the measure instrument (see later).

Discrete Random Variables

- For a discrete random variable $X \in \mathcal{X}$ we also write $p(x)$, $p_X(x)$ or $p(X = x)$ for the probability $\mathbb{P}(X = x)$.
  $p(x)$ is called the (probability) mass function for X.
- For events $A \subset \mathcal{X}$ we then have:

$$\mathbb{P}(X \in A) = \sum_{x \in A} p(x),$$

  meaning that the probability that X takes values in A is the sum of all probabilities that X equals x for $x \in A$.
- If $A = \{a_1, a_2 \dots\}$ then this means:

$$\mathbb{P}(X \in A) = p(a_1) + p(a_2) + \dots,$$

  i.e. the probability that X is a_1 or a_2 or ... is the sum of these single probabilities.

Continuous Random Variables

- For continuous random variables we will usually assume a density function $p(x)$, i.e. for an event $A \subset \mathcal{X}$ we have:

$$\mathbb{P}(X \in A) = \int_A p(x)dx.$$

- This means: $p(x)\delta x \approx \mathbb{P}(X \in [x + \delta x))$ for $\delta x \to 0$.

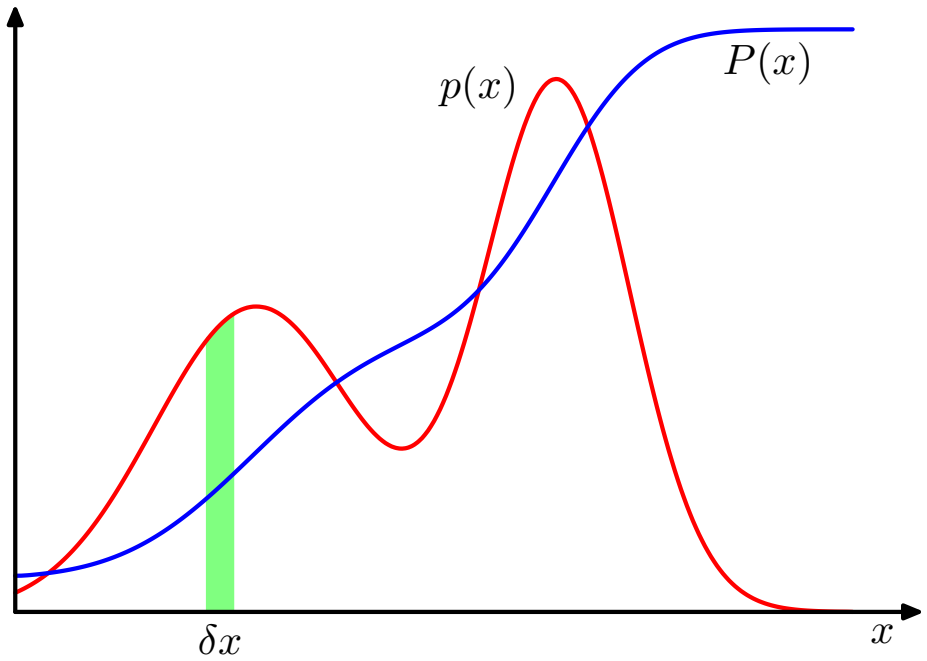
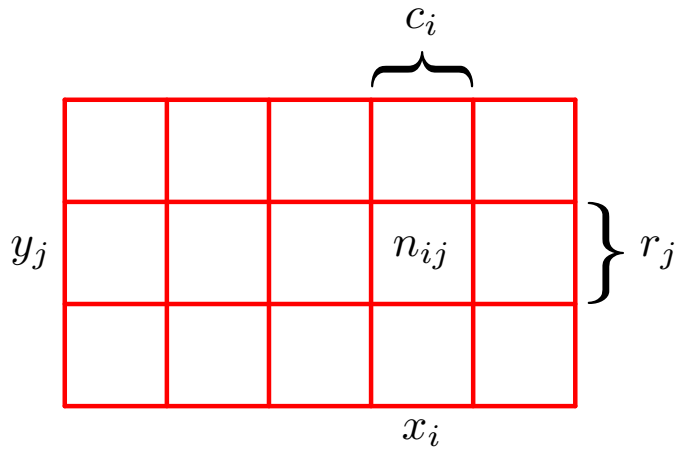


Figure: Bishop 1.12.

