Machine Learning

Lecture 1.1:
Introduction and probability theory

Sami Brandt

Department of Computer Science
IT University of Copenhagen

Based on slides made by Jes Frellsen.

26 August 2019

IT UNIVERSITY OF COPENHAGEN

Welcome to the

Machine Learning and

Advanced Machine Learning

courses!

Outline of lecture

Course overview

Introduction to Machine Learning

Probability Theory (repetition)

Who are the teachers?

Sami: lecturer

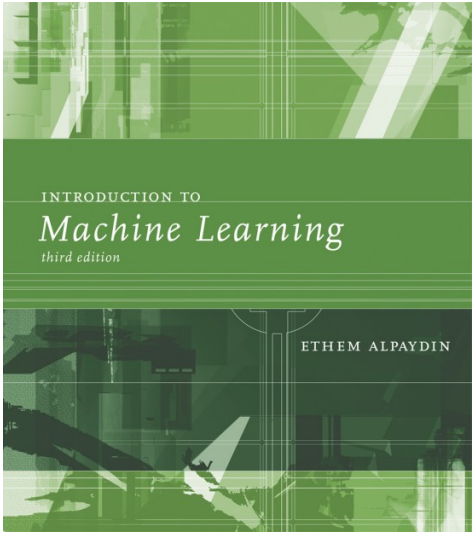
Sanne: lecturer

Stella: lecturer

Najmeh: lecturer

TAs: Stefan (ML), Viktor (ML), and Luis (AML).

Textbooks



Main textbook: “Alpaydin”

- A good textbook for introducing ML
- Fits well to this course
- Not too heavy if compared to e.g. Bishop’s book.

Course structure

- **Weekly activities** in the **first 10 weeks**:
  - **Lecture 1**: Mondays at 12:00–14:00 (Aud 1; except W37 when in Aud3)
  - **Exercises (ML)**: Thursdays at 12:00–14:00 (Anders–Jeppe: 2A52, Ji–Ziming: 2A54)
  - **Exercises (AML)**: Wednesdays at 14:00–16:00 (3A52)
  - **Lecture 2**: Thursdays at 10:00–12:00 (Aud 1; except W37,W40,W45 when in Aud3)
  - **Mandatory Exercises (ML)**: Mondays at 14:00–16:00 (Anders–Jeppe: 2A52, Ji–Ziming: 3A18)
  - **Mandatory Exercises (AML)**: Fridays at 12:00–14:00 (3A52)
- The **last 4 weeks**: project in groups of 2–3 persons.
  - Part of the final exam
  - Will count towards you final grade together with the oral exam
- **Communication** will happen through the **LearnIT** page
  - The **course outline** gives a overview of the course
  - **Separate section for each week** below (detailed reading plan, exercises etc.)
  - **Student forum** for Q&A

Preparation for Weekly Activities

- **How to prepare for weekly activities?**
  - **We do not assume that you have read before the lectures**, but we expect that you have looked at the material.
  - We do not expect you to have solved the exercises before the Wed/Thu exercise session.
  - You have to have solved the exercises before the Fri/Mon mandatory exercise session.

Mandatory activities

- **9 mandatory exercise sessions**
  - You have to **offer** to present at least 50% of the exercises on average during the course.
  - **Before** you come to the mandatory exercises on Mondays (ML) or Fridays (AML) you have to fill in the checklist on the LearnIT page.
  - The TA will select who presents the exercise.
  - *If you have indicated that you can present an exercise and you have not done the exercises you fail this mandatory activity.*
- You need to **pass the mandatory activities to go to the exam**.
- If you have not passed before the project starts: an **oral test**

Exam

The exam exists of two parts:

- A **4 week project** to be done in groups and be handed in by the end of the semester (more information later)
  - ML: Covers many of the subjects that we discussed in the first 10 weeks
  - AML: You will implement a recent machine learning method from a research article.
- A **20 min oral exam** about the project and the topics of the course.

