Statistical Machine Learning
Basics of probability and statistics

Luca Bortolussi

Data Science and Scientific Computing

1 Introduction

Machine learning - the perspective that we take on it at least - is all about learning probability models:

- Full data generating distribution $p(x, y)$
- Input data generating distribution $p(x)$
- Output conditional distribution $p(y \mid x)$
- Input conditional distribution $p(x \mid y)$

First goal is to recall what are these models, and how we can manipulate them and reason about them.

2 Basics of probability theory

2.1 Probability: an intuition

- Probabilities are a mathematical tool to describe *uncertain* phenomena.
- Uncertainty is present either because the phenomenon is intrinsically random, like in quantum mechanics (*aleatoric uncertainty*) or because of our incomplete knowledge about the phenomenon (*epistemic uncertainty*).
- The probability p_i of an experiment taking a certain value i is the frequency with which that value is taken in the limit of infinite experimental trials (*frequentist viewpoint* - aleatoric uncertainty)
- Alternatively, we can take probability to be our belief that a certain value will be taken (*Bayesian viewpoint* - epistemic uncertainty)

2.2 Random Variables and probability distributions

- Random variables: results of non exactly reproducible experiments
- Let x and y be two random variables, $p(x = i, y = j)$ is the *joint probability* of x taking value i and y taking value j (with i and j in the respective spaces of possible values).
- Often just written $p(x, y)$ to indicate the function (as opposed to its evaluation over the outcomes i and j).
- $p(x|y)$ is the conditional probability, i.e. the probability of x if you know y has a certain value
- Example: a discrete random variable, with values in a countable state space S, e.g. $\mathbb{N}$, like the Poisson random variable.
- Example: real-valued random variables, with values in $\mathbb{R}$, like the Gaussian random variable.

2.2.1 Rules of probability (discrete rv)

- *Normalisation*: the sum of the probabilities of all possible experimental outcomes must be 1, $\sum_{x \in \mathcal{X}} p(x) = 1$
- *Sum rule*: the marginal probability $p(x)$ is given by summing the joint $p(x, y)$ over all possible values of y,

$$p(x) = \sum_{y \in \mathcal{Y}} p(x, y)$$

- *Product rule*: the joint is the product of the conditional and the marginal, $p(x, y) = p(x|y)p(y)$
- *Bayes rule*: the posterior is the ratio of the joint and the marginal

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

2.2.2 Independence

- Two random variables x and y are *independent* if their joint probability factorises in terms of marginals

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

- Using the product rule, this is equivalent to the conditional being equal to the marginal

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

2.2.3 Continuous random variables

- If the state space $\mathcal{X}$ is continuous some of the previous definitions must be modified
- The general case is mathematically difficult; we restrict ourselves to $\mathcal{X} = \mathbb{R}^n$ and to distributions which admit a *density*, i.e. a function

$$p : \mathcal{X} \to \mathbb{R} \quad \text{s.t.} \quad p(x) \geq 0, \ \forall x \quad \text{and} \quad \int_{\mathcal{X}} p(x)dx = 1$$

- Formally, these are absolute continuous measures with respect to the Lebesgue measure, with density function playing the role of a Radon-Nikodym derivative.
- It can be shown that the rules of probability distributions hold also for probability densities
- Notice that $p(x)$ is NOT the probability of the random variable being in state x (that is always zero for bounded densities); probabilities are only defined as integrals over subsets of $\mathcal{X}$

2.3 Distributions and expectations

- A probability distribution for finite state space can be given by a table, in general is given by a functional form
- Probability distributions (over numerical objects) are useful to compute expectations of functions

$$\langle f \rangle = \sum_{x \in \mathcal{X}} f(x)p(x)$$

- Important expectations are the *mean* $\langle x \rangle$ and *variance* $var(x) = \langle (x - \langle x \rangle)^2 \rangle$.
- For more variables, also the *covariance* $cov(x, y) = \langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle$ or its scaled relative the *correlation* $corr(x, y) = cov(x, y)/\sqrt{var(x)var(y)}$

Exercise

If two variables are independent, then their correlation is zero. **NOT TRUE** viceversa (no correlation does not imply independence)

2.3.1 Computing expectations

- If you know analytically the probability distribution and can compute the sums (integrals), no problem
- If you know the distribution but cannot compute the sums (integrals), enter the magical realm of approximate inference (fun but out of scope)
- If you know nothing but have N_S samples, then use a sample approximation
- Approximate the probability of an outcome with the *frequency* in the sample

$$\langle f(x) \rangle \simeq \sum_x \frac{n_x}{N_S} f(x) = \frac{1}{N_S} \sum_{i=1}^{N_S} f(x_i)$$

3 Formal definition of probability

In this section we introduce probabilities from the more formally correct point of view of sigma algebras. The problem is that, for a continuous random variable, not every subset of real numbers can have a probability attached to it. Sets must be well-behaved, and this is captured by the notion of sigma algebra.

3.1 Sigma Algebras

Let Ω be a set, $\mathcal{S} \subseteq 2^{\Omega}$ is a σ-algebra iff

1. $\emptyset, \Omega \in \mathcal{S}$;
2. $A \in \mathcal{S} \Rightarrow A^c \in \mathcal{S}$;
3. $A_n \in \mathcal{S},\ n \in \mathbb{N} \Rightarrow \bigcup_n A_n \in \mathcal{S}$;

$(\Omega, \mathcal{S})$ is called *measurable space*. Example: the Borel sigma algebra $\mathcal{B}$ in $\mathbb{R}^n$, the smallest σ-algebra containing all open sets.

