Some basics in probability and statistics

Course of Machine Learning
Master Degree in Computer Science
University of Rome ``Tor Vergata''

Giorgio Gambosi

a.a. 2016-2017

A discrete random variable X can take values from some finite or countably infinite set $\mathcal{X}$. A probability mass function (pmf) associates to each event $X = x$ a probability $p(X = x)$.

Properties

- $0 \leq p(x) \leq 1$ for all $x \in \mathcal{X}$
- $\sum_{x \in \mathcal{X} p(x) = 1}$

Note: we shall denote as x the event $X = x$

Joint and conditional probabilities

Given two events x, y, it is possible to define:

- the probability $p(x, y) = p(x \wedge y)$ of their joint occurrence
- the conditional probability $p(x|y)$ of x under the hypothesis that y has occurred

Union of events

Given two events x, y, the probability of x or y is defined as

$$p(x \vee y) = p(x) + p(y) - p(x, y)$$

in particular,

$$p(x \vee y) = p(x) + p(y)$$

The same definitions hold for probability distributions.

Product rule

The product rule relates joint and conditional probabilities

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

where $p(x)$ is the marginal probability.

In general,

$$\begin{aligned} p(x_1, \dots, x_n) &= p(x_2, \dots, x_n|x_1)p(x_1) \\ &= p(x_3, \dots, x_n|x_1, x_2)p(x_2|x_1)p(x_1) \\ &= \cdots \\ &= p(x_n|x_1, \dots, x_{n-1})p(x_{n-1}|x_1 \dots x_{n-2}) \cdots p(x_2|x_1)p(x_1) \end{aligned}$$

Sum rule and marginalization

The sum rule relates the joint probability of two events x, y and the probability of one such events $p(y)$ (or $p(y)$)

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

Applying the sum rule to derive a marginal probability from a joint probability is usually called marginalization

Discrete random variables

Bayes rule

Since

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

and

$$p(y) = \sum_{x \in \S} p(x, y) = \sum_{x \in \mathcal{X}} p(y|x)p(x)$$

it results

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} = \frac{p(y|x)p(x)}{\sum_{x \in \mathcal{X}} p(y|x)p(x)}$$

Terminology

- $p(x)$: Prior probability of x (before knowing that y occurred)
- $p(x|y)$: Posterior of x (if y has occurred)
- $p(y|x)$: Likelihood of y given x
- $p(y)$: Evidence of y

Independence

Definition

Two random variables X, Y are independent ($X \perp\!\!\!\perp Y$) if their joint probability is equal to the product of their marginals

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

or, equivalently,

$$p(x|y) = p(x) \qquad\qquad p(y|x) = p(y)$$

The condition $p(x|y) = p(x)$, in particular, states that, if two variables are independent, knowing the value of one does not add any knowledge about the other one.

Conditional independence

Two random variables X, Y are conditionally independent w.r.t. a third r.v. Z ($X \perp\!\!\!\perp Y | Z$) if

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

Conditional independence does not imply (absolute) independence, and vice versa.

A continuous random variable X can take values from a continuous infinite set $\mathcal{X}$. Its probability is defined as cumulative distribution function (cdf) $F(x) = p(X \leq x)$.

The probability that X is in an interval $(a, b]$ is then $p(a < X \leq b) = F(b) - F(a)$.

Probability density function

The probability density function (pdf) is defined as $f(x) = \dfrac{dF(x)}{dx}$. As a consequence,

$$p(a < X \leq b) = \int_a^b f(x)dx$$

and

$$p(x < X \leq x + dx) \approx f(x)dx$$

for a sufficiently small dx.

Sum rule and continuous random variables

In the case of continuous random variables, their probability density functions relate as follows.

$$f(x) = \int_{\mathcal{Y}} f(x, y) dy = \int_{y \in \mathcal{Y}} p(x|y) p(y) dy$$

Definition

Let x be a discrete random variable with distribution $p(x)$, and let $g : \mathbb{R} \mapsto \mathbb{R}$ be any function: the expectation of $g(x)$ w.r.t. $p(x)$ is

$$E_p[g(x)] = \sum_{x \in V_x} g(x)p(x)$$

If x is a continuous r.v., with probability density $f(x)$, then

$$E_f[g(x)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

Mean value

Particular case: $g(x) = x$

$$E_p[x] = \sum_{x \in V_x} xp(x) \qquad\qquad E_f[x] = \int_{-\infty}^{\infty} xf(x)dx$$