Example: Two Random Variables: Example

Consider $X \in \{x_1, \ldots, x_M\}$, $Y \in \{y_1, \ldots, y_L\}$ with joint distribution $p(x, y)$, the probability that $X = x$ and $Y = y$ occurs.

- N: number of trials ($N \to \infty$).
- n_{ij}: number of trials with $X = x_i$ and $Y = y_j$.
- c_i: number of trials with $X = x_i$.
- r_j: number of trials with $Y = y_j$.



Bishop 1.10

We then get: $p(x_i, y_j) = \frac{n_{ij}}{N}$.
The marginal probability is:

$$p(x_i) = \frac{c_i}{N} = \frac{\sum_j n_{ij}}{N} = \sum_{j=1}^{L} \frac{n_{ij}}{N} = \sum_{j=1}^{L} p(x_i, y_j).$$

Example: Two Random Variables (II)

- The conditional probability is: $p(x_i|y_j) = \frac{n_{ij}}{r_j}$.
- We then get:

$$p(x_i|y_j) * p(y_j) = \frac{n_{ij}}{r_j} * \frac{r_j}{N} = \frac{n_{ij}}{N} = p(x_i, y_j).$$

- And summing over all i:

$$\sum_i p(x_i|y_j) = \sum_i \frac{n_{ij}}{r_j} = \frac{r_i}{r_i} = 1.$$

- Similarly:

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

Example: Two Random Variables (III)