3.2 Measurable function

A function $f : (X, \mathcal{A}) \to (Y, \mathcal{B})$ is measurable iff $f^{-1}(B) \in \mathcal{A}$ for each $B \in \mathcal{B}$

3.3 Probability measure

Let $(\Omega, \mathcal{S})$ be a measurable space. A probability measure on $(\Omega, \mathcal{S})$ is a function $\mu : \mathcal{S} \to [0, 1]$ such that

1. $\mu(\emptyset) = 0$
2. $\mu(A^c) = 1 - \mu(A)$
3. If $A_n \in \mathcal{S}$ disjoint, then $\mu(\bigcup_n A_n) = \sum_{n=0}^{\infty} \mu(A_n)$

3.4 Probability space

$(\Omega, \mathcal{S}, \mu)$, with $\mathcal{S}$ σ-algebra and μ probability measure on $(\Omega, \mathcal{S})$, is a probability space.

3.5 Random Variables

- Let $(\Omega, \mathcal{S}, \mu)$ be a probability space (the *sample space*) and $(\mathcal{X}, \mathcal{A})$ be a measurable space. A measurable function $x : (\Omega, \mathcal{S}) \to (\mathcal{X}, \mathcal{A})$ is called a *random variable*.
- The law of x is $\mathbb{P}\{x \in A\} = \mu(x^{-1}(A))$, for each $A \in \mathcal{A}$, and it is a probability distribution in $(\mathcal{X}, \mathcal{A})$.
- Example: discrete random variables, with values in a countable state space S, with the σ-algebra 2^S.

- Example: real-valued random variables, with values in $\mathbb{R}$, with the Borel σ-algebra.

4 Probabilistic inference

- In logics, an inference system is given by a set of inference rules, allowing to infer logical consequences from a set of facts/ axioms.
- The rules of probability define an inference system generalising logical ones to reason under uncertainty.
- Typically, we have a probabilistic model, and possibly evidence (e.g. experimental observations), and we want to deduce consequences – here compute probabilities.
- We do this by consistent applications of the rules of probability.

Example

- Scientists[1] found that people that enjoy working 14 hours per day (HW) almost inevitably eat Frico (F): $p(F|HW) = 0.8$. The probability of being a HW is rather low, about 10^{-4}.
- Assuming eating Frico is quite common, $p(F) = 0.4$, what is the probability that a Frico eater is a HW? By Bayes rule:

$$p(HW|F) = \frac{p(F|HW)p(HW)}{p(F)} = \frac{0.8 \cdot 10^{-4}}{0.4} = 2 \cdot 10^{-4}$$

- As Frico eating is rare worldwide, say $p(F) = 2 \cdot 10^{-4}$

$$p(HW|F) = \frac{p(F|HW)p(HW)}{p(F)} = \frac{0.8 \cdot 10^{-4}}{2 \cdot 10^{-4}} = 0.4$$

[1] cf. Monon Behaviour

Example

Example 1.4 (Who's in the bathroom?)**.** Consider a household of three people, Alice, Bob and Cecil. Cecil wants to go to the bathroom but finds it occupied. He then goes to Alice's room and sees she is there. Since Cecil knows that only either Alice or Bob can be in the bathroom, from this he infers that Bob must be in the bathroom.

To arrive at the same conclusion in a mathematical framework, we define the following events

$$A = \text{Alice is in her bedroom}, \quad B = \text{Bob is in his bedroom}, \quad O = \text{Bathroom occupied} \tag{1.2.11}$$

We can encode the information that if either Alice or Bob are not in their bedrooms, then they must be in the bathroom (they might both be in the bathroom) as

$$p(O = \mathsf{tr}|A = \mathsf{fa}, B) = 1, \qquad p(O = \mathsf{tr}|A, B = \mathsf{fa}) = 1 \tag{1.2.12}$$

The first term expresses that the bathroom is occupied if Alice is not in her bedroom, wherever Bob is. Similarly, the second term expresses bathroom occupancy as long as Bob is not in his bedroom. Then

$$p(B = \mathsf{fa}|O = \mathsf{tr}, A = \mathsf{tr}) = \frac{p(B = \mathsf{fa}, O = \mathsf{tr}, A = \mathsf{tr})}{p(O = \mathsf{tr}, A = \mathsf{tr})} = \frac{p(O = \mathsf{tr}|A = \mathsf{tr}, B = \mathsf{fa})p(A = \mathsf{tr}, B = \mathsf{fa})}{p(O = \mathsf{tr}, A = \mathsf{tr})} \tag{1.2.13}$$

where

$$\begin{aligned} p(O = \mathsf{tr}, A = \mathsf{tr}) = p(O = \mathsf{tr}|A = \mathsf{tr}, B = \mathsf{fa})p(A = \mathsf{tr}, B = \mathsf{fa}) \\ + p(O = \mathsf{tr}|A = \mathsf{tr}, B = \mathsf{tr})p(A = \mathsf{tr}, B = \mathsf{tr}) \end{aligned} \tag{1.2.14}$$

Using the fact $p(O = \mathsf{tr}|A = \mathsf{tr}, B = \mathsf{fa}) = 1$ and $p(O = \mathsf{tr}|A = \mathsf{tr}, B = \mathsf{tr}) = 0$, which encodes that if Alice is in her room and Bob is not, the bathroom must be occupied, and similarly, if both Alice and Bob are in their rooms, the bathroom cannot be occupied,

$$p(B = \mathsf{fa}|O = \mathsf{tr}, A = \mathsf{tr}) = \frac{p(A = \mathsf{tr}, B = \mathsf{fa})}{p(A = \mathsf{tr}, B = \mathsf{fa})} = 1 \tag{1.2.15}$$

This example is interesting since we are not required to make a full probabilistic model in this case thanks to the limiting nature of the probabilities (we don't need to specify $p(A, B)$). The situation is common in limiting situations of probabilities being either 0 or 1, corresponding to traditional logic systems.