Elementary properties of expectation

- $E[a] = a$ for each $a \in \mathbb{R}$
- $E[af(x)] = aE[f(x)]$ for each $a \in \mathbb{R}$
- $E[f(x) + g(x)] = E[f(x)] + E[g(x)]$

Variance

Definition

$$\mathit{Var}[X] = \mathit{E}[(x - \mathit{E}[x])^2]$$

We may easily derive:

$$\begin{aligned} \mathit{E}[(x - \mathit{E}[x])^2] &= \mathit{E}[x^2 - 2\mathit{E}[x]x + \mathit{E}[x]^2] \\ &= \mathit{E}[x^2] - 2\mathit{E}[x]\mathit{E}[x] + \mathit{E}[x]^2 \\ &= \mathit{E}[x^2] - \mathit{E}[x]^2 \end{aligned}$$

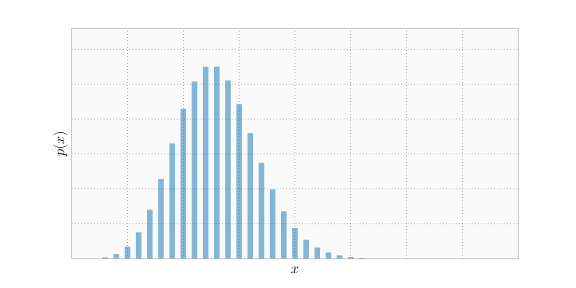
Some elementary properties:

- $\mathit{Var}[a] = 0$ for each $a \in \mathbb{R}$
- $\mathit{Var}[af(x)] = a^2\mathit{Var}[f(x)]$ for each $a \in \mathbb{R}$

Probability distribution

Given a discrete random variable $X \in V_X$, the corresponding probability distribution is a function $p(x) = P(X = x)$ such that

- $0 \leq p(x) \leq 1$
- $\sum_{x \in V_X} p(x) = 1$
- $\sum_{x \in A} p(x) = P(x \in A)$, with $A \subseteq V_X$

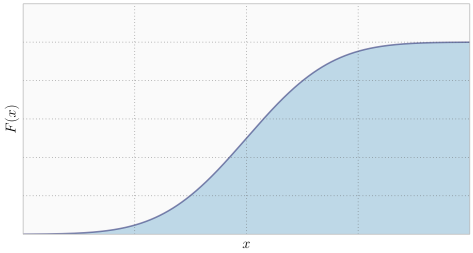


Some definitions

Cumulative distribution

Given a continuous random variable $X \in \mathbb{R}$, the corresponding cumulative probability distribution is a function $F(x) = P(X \leq x)$ such that:

- $0 \leq F(x) \leq 1$
- $\lim_{x \to -\infty} F(x) = 0$
- $\lim_{x \to \infty} F(x) = 1$
- $x \leq y \implies F(x) \leq F(y)$



Probability density

Given a continuous random variable $X \in \mathbb{R}$ with derivable cumulative distribution $F(x)$, the probability density is defined as

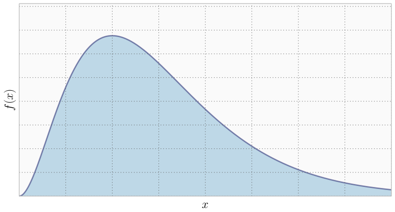
$$f(x) = \frac{dF(x)}{dx}$$

By definition of derivative, for a sufficiently small Δx,

$$Pr(x \leq X \leq x + \Delta x) \approx f(x)\Delta x$$

The following properties hold:

- $f(x) \geq 0$
- $\int_{-\infty}^{\infty} f(x)dx = 1$
- $\int_{x \in A} f(x)dx = P(X \in A)$



Definition

Let $x \in \{0, 1\}$, then $x \sim$ *Bernoulli*(p), with $0 \leq p \leq 1$, if

$$p(x) = \begin{cases} p & \text{se } x = 1 \\ 1 - p & \text{se } x = 0 \end{cases}$$

or, equivalently,

$$p(x) = p^x (1 - p)^{1-x}$$

Probability that, given a coin with head (H) probability p (and tail probability (T) $1 - p$), a coin toss result into $x \in \{H, T\}$.

Mean and variance

$$E[x] = p \qquad\qquad Var[x] = p(1 - p)$$

Extension to multiple outcomes

Assume k possible outcomes (for example a die toss).

