TECHNISCHE UNIVERSITÄT DARMSTADT

Statistical Machine Learning

Lecture 1 b: Statistics

Simone Schaub-Meyer
Department of Computer Science
TU Darmstadt

Summer Term 2025

Today's Objectives

- Refresher of statistics & probabilities.
- Focus on the Gaussian distribution and the exponential family.
- Notion of information and entropy.

Check also
www.statlect.com/probability-distributions/

Outline

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

Random Variables

- What is a random variable?
  - Is a random number determined by chance – i.e., its value is unknown and/or could change
    - e.g., the temperature outside the room at the current time
  - More formally, drawn according to a probability distribution
  - Typical random variables in statistical learning: input data, output data, noise
- What is a probability distribution?
  - Describes the probability (mass / density) that the random variable will be equal to a certain value.
  - The probability distribution can be given by the physics of an experiment (e.g., throwing dice)

Random Variables

- Important concept: The data generating model
  - E.g., what is the data generating model for:
    i) throwing dice
    ii) regression
    iii) classification
    iv) visual perception?

Discrete / Continuous Random Variables

Let $\mathcal{X}$ denote the set of possible values that a random variable X can take, i.e., the **sample space** or **state space**.

- **Discrete random variable** i.e. $\mathcal{X}$ is finite (or countably finite).
  - We define the probability mass function (pmf) as the probability that X would be *equal* to a sample x.

$$p(x) = P(X = x) \text{ with } 0 \leq p(x) \leq 1 \text{ and } \sum_{x \in \mathcal{X}} p(x) = 1$$

- **Continuous random variable** i.e. $\mathcal{X}$ is infinite and uncountable.
  - We define the probability density function (pdf) or density as the *relative likelihood* that X would be *close* to a sample x.

$$p(x), \text{ with } p(x) \geq 0 \text{ and } \int_{\mathcal{X}} p(x) dx = 1$$

.

Cumulative distribution function

Let $\mathcal{X} = \mathbb{R}$.

Cumulative distribution function (cdf) is an increasing differentiable function F mapping $\mathbb{R}$ to $[0, 1]$ with $F(-\infty) = 0$ and $F(+\infty) = 1$.

- $F_X(x) = P(X \leq x) = \int_{-\infty}^{x} p(x')dx'$
- $p(x) = \frac{\mathrm{d}F_X}{\mathrm{d}x}(x)$

A good way to sample ANY random variable from a uniform distribution!

Outline

2. Basic Rules of Probability

- Joint distribution: $p(x, y)$
- Marginal distribution: $p(y) = \int p(x, y) dx$
- Conditional distribution: $p(y|x) = \frac{p(x,y)}{p(x)}$, if $p(x) > 0$
- Chain rule of probabilities

$$\begin{aligned} p(x_1, \ldots, x_n) &= p(x_1|x_2, \ldots, x_n) p(x_2, \ldots, x_n) \\ &= p(x_1|x_2, \ldots, x_n) p(x_2|x_3, \ldots, x_n) \ldots p(x_{n-1}|x_n) p(x_n) \end{aligned}$$

TECHNISCHE UNIVERSITÄT DARMSTADT

Bayes Rule

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

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

Outline

Expectations

- **Expectation**: The average value of some function $f(x)$ under a probability distribution $p(\cdot)$:

$$\mathbb{E}_{X\sim p(\cdot)}[f(X)] = \mathbb{E}_X[f] = \mathbb{E}[f] = \begin{cases} \sum_x p(x)f(x) & \text{discrete case} \\ \int p(x)f(x)\mathrm{d}x & \text{continuous case} \end{cases}$$

We note $\mu = \mathbb{E}[X]$.

- **Conditional Expectation**:

$$\mathbb{E}_{X\sim p(\cdot|y)}[f(X)] = \mathbb{E}_X[f|y] = \begin{cases} \sum_x p(x|y)f(x) & \text{discrete case} \\ \int p(x|y)f(x)\mathrm{d}x & \text{continuous case} \end{cases}$$

TECHNISCHE UNIVERSITÄT DARMSTADT

Expectations

- Approximate Expectation

$$\mathbb{E}[f] = \int f(x)p(x)\mathrm{d}x \approx \frac{1}{N}\sum_{n=1}^{N} f(x_n)$$

When there is no analytical solution, we can use this to approximate integrals by sampling!