Example

Exercise 1.3 (Adapted from [181])**.** *There are two boxes. Box 1 contains three red and five white balls and box 2 contains two red and five white balls. A box is chosen at random* $p(box = 1) = p(box = 2) = 0.5$ *and a ball chosen at random from this box turns out to be red. What is the posterior probability that the red ball came from box 1?*

5 Some probability distributions

In this section we introduce some of the most fundamental probability distributions, both discrete and continuous.

5.1 Discrete probability distributions

5.1.1 Discrete/ categorical distribution

A random variable can take N distinct values with probability $p_i, \quad i = 1, \ldots, N$. Formally

$$p(x = i) = p_i \quad \sum_{i=1}^{N} p_i = 1$$

5.1.2 Bernoulli distribution

A discrete random variable x with two outcomes: 1, with probability $p(x = 1|\theta) = \theta$ and 0, with probability $1 - \theta$. Compute mean and variance.

- $\langle x \rangle = \theta$
- $\text{var}(x) = \theta(1 - \theta)$

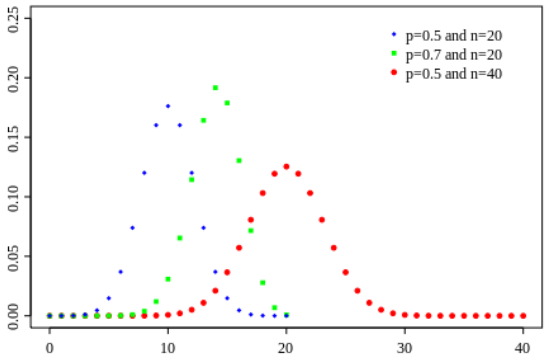
5.1.3 Binomial distribution

Describes the outcome of n Bernoulli trials. The probability of k successes is

$$p(y = k|\theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}$$

Compute mean and variance.

- $\langle x \rangle = n\theta$
- $\text{var}(x) = n\theta(1 - \theta)$



5.1.4 Multinomial distribution

Describes the outcome of n trials of a categorical distribution on $\{1, \ldots, K\}$ with probabilities $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$. The probability of observing y_i outcomes of type i is

$$p(y_1, \ldots, y_K|\boldsymbol{\theta}) = \frac{n!}{y_1! \cdots y_K!} \prod_{i=1}^{K} \theta_i^{y_i}$$

Compute mean and variance.

- $\langle y_i \rangle = n\theta_i$
- $\text{var}(y_i) = n\theta_i(1 - \theta_i)$
- $\text{cov}(y_i, y_j) = -n\theta_i\theta_j$

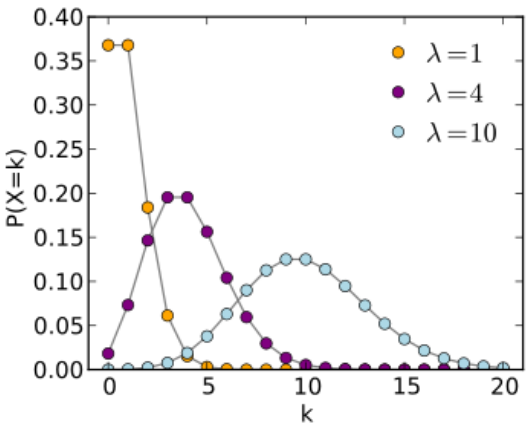
5.1.5 Poisson distribution

A distribution over non-negative integers

$$p\,(n|\lambda) = \frac{\lambda^n}{n!} \exp[-\lambda]$$

The parameter λ is often called the *rate* of the distribution. The Poisson distribution is often used for *rare events*, e.g. decaying of particles or binding of DNA fragments to a probe.
Compute mean and variance.

- $\langle x \rangle = \lambda$
- $\text{var}(x) = \lambda$
- A binomial with parameter $\theta = \lambda/n$ converges to a Poisson for $n \to \infty$

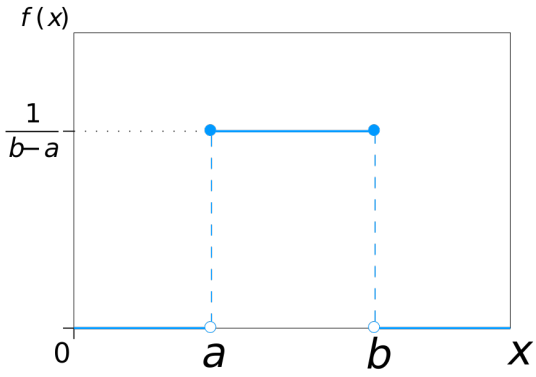


5.2 Continuous probability distributions

5.2.1 Uniform distribution

A variable x with constant density, over its domain of definition, which is an interval $[a, b] \subset \mathbb{R}$. Hence $p(x) = 1/(b-a)$, if $x \in [a, b]$.