Intended learning outcome

After the course, the student should be able to:

- **Discuss**, clearly **explain**, and **reflect** upon central machine learning concepts and algorithms.
- **Choose among** and **make use** of the most important machine learning approaches in order to apply (match) them to practical problems.
- **Implement** abstractly specified machine learning methods in an imperative programming language (Python)
- **Combine** and **modify** machine learning methods to **analyse** practical dataset and **covey** the results.

Intended learning outcome

After the course, the student should be able to:

- **Discuss**, clearly **explain**, and **reflect** upon central machine learning concepts and algorithms.
  - **Main learning activity:** mandatory exercises,
  - **Main evaluation:** oral exam
- **Choose among** and **make use** of the most important machine learning approaches in order to apply (match) them to practical problems.
- **Implement** abstractly specified machine learning methods in an imperative programming language (Python)
- **Combine** and **modify** machine learning methods to **analyse** practical dataset and **covey** the results.

Intended learning outcome

After the course, the student should be able to:

- **Discuss**, clearly **explain**, and **reflect** upon central machine learning concepts and algorithms.
  - **Main learning activity:** mandatory exercises,
  - **Main evaluation:** oral exam
- **Choose among** and **make use** of the most important machine learning approaches in order to apply (match) them to practical problems.
  - **Main learning activity:** exercises, project
  - **Main evaluation:** project
- **Implement** abstractly specified machine learning methods in an imperative programming language (Python)

- **Combine** and **modify** machine learning methods to **analyse** practical dataset and **covey** the results.

Intended learning outcome

After the course, the student should be able to:

- **Discuss**, clearly **explain**, and **reflect** upon central machine learning concepts and algorithms.
  - **Main learning activity:** mandatory exercises,
  - **Main evaluation:** oral exam
- **Choose among** and **make use** of the most important machine learning approaches in order to apply (match) them to practical problems.
  - **Main learning activity:** exercises, project
  - **Main evaluation:** project
- **Implement** abstractly specified machine learning methods in an imperative programming language (Python)
  - **Main learning activity:** exercises, project,
  - **Main evaluation:** project
- **Combine** and **modify** machine learning methods to **analyse** practical dataset and **covey** the results.

Intended learning outcome

After the course, the student should be able to:

- **Discuss**, clearly **explain**, and **reflect** upon central machine learning concepts and algorithms.
  - **Main learning activity:** mandatory exercises,
  - **Main evaluation:** oral exam
- **Choose among** and **make use** of the most important machine learning approaches in order to apply (match) them to practical problems.
  - **Main learning activity:** exercises, project
  - **Main evaluation:** project
- **Implement** abstractly specified machine learning methods in an imperative programming language (Python)
  - **Main learning activity:** exercises, project,
  - **Main evaluation:** project
- **Combine** and **modify** machine learning methods to **analyse** practical dataset and **covey** the results.
  - **Main learning activity:** project
  - **Main evaluation:** project and oral exam

What is this course not?

- It is not a *deep learning* course
- It is not a Tensorflow / PyTorch / scikit-learn course

Outline of lecture

Course overview

Introduction to Machine Learning

Probability Theory (repetition)

What is *Machine Learning*?

Picture from: https://history-computer.com/ModernComputer/thinkers/Samuel.html

What is *Machine Learning*?

The term was coined by Arthur Lee Samuel in the 1950s.

> "Programming computers to learn from experience should eventually eliminate the need for much of this detailed programming effort."

(Samuel, 1959)

This quote is often loosely rephrased into the definition

> Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed

(e.g. Wikipedia)

Picture from: https://history-computer.com/ModernComputer/thinkers/Samuel.html

What is *Machine Learning*?

The term was coined by Arthur Lee Samuel in the 1950s.

> "Programming computers to learn from experience should eventually eliminate the need for much of this detailed programming effort."

(Samuel, 1959)

This quote is often loosely rephrased into the definition

> Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed

(e.g. Wikipedia)

What is the difference between ML and stats?