- $X \mapsto \mathbb{E}[X]$ is a linear function i.e. $\mathbb{E}[\alpha X + Y] = \alpha\mathbb{E}[X] + \mathbb{E}[Y]$.
  In general: $\mathbb{E}[g(X)] \neq g\left(\mathbb{E}[X]\right)$

TECHNISCHE UNIVERSITÄT DARMSTADT

Variance and Covariance

- **Variances** give a measure of dispersion - the expected spread of the variable in relation to its mean

$$\mathrm{var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$$
$$\mathrm{std}[X] = \sqrt{\mathrm{var}[X]} = \sigma$$

- **Covariances** give a measure of correlation - how much two variables change together

$$\begin{aligned}\mathrm{cov}[X, Y] &= \mathbb{E}_{X,Y}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right] \\ &= \mathbb{E}_{X,Y}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$

Variance and Covariance

- Note the very important rule:

$$\mathbb{E}\left[\mathbf{X}\mathbf{X}^T\right] = \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{X}]^T + \text{cov}[\mathbf{X}, \mathbf{X}]$$
$$= \boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma}$$

TECHNISCHE UNIVERSITÄT DARMSTADT

Moments of Random Variables

- Definition of a Moment

  $m_n = \mathbb{E}\left[x^n\right]$

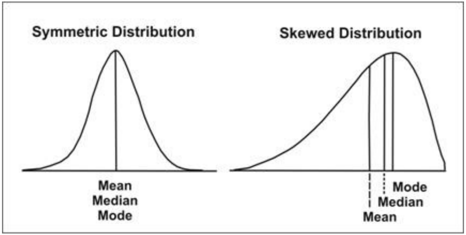
- Definition of a Central Moment (deviation from the mean)

  $cm_n = \mathbb{E}\left[(x-\mu)^n\right]$

- cm_2: variance

- cm_3: skewness (measure of asymmetry)

- cm_4: kurtosis (measure of heavy tailed-ness and light tailed-ness)



Quiz

When there is no analytical solution to compute the expectation, what can we do?

- Approximate Expectation:

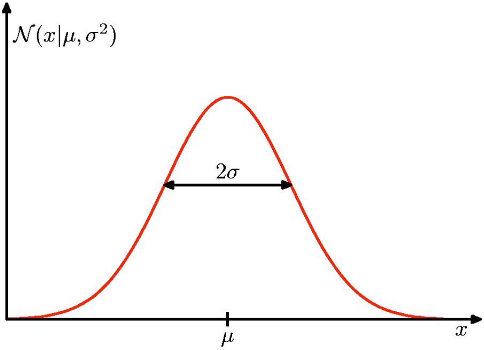
$$\mathbb{E}[f] = \int f(x)p(x)\mathrm{d}x \approx \frac{1}{N}\sum_{n=1}^{N} f(x_n)$$

Also know as Monte-Carlo sampling.

TECHNISCHE UNIVERSITÄT DARMSTADT

Outline

The Gaussian Distribution



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

TECHNISCHE UNIVERSITÄT DARMSTADT

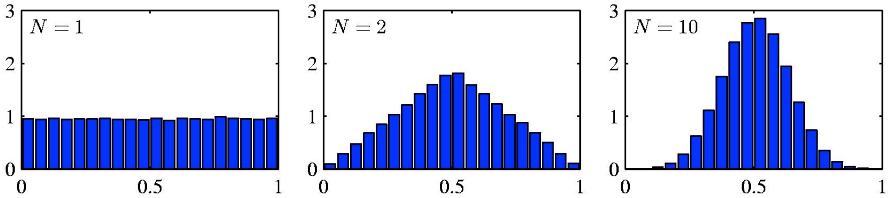
Central Limit Theorem

- The distribution of the sum of N i.i.d. (independent and identically distributed) random variables becomes increasingly Gaussian as N grows.
- Application: What would be the "shape" of the mean of samples drawn from ANY random variable?
  $\rightarrow$ Gaussian, if we draw enough samples.

This is why Gaussians are SO important!