- $\langle x \rangle = (a+b)/2$
- $\text{var}(x) = (a^2 + b^2 + ab)/3$



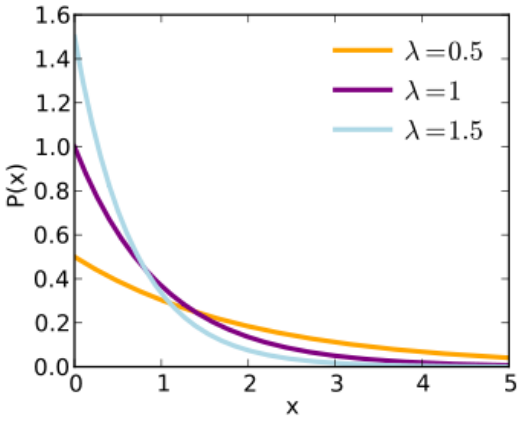
5.2.2 Exponential distribution

Typically used for the time of occurrence of a random event, like the decay of a particle or the arrival of a customer in a shop. For $x \geq 0$,

$$p(x|\lambda) = \lambda e^{-\lambda x}$$

with λ known as the rate of the distribution.

- $\langle x \rangle = 1/\lambda$
- $\text{var}(x) = 1/\lambda^2$



5.2.3 Gamma distribution

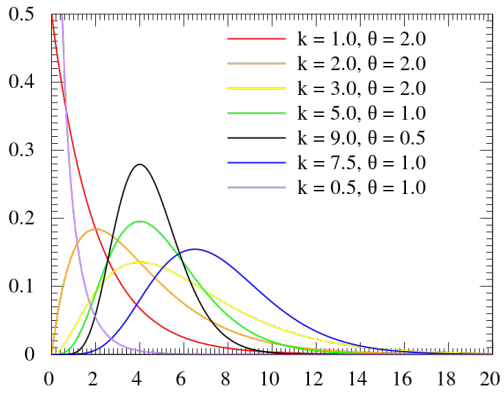
Defined for $x \geq 0$, by the density

$$p(x|\alpha,\beta) = \frac{1}{\beta\Gamma(\alpha)}(x/\beta)^{\alpha-1}e^{-x/\beta}$$

where α is the shape parameter, β is the scale parameter, and $\Gamma(\alpha)$ is the gamma function

$$\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt$$

- $\langle x \rangle = \alpha/\beta$
- $\text{var}(x) = \alpha/\beta^2$
- $Gamma(1,\lambda) \equiv Exp(\lambda)$



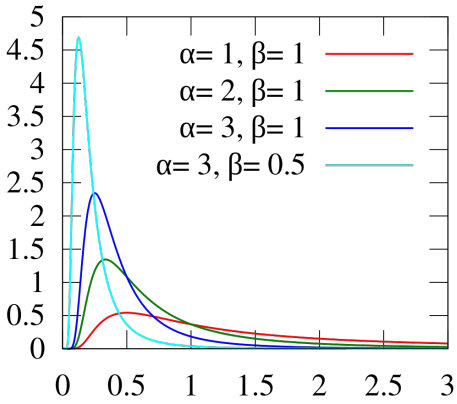
5.2.4 Inverse Gamma distribution

Defined for $x \geq 0$, by the density

$$p(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}(1/x)^{\alpha+1}e^{-\beta/x}$$

where α is the shape parameter, β is the scale parameter, and $\Gamma(\alpha)$ is the gamma function

- $\langle x \rangle = \beta/(\alpha-1)$, $\alpha > 1$
- $\text{var}(x) = \beta^2/((\alpha-1)^2(\alpha-2))$



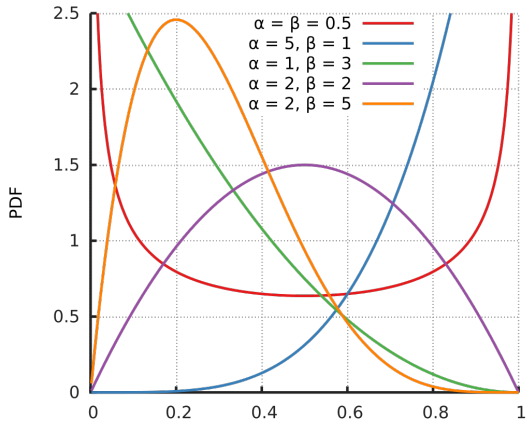
5.2.5 Beta distribution

Defined for $x \in [0, 1]$ by

$$p(x|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

where the Beta function $B(\alpha,\beta)$ is defined by $[B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$

- $\langle x \rangle = \alpha/(\alpha+\beta)$
- $\text{var}(x) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$



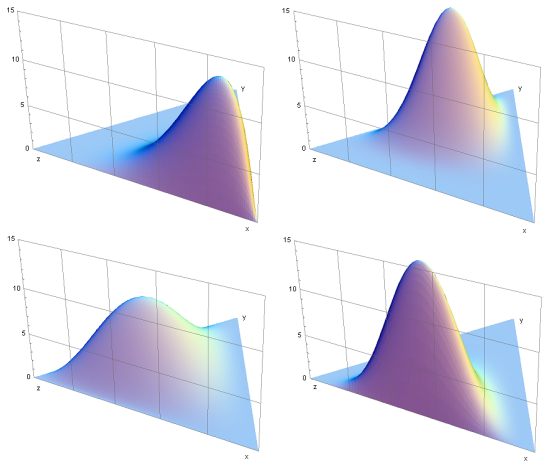
5.2.6 Dirichlet distribution

It is a distribution over discrete probability distributions on k elements. It depends on k parameters $\mathbf{u} = u_1, \ldots, u_k$. Its pdf is

$$p_{dir}(\mathbf{x}|\mathbf{u}) = p(\mathbf{x}|\mathbf{u}) = \frac{1}{B(\mathbf{u})} \prod_{i=1}^{k} x_i^{u_i - 1}$$

with $B(\mathbf{u}) = \frac{\prod_{i=1}^{k} \Gamma(u_i)}{\Gamma(\sum_{i=1}^{k} u_i)}$