- Stats is mainly concerned with estimation
- ML is mainly concerned with prediction

Picture from: https://history-computer.com/ModernComputer/thinkers/Samuel.html

A brief history of machine learning

Bishop (2009)

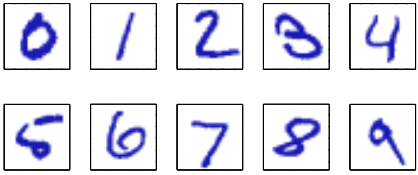
- **First Generation**: Expert systems[1] (1970s and 1980s)
  - Rule based (hand crafted by humans)
  - **Drawback:** Combinatorial explosion
- **Second Generation:** "Black-box" statistical models (1980s and 1990s)
  - Neural networks, Support Vector Machines
  - **Drawback:** Difficult to incorporate domain knowledge
- **Third Generation:** Integration of domain knowledge using the language of probability theory (1990s and 2000s)
  - Bayesian framework, efficient scalable inference algorithms
  - Ability to quantify uncertainties
  - **Drawback:** High computational complexity
- **Forth Generation:** Deep Learning (2010s)
  - Driven by faster computers and larger data sets
  - **Drawback:** Challenging to quantify uncertainties

[1] Aka *Symbolic Artificial Intelligence* or **Good Old-Fashioned Artificial Intelligence**

Example of machine learning: hand-written digit recognition

Task: Given an input image $\mathbf{x} \in \mathbb{R}^{28\times 28}$ produce the identity of the digit $t \in \{0, 1, \ldots, 9\}$.

- $\mathbf{x}$: input or feature vector
- t: target, output or label (vector)



The ML approach: from a **training set**

$$\mathcal{D} = \{(\mathbf{x}_1, t_1), \ldots, (\mathbf{x}_N, t_N)\}$$

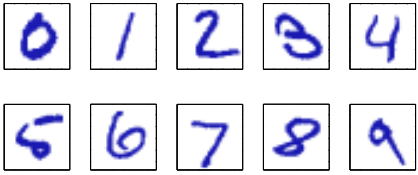
tune the parameters $\mathbf{w}$ of an adaptive model y such that $y_{\mathbf{w}}(\mathbf{x}) = t$.

This is called the **training** or **learning** phase.

Figure(s) from Bishop.

Example of machine learning: hand-written digit recognition

Task: Given an input image $\mathbf{x} \in \mathbb{R}^{28\times 28}$ produce the identity of the digit $t \in \{0, 1, \ldots, 9\}$.



- $\mathbf{x}$: input or feature vector
- t: target, output or label (vector)

The ML approach: from a **training set**

$$\mathcal{D} = \{(\mathbf{x}_1, t_1), \ldots, (\mathbf{x}_N, t_N)\}$$

tune the parameters $\mathbf{w}$ of an adaptive model y such that $y_{\mathbf{w}}(\mathbf{x}) = t$.

This is called the **training** or **learning** phase.

When the model is learned, we can evaluate it on a **test set**.

Generalization: the ability to correctly categorize new examples (not in the training set)

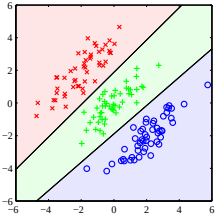
Feature extraction: Pre-processing of the training and test data, e.g. normalization, standardization, dimensionality reduction...

Figure(s) from Bishop.

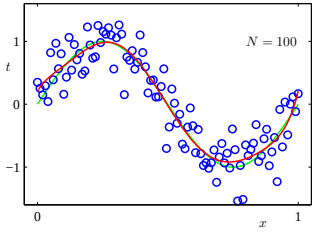
Categories of Machine Learning (I)

Supervised learning: $\mathcal{D} = \{(\mathbf{x}_1, t_1), \ldots, (\mathbf{x}_n, t_n)\}$