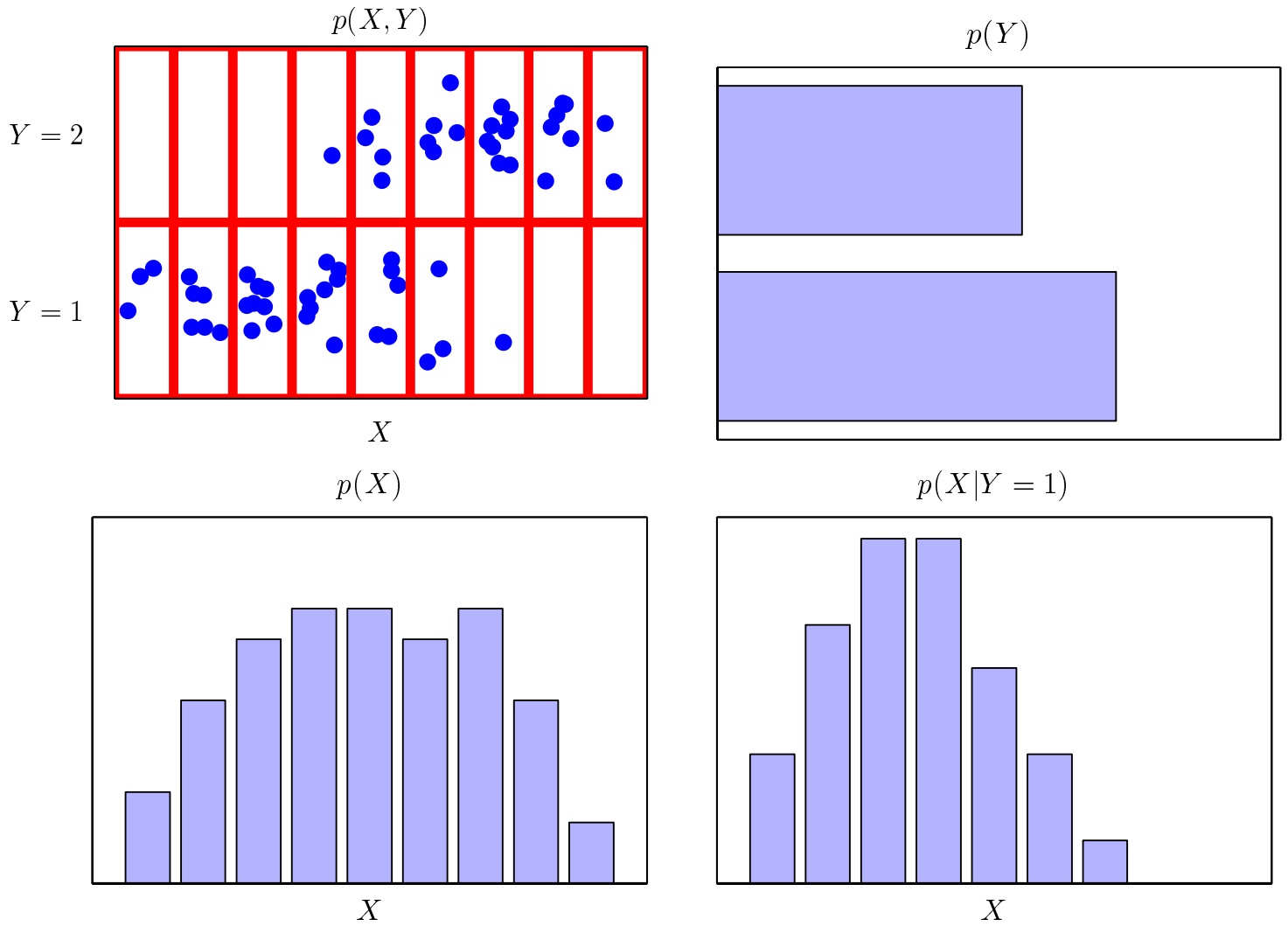


Table: Bishop 1.11

The Rules of Probability Theory

Theorem (The Rules of Probability Theory)

For random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ we have the following rules:

	discrete RV	*continuous RV*
σ-Additivity	$\mathbb{P}(X \in A) = \sum_{x \in A} p(x)$	$\mathbb{P}(X \in A) = \int_A p(x)dx$
Positivity	$p(x) \geqslant 0$	$p(x) \geqslant 0$
Normalization	$\sum_{x \in \mathcal{X}} p(x) = 1$	$\int_{\mathcal{X}} p(x)dx = 1$
Sum Rule	$p(x) = \sum_{y \in \mathcal{Y}} p(x, y)$	$p(x) = \int_{\mathcal{Y}} p(x, y)dy$
Product Rule	$p(x, y) = p(x\|y) \cdot p(y)$	$p(x, y) = p(x\|y) \cdot p(y)$

Bayes' Rule

Theorem (Bayes)

For random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ we have the following rule:

$$p(y|x) = \frac{p(x|y) \cdot p(y)}{p(x)} = \begin{cases} \frac{p(x|y) \cdot p(y)}{\sum_{y' \in \mathcal{Y}} p(x|y') \cdot p(y')} & \text{for discrete RV} \\ \frac{p(x|y) \cdot p(y)}{\int_{\mathcal{Y}} p(x|y') \cdot p(y') dy'} & \text{for continuous RV} \end{cases}$$

I.a.w. the conditioning can be "exchanged" by this rule.

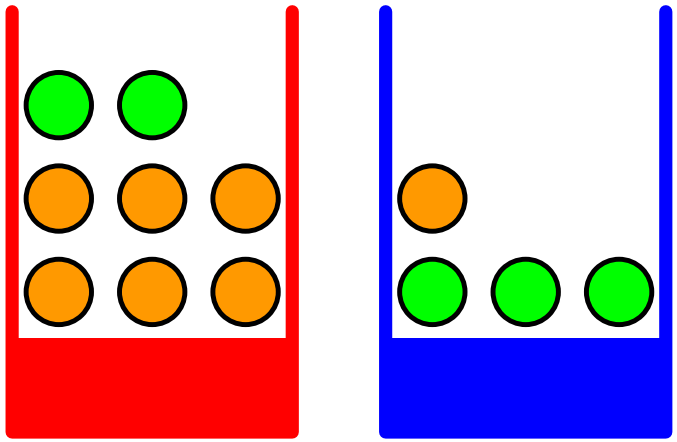
In the context of Bayesian inference (see later) we call:

- $p(y)$: the prior probability of Y (i.e. before observing x).
- $p(y|x)$: the posterior probability of Y (i.e. after observing x).
- $p(x|y)$: the likelihood of $X = x$ given $Y = y$.
- $p(x)$: the evidence for $X = x$.

Example: Fruits in Boxes

We have $B \in \{r, b\}$ and $F \in \{a, o\}$ and given:

- $p(B = r) = 4/10$
- $p(B = b) = 6/10$.
- $p(F = a|B = r) = 2/8$.
- $p(F = o|B = r) = 6/8$.
- $p(F = a|B = b) = 3/4$.
- $p(F = o|B = b) = 1/4$.