- $\langle x_i \rangle = u_i/(u_i + \bar{u}_i)$, $\bar{u}_i = \sum_{j \neq i} u_j$
- $\text{var}(x_i) = \frac{u_i \bar{u}_i}{(u_i+\bar{u}_i)^2(u_i+\bar{u}_i+1)}$
- $p_{dir}(\mathbf{x}|\mathbf{u_1})p_{dir}(\mathbf{x}|\mathbf{u_2}) = p_{dir}(\mathbf{x}|\mathbf{u_1} + \mathbf{u_2})$
- $\int_{x_j} p_{dir}(\mathbf{x}|\mathbf{u}) dx_j = p_{dir}(\mathbf{x_{/j}}|\mathbf{u_{/j}})$
- $\int_{\mathbf{x_{/j}}} p_{dir}(\mathbf{x}|\mathbf{u}) d\mathbf{x_{/j}} = p_{beta}(x_j|u_j, \sum_{i \neq j} u_j)$



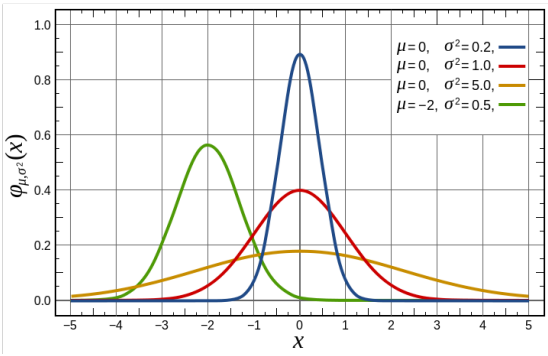
5.3 Univariate Normal distribution

One of the most used distributions in Machine Learning

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

- $\langle x \rangle = \mu$
- $\text{var}(x) = \sigma^2$
- For $\mu = 1$, $\sigma^2 = 1$ we talk about a standard normal distribution.

- $1/\sigma^2$ is known as *precision*.



5.3.1 Student's t-distribution

$$p(x|\mu,\lambda,\nu) = Student\,(x|\mu,\lambda,\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})}\left(\frac{\lambda}{\nu\pi}\right)^{\frac{1}{2}}\left[1+\frac{\lambda\,(x-\mu)^2}{\nu}\right]^{-\frac{\nu+1}{2}} \tag{8.3.24}$$

where μ is the mean, ν the degrees of freedom, and λ scales the distribution. The variance is given by

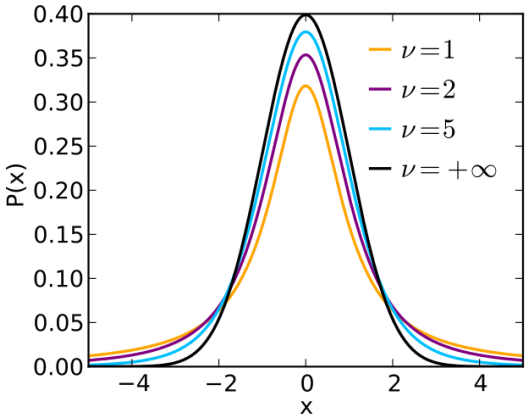
$$\text{var}(x) = \frac{\nu}{\lambda\,(\nu-2)},\ \text{for}\ \nu > 2 \tag{8.3.25}$$

For $\nu \to \infty$ the distribution tends to a Gaussian with mean μ and variance $1/\lambda$. As ν decreases the tails of the distribution become fatter.

The t-distribution can be derived from a *scaled mixture*

$$p(x|\mu,a,b) = \int_{\tau=0}^{\infty} \mathcal{N}\left(x|\mu,\tau^{-1}\right) Gam^{is}\left(\tau|a,b\right) d\tau$$

With $\nu = 2a$ degrees of freedom and scale $\lambda = a/b$.

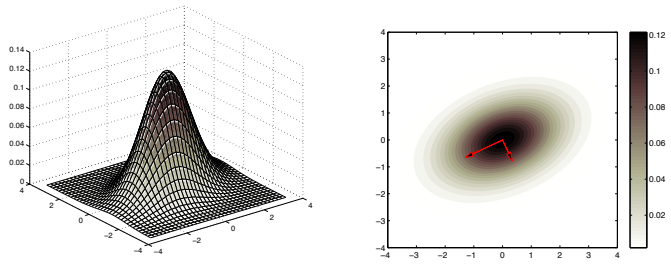


5.4 Multivariate normal distribution

This is the most important distribution we will use, and generalises the 1d normal. In d dimensions

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})^T\right)$$

It holds $\boldsymbol{\mu} = \langle \mathbf{x} \rangle$, and $\boldsymbol{\Sigma} = \text{cov}(\mathbf{x}, \mathbf{x}) = \langle (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T \rangle$