Classification

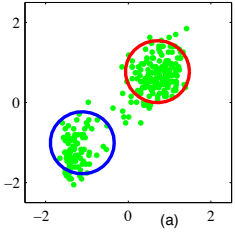


Regression

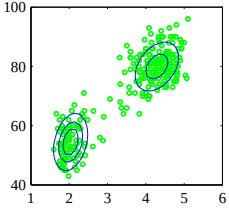


Unsupervised learning: $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$

Clustering



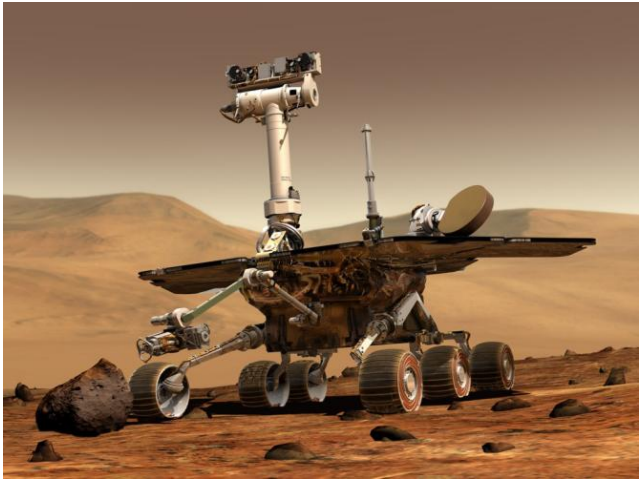
Density estimation



Figure(s) from Bishop.

Categories of Machine Learning (II)

Reinforcement learning: find suitable actions in a given situation to maximize a reward.



Picture from: https://photojournal.jpl.nasa.gov/catalog/PIA04413

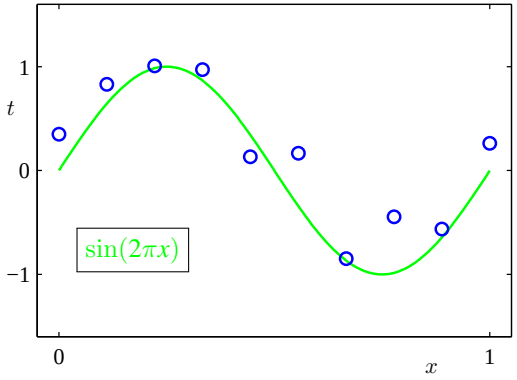
Categories of Machine Learning (II)

Reinforcement learning: find suitable actions in a given situation to maximize a reward.

Picture from: https://medicalfuturist.com/alphago-artificial-intelligence-in-healthcare

Example: Polynomial Regression

Assume we are given a dataset $\mathcal{D} = \{(x_1, t_1), \ldots, (x_N, t_N)\}$ where $x_n, t_n \in \mathbb{R}$ and $N = 10$.



We want to fit the data using a polynomial function

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=1}^{M} w_j x^j$$

Figure(s) from Bishop.

Outline of lecture

Course overview

Introduction to Machine Learning

Probability Theory (repetition)

Sample spaces and outcomes

The **sample space** for an **experiment** is a **set**, where

- each element of the sample space is an **outcome** of the **experiment**, and
- all possible **outcomes** are included in the set.

A **sample space** is typically denoted Ω.

Example (Coin flipping)

The experiment of flipping a coin has two possible outcomes: heads and tails.
The sample space for this experiment is $\Omega = \{\mathsf{H}, \mathsf{T}\}$.

Probability Function

Definition

A **probability function** P on a **finite sample space** Ω assigns to each event A in Ω **a number** $P(A)$ **in** $[0, 1]$ such that

i. $P(\Omega) = 1$, and

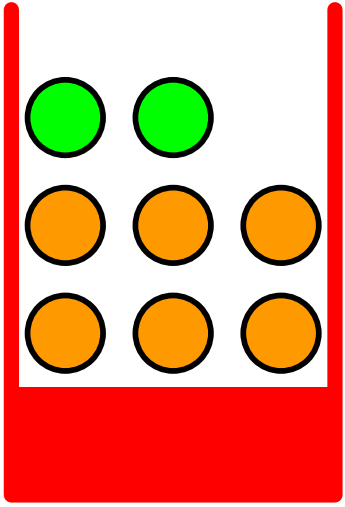
ii. $P(A \cup B) = P(A) + P(B)$ if A and B are **disjoint** (**called the additive property**).

The number $P(A)$ is called the probability that A occurs.

The additive rule implies that

$$P(A \cup B \cup C) = P(A \cup B) + P(C) = P(A) + P(B) + P(C).$$

Frequentist Notion of Probability



Apples and oranges

Frequentist definition of probability:

The probability of an event is the number of times the event occurs out of the total number of trails, as the number of trials goes to infinity.