Central Limit Theorem

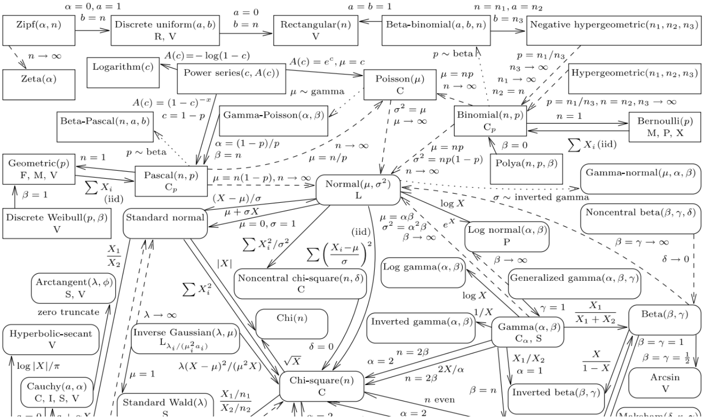
- Example: *N* uniform [0,1] random variables



- Gaussians are often a *good* model of data
- Working with Gaussians leads to analytic solutions for complex operations

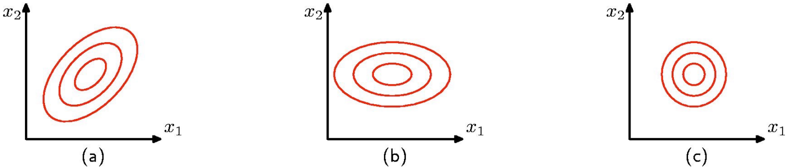
TECHNISCHE UNIVERSITÄT DARMSTADT

Distributions' landscape



Taken from http://www.math.wm.edu/ leemis/2008amstat.pdf.

The Multivariate Gaussian Distribution



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

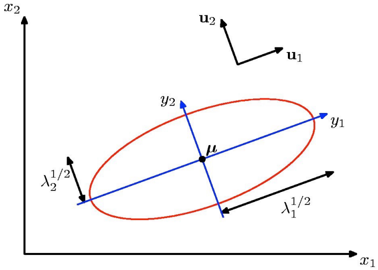
The Multivariate Gaussian Distribution

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

- **To clear some confusion**: for a chosen vector $\mathbf{x}$, $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a real number indicating the relative likelihood of $\mathbf{X}$ to be close to $\mathbf{x}$. The mean $\boldsymbol{\mu}$ is just a specific vector amongst all the possible vectors. The covariance matrix $\boldsymbol{\Sigma}$ tells us how two dimensions of a vector are related to each other.

Geometry of the Multivariate Gaussian

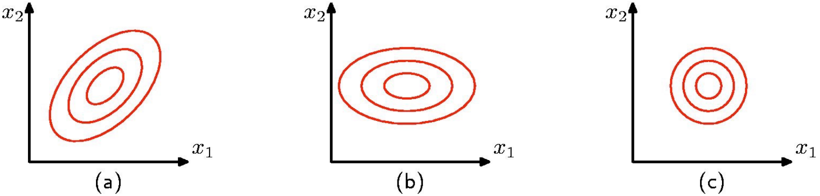
$$
\begin{aligned}
\Delta^2 &= (\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \\
\boldsymbol{\Sigma}^{-1} &= \sum_{i=1}^{D}\frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^{\mathsf{T}} \\
\Delta^2 &= \sum_{i=1}^{D}\frac{y_i^2}{\lambda_i} \\
y_i &= \mathbf{u}_i^{\mathsf{T}}(\mathbf{x}-\boldsymbol{\mu})
\end{aligned}
$$



Δ is the Mahalanobis distance.

Moments of the Multivariate Gaussian

$$\mathrm{var}[\mathbf{X}] = \mathrm{cov}[\mathbf{X}, \mathbf{X}] = \boldsymbol{\Sigma}$$



TECHNISCHE UNIVERSITÄT DARMSTADT

Partitioned Gaussian Distributions

- We partition $\mathbf{x}$ into two disjoint subsets

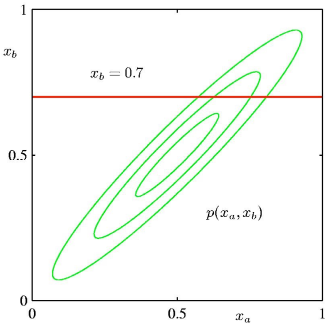
$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$$

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}$$

$$\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1} \qquad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$$

$\boldsymbol{\Lambda}$ is the precision matrix.