(a) (b)

We will now present some properties of the multivariate normal distribution

5.4.1 Completing the square

A useful technique in manipulating Gaussians is completing the square. For example, the expression

$$\exp\left(-\frac{1}{2}\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} + \mathbf{b}^\mathsf{T}\mathbf{x}\right) \tag{8.4.10}$$

can be transformed as follows. First we complete the square:

$$\frac{1}{2}\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} - \mathbf{b}^\mathsf{T}\mathbf{x} = \frac{1}{2}\left(\mathbf{x} - \mathbf{A}^{-1}\mathbf{b}\right)^\mathsf{T}\mathbf{A}\left(\mathbf{x} - \mathbf{A}^{-1}\mathbf{b}\right) - \frac{1}{2}\mathbf{b}^\mathsf{T}\mathbf{A}^{-1}\mathbf{b} \tag{8.4.11}$$

Hence

$$\exp\left(-\frac{1}{2}\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} - \mathbf{b}^\mathsf{T}\mathbf{x}\right) = \mathcal{N}\left(\mathbf{x}|\mathbf{A}^{-1}\mathbf{b}, \mathbf{A}^{-1}\right)\sqrt{\det\left(2\pi\mathbf{A}^{-1}\right)}\exp\left(\frac{1}{2}\mathbf{b}^\mathsf{T}\mathbf{A}^{-1}\mathbf{b}\right) \tag{8.4.12}$$

$p(\mathbf{x}|\mathbf{A}, \mathbf{b}) = c\exp\left(-\frac{1}{2}\mathbf{x}\mathbf{A}\mathbf{x}^T + \mathbf{b}^T\mathbf{x}\right)$ is known as the canonical representation, and it is normal with mean $\mathbf{A}^{-1}\mathbf{b}$ and covariance $\mathbf{A}^{-1}$.

5.4.2 Linear transformation

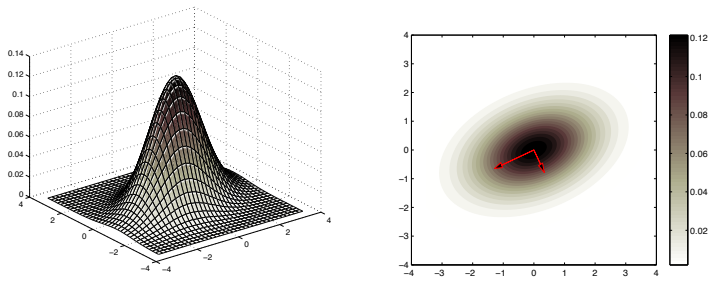
Result 8.3 (Linear Transform of a Gaussian)**.** Let $\mathbf{y}$ be linearly related to $\mathbf{x}$ through

$$\mathbf{y} = \mathbf{M}\mathbf{x} + \boldsymbol{\eta} \tag{8.4.14}$$

where $\mathbf{x} \perp\!\!\!\perp \boldsymbol{\eta}$, $\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$. Then the marginal $p(\mathbf{y}) = \int_\mathbf{x} p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$ is a Gaussian

$$p(\mathbf{y}) = \mathcal{N}\left(\mathbf{y}|\mathbf{M}\boldsymbol{\mu}_x + \boldsymbol{\mu}, \mathbf{M}\boldsymbol{\Sigma}_x\mathbf{M}^\mathsf{T} + \boldsymbol{\Sigma}\right) \tag{8.4.15}$$

5.4.3 Eigendecomposition



(a) (b)

$$\boldsymbol{\Sigma} = \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}^{\mathsf{T}} \tag{8.4.5}$$

where $\mathbf{E}^{\mathsf{T}}\mathbf{E} = \mathbf{I}$ and $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_D)$. In the case of a covariance matrix, all the eigenvalues λ_i are positive. This means that one can use the transformation

$$\mathbf{y} = \boldsymbol{\Lambda}^{-\frac{1}{2}}\mathbf{E}^{\mathsf{T}}(\mathbf{x} - \boldsymbol{\mu}) \tag{8.4.6}$$

so that

$$(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = (\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} \mathbf{E}\boldsymbol{\Lambda}^{-1}\mathbf{E}^{\mathsf{T}} (\mathbf{x} - \boldsymbol{\mu}) = \mathbf{y}^{\mathsf{T}}\mathbf{y} \tag{8.4.7}$$

So by rescaling, we can obtain a product of d-univariate standard normal distributions, one per dimension.

5.4.4 Marginal and conditional of multivariate Gaussians

Result 8.4 (Partitioned Gaussian)**.** Consider a distribution $\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ defined jointly over two vectors $\mathbf{x}$ and $\mathbf{y}$ of potentially differing dimensions,

$$\mathbf{z} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \tag{8.4.16}$$

with corresponding mean and partitioned covariance

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{pmatrix} \tag{8.4.17}$$

where $\boldsymbol{\Sigma}_{yx} \equiv \boldsymbol{\Sigma}_{xy}^{\mathsf{T}}$. The marginal distribution is given by

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{xx}) \tag{8.4.18}$$

and conditional

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}\left(\mathbf{x}\middle|\boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\mathbf{y} - \boldsymbol{\mu}_y), \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}\right) \tag{8.4.19}$$

5.4.5 Product of multivariate Gaussians

Result 8.2 (Product of two Gaussians)**.** The product of two Gaussians is another Gaussian, with a multiplicative factor, exercise(8.35):