One trial: Pick a random fruit from a new box containing apples and oranges.

$$P(F = a) = \frac{n_a}{N} \text{ as } N \to \infty$$

Figure(s) from Bishop.

Probability as a measure of uncertainty

We can all agree that the probability of **heads** is $\frac{1}{2}$ for a toss of a fair coin.

Here are to common justifications

1. **Frequency argument:**

$$\text{probability} = \text{relative frequency obtained in a long sequence of tosses,}$$

   i.e. $P(H) = \frac{n_H}{N}$ as $N \to \infty$..

2. **Symmetry or exchangeability argument:**

$$\text{probability} = \frac{\text{number of favorable cases}}{\text{number of possibilities}}$$

   assuming equally likely possibilities, i.e. $P(H) = \frac{1}{2}$.

The frequents argument involves the hypothetical notion of a long sequence of experiments.

The symmetry argument allow us to assign probability to a singe experiment.

Picture from gurmeet.net

Bayesian Notion of Probability

In the Bayesian statistics, probability is used as a measure of uncertainty.

What is the probability of Denmark winning the World Cup in 2022?

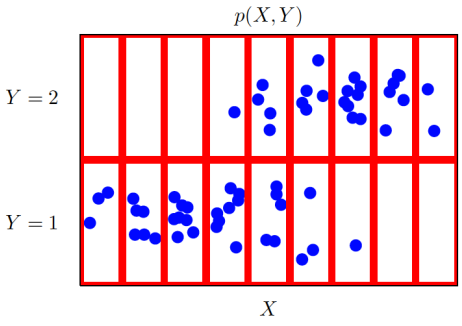
How uncertain am I that Denmark will win the World Cup?

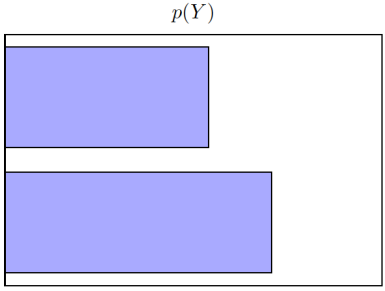
What is the probability that the Arctic ice cap will have melted by the end of the century?

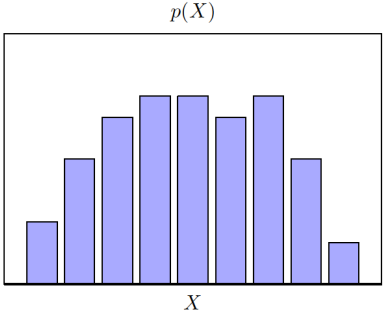
How uncertain am I about the rate of ice loss?

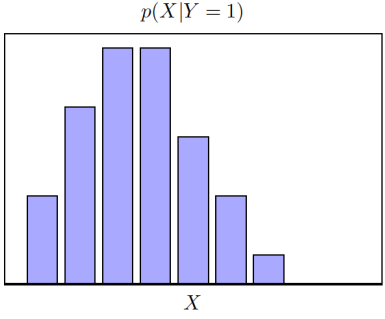
Pictures from wikipedia.org and ecochiccayman.com

Distribution of Two Variables, Discrete





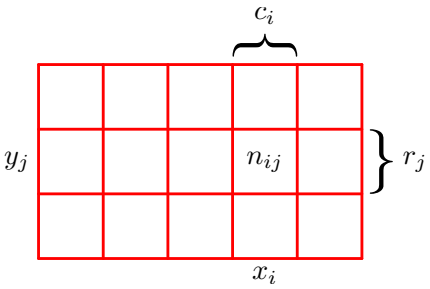




Probability theory

Joint probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$



Figure(s) from Bishop.

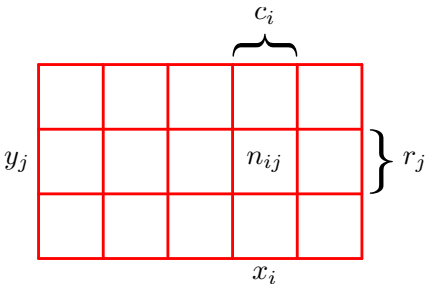
Probability theory

Joint probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Marginal probability

$$p(X = x_i) = \frac{c_i}{N}$$



Figure(s) from Bishop.