Bishop 1.9

We calculate:

- $p(F = a) = p(a|r)p(r) + p(a|b)p(b) = \frac{2}{8} \cdot \frac{4}{10} + \frac{3}{4} \cdot \frac{6}{10} = \frac{11}{20}$.
- $p(F = o) = 1 - p(F = a) = 1 - \frac{11}{20} = \frac{9}{20}$.
- $p(B = r|F = o) = \frac{p(o|r) \cdot p(r)}{p(o)} = \frac{6}{8} \cdot \frac{4}{10} \cdot \frac{20}{9} = \frac{2}{3}$.
- This means that $p(B = r)$ jumps from $\frac{2}{5}$ to $p(B = r|F = o) = \frac{2}{3}$ after observing the fruit $F = o$.

Independent Random Variables

Definition (Independence)

Two random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ are called independent if for all values x, y we have:

$$p(x, y) = p(x) \cdot p(y).$$

This is equivalent to saying that for all x and y (with $p(y) > 0$) we have:

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

In words: X and Y are independent iff measuring X gives no information about Y, and vice versa.

Expectation and Variance

Definition (Expectation)

If $X \in \mathcal{X}$ is random variable and $f : \mathcal{X} \to \mathbb{R}$ a function, then we define the expectation value of f by:

$$\mathbb{E}[f] := \mathbb{E}[f(X)] := \begin{cases} \sum_{x \in \mathcal{X}} f(x)p(x) & \text{for discrete } X \\ \int_{\mathcal{X}} f(x)p(x)dx & \text{for continuous } X \end{cases}$$

Definition (Variance)

In the same situation as above the variance of f is defined by:

$$\mathrm{Var}(f) := \mathrm{Var}(f(X)) := \mathbb{E}[(f - \mathbb{E}[f])^2] = \mathbb{E}[f^2] - \mathbb{E}[f]^2.$$

The variance measures the (quadratic) deviation of f from $\mathbb{E}[f]$.

Covariance

Definition (Covariance)

Let X and Y be real valued random variables. Then we define the covariance of X and Y by:

$$\mathrm{Cov}(X,Y) := \mathbb{E}[(X-\mathbb{E}[X])\cdot(Y-\mathbb{E}[Y])] = \mathbb{E}[X\cdot Y]-\mathbb{E}[X]\cdot\mathbb{E}[Y].$$

The covariance measures how "often" X and Y deviate from their expectation value into the same direction.
If $X = (X_1, \ldots, X_n)$ is a vector of random variables then we define the covariance matrix of X by:

$$\mathrm{Cov}(X) := (\mathrm{Cov}(X_i, X_j))_{i,j}.$$

Note that $\mathrm{Cov}(X_i, X_i) = \mathrm{Var}(X_i)$.

Gaussian distribution

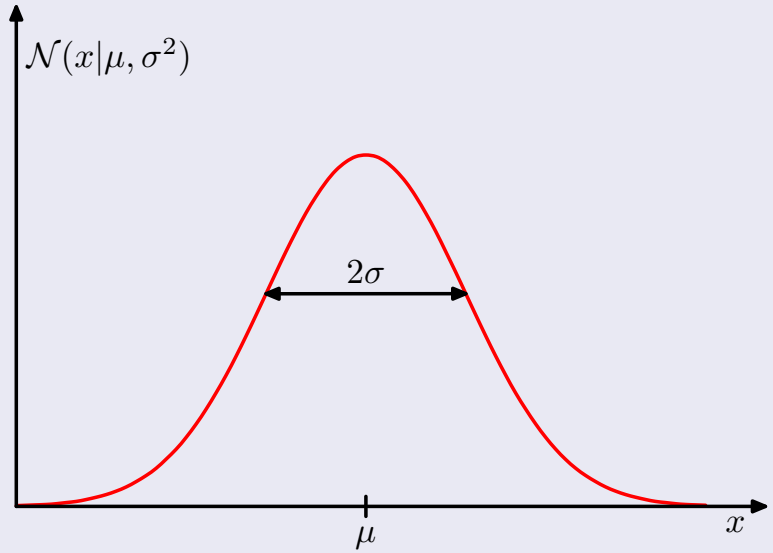
Definition (Gaussion distribution)

A real valued random variable X is said to be Gaussian distributed or normal distributed with parameters μ and σ^2 if X has the density:

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

We have:

- $\mathbb{E}[X] = \mu$.
- $\mathrm{Var}(X) = \sigma^2$.