Partitioned Conditionals



TECHNISCHE UNIVERSITÄT DARMSTADT

Partitioned Conditionals and Marginals

If the joint distribution $p(\mathbf{x}_a, \mathbf{x}_b)$ is Gaussian then:

- the conditional distributions $p(\mathbf{x}_a|\mathbf{x}_b)$ and $p(\mathbf{x}_b|\mathbf{x}_a)$ are also Gaussians.
- the marginal distributions $p(\mathbf{x}_a)$ and $p(\mathbf{x}_b)$ are also Gaussians.

Manipulating Gaussians

- Converting Marginal $p(\mathbf{x})$ and Conditional $p(\mathbf{y}|\mathbf{x})$ to Joint Distribution $p(\mathbf{x}, \mathbf{y})$:

$$\underbrace{\mathcal{N}(\mathbf{x}|\mathbf{a}, \mathbf{A})}_{p(\mathbf{x})} \underbrace{\mathcal{N}(\mathbf{y}|\mathbf{b} + \mathbf{F}\mathbf{x}, \mathbf{B})}_{p(\mathbf{y}|\mathbf{x})} = \underbrace{\mathcal{N}\left(\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \middle| \begin{bmatrix} \mathbf{a} \\ \mathbf{b} + \mathbf{F}\mathbf{a} \end{bmatrix}, \begin{bmatrix} \mathbf{A} & \mathbf{A}^\mathsf{T}\mathbf{F}^\mathsf{T} \\ \mathbf{F}\mathbf{A} & \mathbf{B} + \mathbf{F}\mathbf{A}^\mathsf{T}\mathbf{F}^\mathsf{T} \end{bmatrix} \right)}_{p(\mathbf{x},\mathbf{y})}$$

- Converting Joint Distribution $p(\mathbf{x}, \mathbf{y})$ to Marginal $p(\mathbf{x})$ and Conditional $p(\mathbf{y}|\mathbf{x})$

$$\underbrace{\mathcal{N}\left(\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \middle| \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^\mathsf{T} & \mathbf{B} \end{bmatrix} \right)}_{p(\mathbf{x},\mathbf{y})} = \underbrace{\mathcal{N}(\mathbf{x}|\mathbf{a}, \mathbf{A})}_{p(\mathbf{x})} \underbrace{\mathcal{N}(\mathbf{y}|\mathbf{b} + \mathbf{C}^\mathsf{T}\mathbf{A}^{-1}(\mathbf{x} - \mathbf{a}), \mathbf{B} - \mathbf{C}^\mathsf{T}\mathbf{A}^{-1}\mathbf{C})}_{p(y|x)}$$

TECHNISCHE UNIVERSITÄT DARMSTADT

Quiz

Why are Gaussian distributions important?

- Central Limit Theorem.
- they are linked with many common distributions.
- they ease computations.

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

5. Exponential Family

- All distributions from this family are uni-modal

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x})\right\}$$

  where $\boldsymbol{\eta}$ is the natural parameter and

$$g(\boldsymbol{\eta})\int h(\mathbf{x})\exp\left\{\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x})\right\}\mathrm{d}\mathbf{x} = 1$$

  hence g can be interpreted as a normalization coefficient.

- The exponential family is a large class of distributions that are all analytically appealing, because taking the $\log$ of them decomposes them into simple terms.

Exponential Family - Example

The Bernoulli Distribution:

$$
\begin{aligned}
p(x|\mu) = \mu^x(1-\mu)^{1-x} &= \exp(x \ln \mu + (1-x)\ln(1-\mu)) \\
&= (1-\mu)\exp\left(\ln\left(\frac{\mu}{1-\mu}\right)x\right) \\
&= \frac{1}{1+\exp(\eta)}\exp(\eta x) \\
&= h(x)g(\eta)\exp(\eta u(x))
\end{aligned}
$$

where $\eta = \ln\left(\frac{\mu}{1-\mu}\right)$, $h(x) = 1$, $g(\eta) = \frac{1}{1+\exp(\eta)}$ and $u(x) = x$.

TECHNISCHE UNIVERSITÄT DARMSTADT

Outline

Information Theory - Core Questions

How can we represent information compactly, i.e., using as few bits as possible?

- Compressing text with GZIP, pictures in JPEG, movies in MPEG or sound in MP3.

How can we transmit or store data reliably?

- ECC memory
- Error Correction on CDs
- Communication with space probes

Information Theory - Core Questions

Machine Learning Questions:

- How can we measure complexity?
- How can we measure “distances” between probability distributions?
- How can we reconstruct data?

TECHNISCHE UNIVERSITÄT DARMSTADT

What is Information?