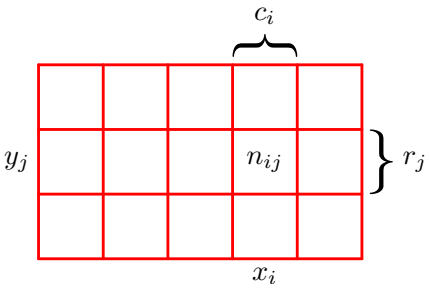
Probability theory

Joint probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Marginal probability

$$p(X = x_i) = \frac{c_i}{N} = \frac{\sum_j n_{ij}}{N}$$



Figure(s) from Bishop.

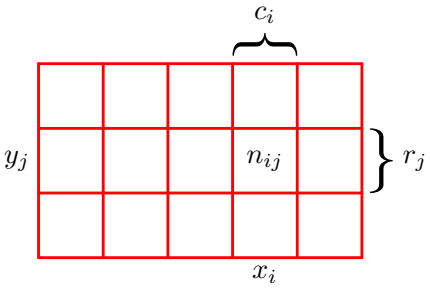
Probability theory

Joint probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Marginal probability

$$p(X = x_i) = \frac{c_i}{N} = \frac{\sum_j n_{ij}}{N} = \sum_j p(X = x_i, Y = y_j)$$



Figure(s) from Bishop.

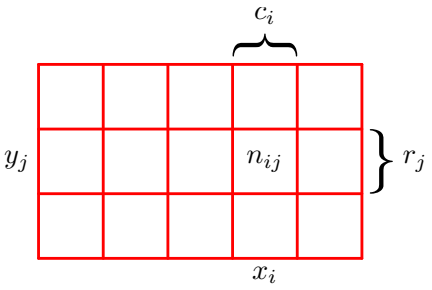
Probability theory

Joint probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Marginal probability

$$p(X = x_i) = \frac{c_i}{N} = \frac{\sum_j n_{ij}}{N} = \sum_j p(X = x_i, Y = y_j) \quad \textbf{Sum rule}$$



Figure(s) from Bishop.

Probability theory

Joint probability

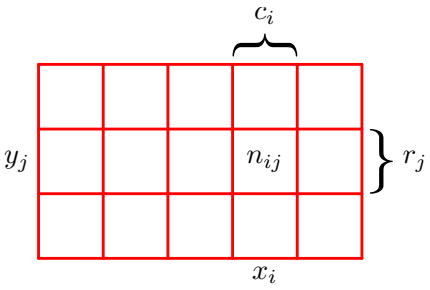
$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Marginal probability

$$p(X = x_i) = \frac{c_i}{N} = \frac{\sum_j n_{ij}}{N} = \sum_j p(X = x_i, Y = y_j) \quad \textbf{Sum rule}$$

Conditional probability

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$



Figure(s) from Bishop.

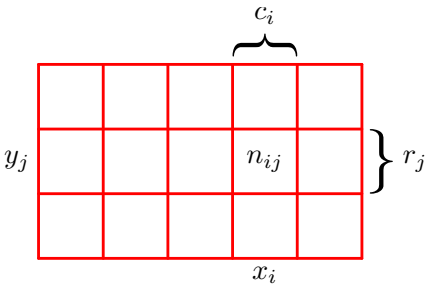
Probability theory

Joint probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Marginal probability

$$p(X = x_i) = \frac{c_i}{N} = \frac{\sum_j n_{ij}}{N} = \sum_j p(X = x_i, Y = y_j) \quad \textbf{Sum rule}$$



Conditional probability

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

We can then derive the **product rule**

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

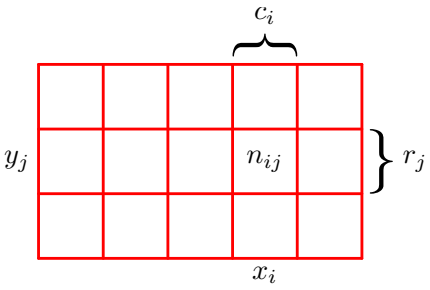
Probability theory

Joint probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Marginal probability

$$p(X = x_i) = \frac{c_i}{N} = \frac{\sum_j n_{ij}}{N} = \sum_j p(X = x_i, Y = y_j) \quad \textbf{Sum rule}$$



Conditional probability

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

We can then derive the **product rule**

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N}$$

Figure(s) from Bishop.