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \frac{\exp\left(-\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\mathsf{T} \mathbf{S}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right)}{\sqrt{\det(2\pi\mathbf{S})}} \tag{8.4.8}$$

where $\mathbf{S} \equiv \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2$ and the mean and covariance are given by

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}_1\mathbf{S}^{-1}\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_2\mathbf{S}^{-1}\boldsymbol{\mu}_1 \qquad \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1\mathbf{S}^{-1}\boldsymbol{\Sigma}_2 \tag{8.4.9}$$

5.4.6 Gaussian average of a quadratic function

Result 8.5 (Gaussian average of a quadratic function)**.**

$$\left\langle \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} \right\rangle_{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})} = \boldsymbol{\mu}^\mathsf{T}\mathbf{A}\boldsymbol{\mu} + \mathrm{trace}(\mathbf{A}\boldsymbol{\Sigma})$$

5.5 The Curse of Dimensionality

When the dimensionality of random variables becomes high (a typical scenario for big data and modern machine learning) some counterintuitive phenomena start to emerge. Here's some as exercises.

Exercise

Suppose you want to explore uniformly a region by gridding it. How many grid points do you need?

Exercise

Suppose you sample from a uniform distribution in d dimensions. What is the probability of finding a point inside the region $[\epsilon, 1-\epsilon]^d$?

Exercise

Suppose you sample from a spherical Gaussian distribution. Where do the points lie as the dimensions increase?

5.6 Mixtures: how to build more distributions

- More general distributions can be built via mixtures: e.g.

$$p(x|\mu_{1\ldots,n}, \sigma^2_{1,\ldots,n}) = \sum_i \pi_i \mathcal{N}(\mu_i, \sigma_i^2)$$

  where the *mixing coefficients* π_i are discretely distributed

- You can interpret this as a two stage hierarchical process: choose one component out of a discrete distribution, then choose the distribution for that component
- **IMPORTANT CONCEPT**: the mixture

$$p(x|\mu_{1\ldots,n}, \sigma^2_{1,\ldots,n}) = \sum_i \pi_i \mathcal{N}(\mu_i, \sigma_i^2)$$

  is an example of *latent variable model*, with a latent class variable and an observed continuous value. The mixture is the marginal distribution for the observations (w.r.t. the latent variable)
- The probability of the latent variables given the observations can be obtained using Bayes' theorem.

5.6.1 Continuous mixtures: some cool distributions

- No need for the mixing distribution (latent variable) to be discrete
- Suppose you are interested in the means of normally distributed samples (possibly with different variances/ precisions): Marginalising the precision in a Gaussian using a Gamma mixing distribution yields a *Student t-distribution*
- Suppose you have multiple rare event processes happening with slightly different rates: Marginalising the rate in a Poisson distribution using a Gamma mixing distribution yields a *negative binomial* distribution

6 Estimation: fitting distributions

6.1 Parameters?

- Many distributions are written as conditional probabilities *given* the parameters: $p(x|\theta)$
- Often the values of the parameters are not known
- If we have observations, we can try to estimate the parameters from such data.
- We assume to have independent and identically distributed (i.i.d.) observations of $p(x|\theta_{true})$: $\mathbf{x} = x_1, \ldots, x_N$.

6.2 Maximum Likelihood

- Likelihood for i.i.d. observations $\mathbf{x} = x_1, \ldots, x_N$:

$$p(\mathbf{x}|\theta) = \prod_{i=1}^{N} p(x_i|\theta)$$

- Choose the parameters that best explain the observations: we pick θ by maximum likelihood:

$$\hat{\theta} = \text{argmax}_\theta \left[\prod_i p(x_i|\theta) \right]$$

6.3 Maximum a posteriori

- Suppose we can encode prior knowledge (or absence of it) in a *prior* distribution over parameters, $p(\theta)$.
- We can then compute the posterior distribution, given i.i.d. observations $\mathbf{x} = x_1, \ldots, x_N$, by Bayes theorem:

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})}$$

where

$$p(\mathbf{x}) = \int_\theta p(\mathbf{x}|\theta)p(\theta)d\theta$$

- Estimate θ_{true} by the *maximum a posteriori (MAP)* estimate

$$\hat{\theta}_{MAP} = \text{argmax}_\theta \left[p(\theta) \prod_i p(x_i|\theta) \right]$$

6.4 Exercise: fitting a discrete distribution

- We have a discrete distribution with values in $K = \{1, \ldots, k\}$, with parameters $\boldsymbol{\mu} = \mu_1, \ldots, \mu_k$, $\sum_i \mu_i = 1$.
- We have independent observations $\mathbf{x} = x_1, \ldots, x_N$, each taking values in K.
- The likelihood is

$$\mathcal{L}(\boldsymbol{\mu}) = p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{i=1}^{N} p\left(x_i|\boldsymbol{\mu}\right)$$

- Compute the Maximum Likelihood estimate of $\boldsymbol{\mu}$. What is the intuitive meaning of the result? What happens if one of the D values is not represented in your sample?

6.5 Exercise: fitting a discrete distribution

6.6 Exercise II: fitting a Gaussian distribution

-4.5cm We have independent, real valued observations $\mathbf{x} = x_1, \ldots, x_N$. Fit a Gaussian by maximum likelihood.