Bishop 1.13

Multivariate Gaussian

Definition (Multivariate Gaussian distribution in D dimensions)

A vector valued random variable $X = (X_1, \ldots, X_D)^T$ is said to be multivariate Gaussian distributed with parameters $\mu = (\mu_1, \ldots, \mu_D)^T$ and $\Sigma = (\Sigma_{ij})_{i,j}$ if X has the density in $x = (x_1, \ldots, x_D)^T$:

$$\mathcal{N}(x|\mu, \Sigma) := \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left(-(x-\mu)^T \Sigma^{-1}(x-\mu)\right),$$

where Σ is a $D \times D$ covariance matrix and $|\Sigma|$ its determinant.
We have: $\mathbb{E}[X] = \mu$ and $\mathrm{Cov}(X) = \Sigma$.
Note that $\mathcal{N}(x|\mu, \Sigma)$ has $D(D+3)/2$ parameters.

Maximum Likelihood Estimation

- Data set $D = (x_1, \ldots, x_N)$ of N independent observations given.
- We are presented with a class of probability distributions $\{p(x|w)|w \in \mathcal{W}\}$ for x, where $\mathcal{W}$ is an index set (in some $\mathbb{R}^d$).
- Goal: Find an index w^* such that $p(x|w^*)$ "best" explains the occurence of the data D,
- i.e. such that D appears to be a realization of i.i.d. (independent and identically distributed) random variables $X_1, \ldots, X_N$ each of which is distributed like $p(x|w^*)$.
- Maximum Likelihood Principle: The most likely "explanation" of D is given by the index w which maximizes the likelihood $p(D|w)$ (joint distribution).

Maximum Likelihood Estimation (II)

- If $X_1, \ldots, X_N$ are i.i.d. $p(x|w)$-distibuted, then their joint distribution is given by:

$$p(x_1, \ldots, x_N|w) = \prod_{i=1}^{N} p(x_i|w).$$

- The Maximum Likelihood Estimator w_{ML} is determined by:

$$\begin{aligned} w_{\mathrm{ML}} &:= \mathrm{argmax}_{w \in \mathcal{W}}\, p(D|w) \\ &= \mathrm{argmax}_{w \in \mathcal{W}} \prod_{i=1}^{N} p(x_i|w) \\ &= \mathrm{argmax}_{w \in \mathcal{W}} \sum_{i=1}^{N} \log p(x_i|w) \\ &= \mathrm{argmin}_{w \in \mathcal{W}} \left\{ - \sum_{i=1}^{N} \log p(x_i|w) \right\} \end{aligned}$$

- Putting $E(x_i; w) := -\log p(x_i|w)$ leaves us with minimizing the error function $E(D; w) := \sum_{i=1}^{N} E(x_i; w)$ w.r.t. w.
- In case $\log p(D|w)$ is differentiable w.r.t. w we can take the derivative w.r.t. w, set it to zero and solve for w to get w_{ML}.

Example: Maximum Likelihood Estimator for Gaussian

- We have $D = (x_1, \ldots, x_N)$ and $p(x|w) = \mathcal{N}(x|\mu, \sigma^2)$ with $w = (\mu, \sigma^2)$. So:

$$p(D|\mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \prod_{i=1}^{N} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right).$$

- Taking log, putting derivatives to zero and solving for μ and σ^2 we get:

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{and} \quad \sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{\mathrm{ML}})^2.$$

- We get: $\mathbb{E}[\mu_{\mathrm{ML}}] = \mu$ and $\mathbb{E}[\sigma^2_{\mathrm{ML}}] = \frac{N-1}{N}\sigma^2$. So the variance is estimated to low, i.e. we have a Bias.
- $\tilde{\sigma}^2 := \frac{N}{N-1}\sigma^2_{\mathrm{ML}}$ is an unbiased estimator for σ^2.