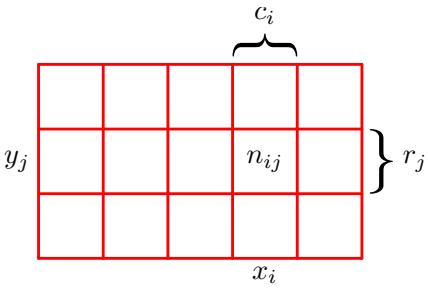
Probability theory

Joint probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Marginal probability

$$p(X = x_i) = \frac{c_i}{N} = \frac{\sum_j n_{ij}}{N} = \sum_j p(X = x_i, Y = y_j) \quad \textbf{Sum rule}$$



Conditional probability

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

We can then derive the **product rule**

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N}$$
$$= p(Y = y_j | X = x_i) P(X = x_i)$$

Figure(s) from Bishop.

The rules of probability and Bayes theorem

The Rules of Probability

sum rule
$$p(X) = \sum_Y p(X, Y)$$

product rule
$$p(X, Y) = p(Y|X)p(X)$$

From these rules follows:

Bayes' theorem

$$p(Y|X) = \frac{p(X|Y)p(Y)}{P(X)} \quad \text{where} \quad p(X) = \sum_Y p(X|Y)p(Y)$$

Continuous case: Probability densities

A **probability density** $p(x)$ over $x \in \mathbb{R}$ must satisfy

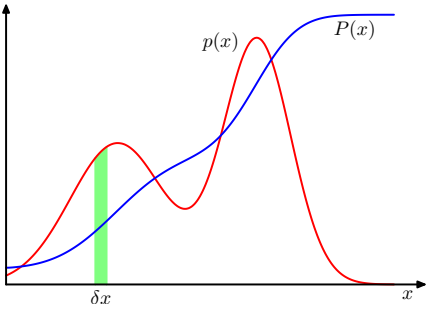
$$p(x) \geq 0 \quad \text{and} \quad \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$

The probability of x lying on the interval (a, b) is

$$p(x \in (a, b)) = \int_{a}^{b} p(x)\,\mathrm{d}x$$

The **cumulative distribution function** (CDF) is defined as

$$P(z) = p(x \in (-\infty, z)) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x$$



The **sum and product rules** also apply to densities

$$p(x) = \int p(x, y)\,\mathrm{d}y \quad \text{and} \quad p(x, y) = p(y|x)p(x)$$

Figure(s) from Bishop.

Expectations

The **expected value** of a function $f(x)$ under a discrete probability mass function or continuous density, both denoted by $p(x)$,

$$\mathrm{E}[f] \equiv \sum_x p(x) f(x) \qquad \text{or} \qquad \mathrm{E}[f] \equiv \int_x p(x) f(x).$$

The **conditional expectation** can be written as

$$\mathrm{E}_x[f|y] \equiv \sum_x p(x|y) f(x).$$

We can approximate the expected value by using **samples** of x drawn from $p(x)$, i.e,

$$\mathrm{E}[f] \simeq \frac{1}{N} \sum_{n=1}^{N} f(x_n).$$

Variance and covariance

The **variance** of $f(x)$ (under $p(x)$) is defined as

$$\begin{aligned}\mathrm{var}[f] &\equiv \mathrm{E}\left[(f(x) - \mathrm{E}[f(x)])^2\right] \\ &= \mathrm{E}[f(x)^2] - \mathrm{E}[f(x)]^2,\end{aligned}$$

and measures the variability of $f(x)$ around its mean.