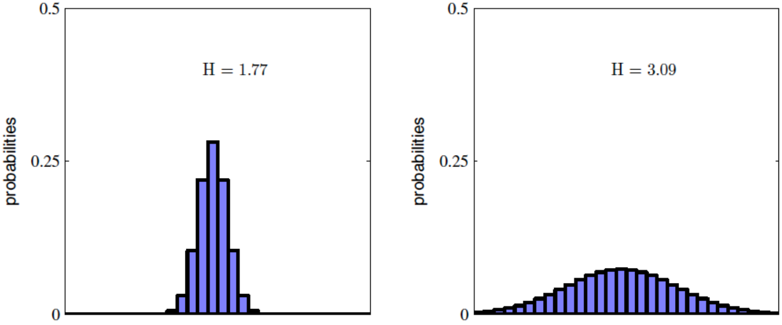
i	a_i	p_i
1	a	.0575
2	b	.0128
3	c	.0263
4	d	.0285
5	e	.0913
6	f	.0173
7	g	.0133
8	h	.0313
9	i	.0599
10	j	.0006
11	k	.0084
12	l	.0335
13	m	.0235
14	n	.0596
15	o	.0689
16	p	.0192
17	q	.0008
18	r	.0508
19	s	.0567
20	t	.0706
21	u	.0334
22	v	.0069
23	w	.0119
24	x	.0073
25	y	.0164
26	z	.0007
27	-	.1928

- All letters in the English alphabet have a very different probability p_i of occurring.
- N bits can encode 2^N characters. The alphabet can be encoded in $\lceil \log_2 27 \rceil \approx \lceil 4.75 \rceil = 5$ bits.
- We define the information in a single character as $h(p_i) = -\log_2 p_i$. Events with a low probability correspond to high information content.
- The *average* information in a character in an English text is $H(p) = \mathbb{E}[h(.)] = -\sum_i p_i \log_2(p_i) \approx 4.1$. This quantity is called the *entropy*. On average, with the right encoding, we can represent each letter with 4.1 bits instead of 4.7.

What is information

- Claude Shannon considered information as a *message* sent by a sender to a receiver, i.e., he wanted to solve the problem of *how to best encode information* that a sender wished to transmit to a receiver
- Shannon gave information a mathematical value based on *probability* defined in terms of the concept of information **entropy** more commonly known as **Shannon entropy**.
- Information is defined as the measure of the increase of uncertainty for a receiver
- Entropy quantifies the amount of uncertainty involved in the value of a random variable or the outcome of a random process
- E.g., identifying the outcome of a **fair coin** flip provides less information (**lower entropy**) than specifying the outcome from a roll of a dice. Indeed, $-\ln(\frac{1}{2}) < -\ln(\frac{1}{6})$.

Entropy of Distributions



What is the “difference” between these distributions?

TECHNISCHE UNIVERSITÄT DARMSTADT

Kullback-Leibler Divergence

- The Kullback-Leibler Divergence - KL Divergence - is a similarity measure between two distributions, and is defined as:

$$\mathrm{KL}(p||q) = -\int p(x) \ln q(x) \mathrm{d}x - \left(-\int p(x) \ln p(x) \mathrm{d}x \right)$$
$$= -\int p(x) \ln \frac{q(x)}{p(x)} \mathrm{d}x$$

- It represents the average additional amount of extra bits required to specify a symbol X, given that its underlying probability distribution is the estimated $q(x)$ and not the true one $p(x)$

TECHNISCHE UNIVERSITÄT DARMSTADT

Kullback-Leibler Divergence

- It is not a distance: $\mathrm{KL}(p||q) \neq \mathrm{KL}(q||p)$
- It is non-negative: $\mathrm{KL}(p||q) \geq 0$ with equality iff $\forall x, p(x) = q(x)$

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

7. Wrap-Up

You know now:

- What random variables are (both continuous and discrete)
- What probability distributions are
- What expectation and variance are
- What a Gaussian distribution is and why it is so important
- What information and entropy are
- How to measure the similarity between two probability distributions

Self-Test Questions

- What is a random variable?
- What is a distribution?
- What is a Gaussian distribution?
- What is an expectation?
- What is a joint distribution?
- What is a conditional distribution?
- What is a distribution with a lot of information?
- How to measure the difference between distributions?

Homework

Reading Assignment for next lecture:

- Bishop ch 1.5
- Murphy (2021) ch. 5