6.7 Bayesian estimation

- The Bayesian approach fully quantifies uncertainty
- The parameters are treated as additional random variables with their own *prior* distribution $p(\theta)$
- The observation likelihood is combined with the prior to obtain a *posterior* distribution via Bayes' theorem

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})}$$

- The distribution of the observable x (*predictive* distribution) is obtained as

$$p(x|\mathbf{x}) = \int p(x|\theta)p(\theta|\mathbf{x})d\theta$$

6.8 Exercise: Bayesian fitting of Gaussians

- Let data $x_i \quad i = 1, \ldots, N$ be distributed according to a Gaussian with mean μ and variance σ^2
- Let the prior distribution over the mean μ be a Gaussian with mean m and variance v^2
- Compute the posterior (and predictive distribution, Exercise)

6.9 Exercise: Bayesian fitting of Gaussians

6.10 Estimators

- A procedure to calculate an expectation is called an *estimator*
- e.g., fitting a Gaussian to data by maximum likelihood provides the M.L. estimator for mean and variance, or Bayesian posterior mean
- An estimator will be a noisy estimate of the true value, due to finite sample effects
- An estimator $\hat{f}$ is *unbiased* if its expectation (under the joint distribution of the data set) coincides with the true value
- An estimator $\hat{f}$ is *consistent* if it converges to the true value when the number of data goes to infinity.

6.11 Exercise: biased estimator

-4.5cm The ML estimator of variance, $\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})^2$ is biased: $\langle\hat{\sigma}^2\rangle = \frac{N-1}{N}\sigma^2$.

6.12 Bootstrap

- For an estimator, in theory we can compute its mean and its variance under the joint distribution of the datasets. In practice, getting the variance may be very hard. Bootstrapping can be used instead.
- Given the dataset $\mathtt{x} = x_1, \ldots, x_N$, construct from it K new datasets $\mathtt{x_i}$, also of size N, by sampling with repetitions.
- compute the estimator $\hat{\theta}_i$ for each $\mathtt{x_i}$.
- Compute the empirical variance (or any other statistics) from $\mathtt{x_1}, \ldots, \mathtt{x_K}$.
- This is an estimate of the actual variance of the estimator.

6.13 Conjugate priors

- The Bayesian way has advantages in that it quantifies uncertainty and regularizes naturally
- BUT computing the normalisation in Bayes theorem is very hard
- The case when it is possible is when the prior and the posterior are of the same form (*conjugate*)
- Example + Exercise: Bernoulli and Beta.
- Example: discrete and Dirichlet
- Exercise: conjugate priors for the univariate normal (mean)

6.14 Conjugate priors: Binomial and Beta

-4.5cm Show that the Beta is the conjugate prior for the Bernoulli distribution.

7 Information theory

7.1 Entropy

- Probability theory is the basis of information theory (interesting, but not the topic of this course).
- An important quantity is the *entropy* of a distribution

$$H[p] = -\sum_i p_i \log_2 p_i$$

Or for continuous distributions:

$$H[p] = -\int p(x) \log p(x) dx$$

- Entropy measures the level of disorder of a distribution; for discrete distributions, it is always ≥ 0 and 0 only for deterministic distributions. The maximum is $\log K$, if K is the size of the support of the discrete distribution, and is achieved by the uniform distribution.

7.2 Divergence

- The *relative entropy* or *Kullback-Leibler (KL) divergence* between two distributions is
$$KL[q||p] = \sum_i q_i \log \frac{q_i}{p_i}$$
Of in the continuous case
$$KL[q||p] = \int q(x) \log \frac{q(x)}{p(x)} dx$$
- Fact: KL is convex and ≥ 0 (by Jensen ineq)
- Fact: KL is zero if and only if $p = q$.

7.3 Conditional Entropy and mutual information

- Conditional entropy is defined as
$$H[p(x|y)] = -\int\int p(x,y) \log p(x|y) dxdy = H[p(x,y)] - H[p(y)]$$
and captures the residual uncertainty on x once y is known.
- Mutual information between r.v. x and y is defined as
$$I[x,y] = KL[p(x,y)|p(x)p(y)] = H[p(x)] - H[p(x|y)]$$
and captures the reduction in uncertainty about x by knowing y, i.e. it is a measure of how much y brings information about x, and viceversa.

7.4 Justification for maximum likelihood

- Given a data set $\mathtt{x} = \{x_i\}, \quad i = 1, \ldots, N$, let the empirical distribution be
$$p_{emp}(x) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(x_i)$$
with $\mathbb{I}$ the indicator function of a set
- To find a suitable distribution q to model the data, one may wish to minimize the Kullback-Leibler divergence
$$KL[p_{emp}||q] = H[p_{emp}] - \langle \log q(x) \rangle_{p_{emp}} = -\frac{1}{N} \sum \log q(x_i)$$
- *Maximum likelihood is equivalent to minimizing a KL divergence with the empirical distirbution*

8 Decision theory

8.1 An overview

- Suppose we have a classification problem, and we are able to learn a model of the joint distribution $p(x, y)$, where y is a categorical variable. Given a new input x^*, for which we want to make a prediction, to which class should we assign it?
- We may choose to assign it to class j if $p(y = j|x^*)$ is the maximum one. However, suppose y models having or not a cancer, and that $p(y = 0|x^*) = 0.51 > 0.49 = p(y = 1|x^*)$.
- To be more flexible, we can specify a loss function (or utility function), which is the cost $c_{k,j}$ of assigning x^* to class j when the true class is k.
- Then we can assign a point x^* to the class j minimising the expected loss w.r.t. the learned joint distribution (i.e. $\sum_k c_{k,j} p(y = k|x^*)$).