The **covariance** of two random variables x and y is defined as

$$\begin{aligned}\mathrm{cov}[x, y] &\equiv \mathrm{E}_{x,y}\left[(x - \mathrm{E}[x])(y - \mathrm{E}[y])\right] \\ &= \mathrm{E}[xy] - \mathrm{E}[x]\mathrm{E}[y],\end{aligned}$$

and measures the extend to which x and y vary together

Examples of Probability Distributions

We will review two **parametric distributions**, that are used as a building blocks in machine learning methods:

- Binomial distribution (Discrete)
- Gaussian distribution (Continous)

A distribution is said to be parametric, if it can be described using **a finite set of parameters**.

Binomial distribution

Figure(s) from Bishop.

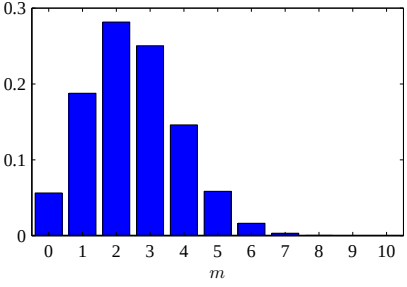
Now suppose that we perform N **Bernoulli trials**, e.g. $N = 10$ coin tosses.

What is the probability of obtaining m **successes in** N **trials**?
E.g. $m = 3$ heads in $N = 10$ trials.

$$\mathrm{Bin}(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m}$$
$$\mathrm{E}[E] = N\mu$$
$$\mathrm{var}[m] = N\mu(1-\mu)$$

Picture from gurmeet.net

$\mathrm{Bin}(m|N=10,\mu=0.25)$

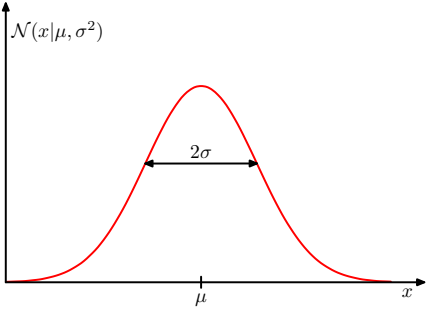


The Gaussian distribution

The **Gaussian** is widely used in ML and is defined by

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

where μ is the **mean** and σ^2 is the **variance**.



Law of Large Numbers

Law of Large Numbers

The law of large numbers says that the mean of i.i.d. samples

$$\bar{x}_n = \frac{1}{n}\left(x_1 + x_2 + \ldots + x_n\right)$$

converges to the expected value, or

$$\bar{x}_n \longrightarrow E\{\bar{X}_n\}, \qquad \text{as } n \to \infty.$$

(What are the required conditions?)

Central limit theorem

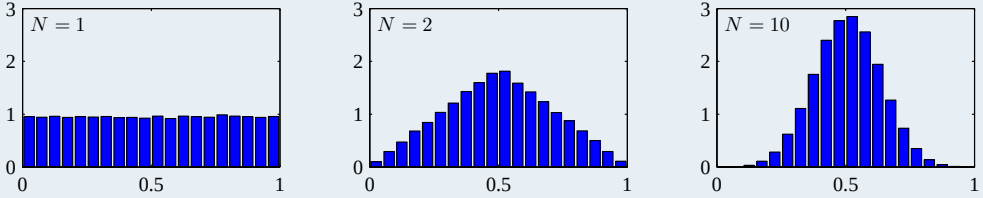
Figure(s) from Bishop.

Central limit theorem

The distribution of the sum of N i.i.d. random variables **becomes increasingly Gaussian as N grows**.

N uniform $[0, 1]$ random variables

The distribution of the mean $(x_1 + \cdots + x_N)/N$ for the N variables $x_1, \ldots, x_N \sim \mathcal{U}(0, 1)$:



Figure(s) from Bishop.

Next lecture

- Introduction to Supervised learning

References I

- Alpaydin, E. (2014). *Introduction to Machine Learning*. Third Edition. The MIT Press.
- Bishop, C. (2009). *Introduction to Bayesian inference*. URL: http://videolectures.net/mlss09uk_bishop_ibi/.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Samuel, A. L. (1959). “Some studies in machine learning using the game of checkers”. In: *IBM Journal of research and development* 3.3, pp. 210–229.