In this case, a generalization of the Bernoulli distribution is considered, usualy named categorical distribution.

$$p(x) = \prod_{j=1}^{k} p_j^{x_j}$$

where $(p_1, \ldots, p_k)$ are the probabilites of the different outcomes ($\sum_{j=1}^{k} p_j = 1$) and $x_j = 1$ iff the k-th outcome occurs.

Binomial distribution

Definition

Let $x \in \mathbb{N}$, then $x \sim \mathit{Binomial}(n, p)$, with $0 \leq p \leq 1$, if

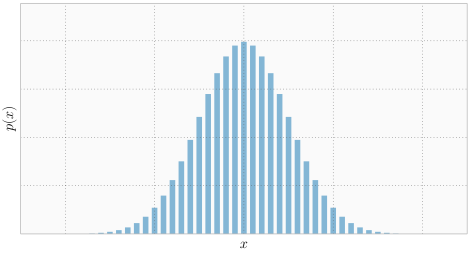
$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

Probability that, given a coin with head (H) probability p, a sequence of n independent coin tosses result into x heads.

Mean and variance

$$E[x] = np$$

$$Var[x] = np(1-p)$$



Definition

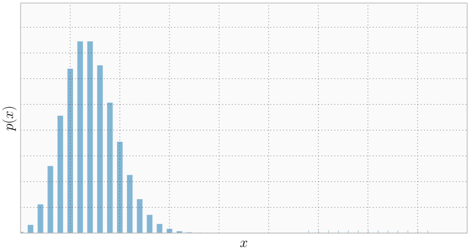
Let $x_i \in \mathbb{N}$, then $x \sim \mathit{Poisson}(\lambda)$, with $\lambda > 0$, if

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

Probability that an event with average frequency λ occurs x times in the next time unit.

Mean and variance

$$E[x] = \lambda$$
$$Var[x] = \lambda$$



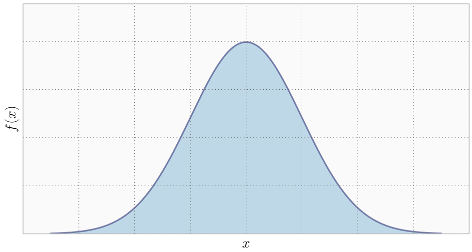
Definition

Let $x \in \mathbb{R}$, then $x \sim \mathit{Normal}(\mu, \sigma^2)$, with $\mu, \sigma \in \mathbb{R}$, $\sigma \geq 0$, if

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

Mean and variance

$$E[x] = \mu$$
$$Var[x] = \sigma^2$$



Definition

Let $x \in [0, 1]$, then $x \sim Beta(\alpha, \beta)$, with $\alpha, \beta > 0$, if

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

where

$$\Gamma(x) = \int_0^\infty u^{x-1} e^u du$$

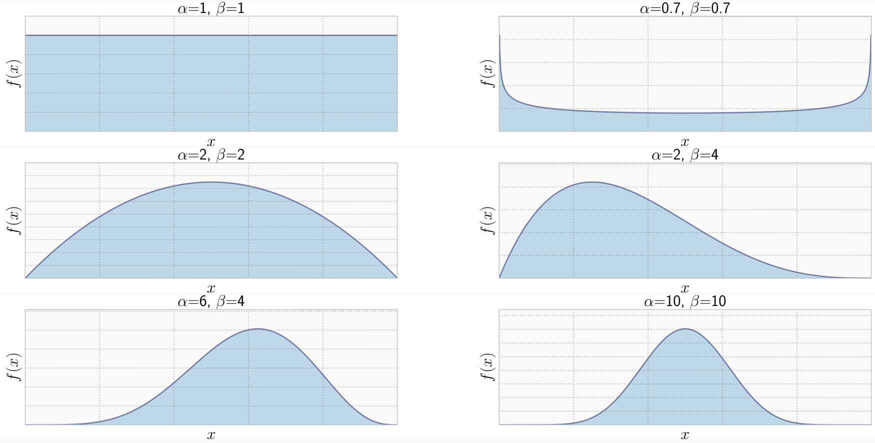
is a generalization of the factorial to the real field $\mathbb{R}$: in particolar, $\Gamma(n) = (n-1)!$ if $n \in \mathbb{N}$

Mean and variance

$$E[x] = \frac{\beta}{\alpha+\beta}$$

$$Var[x] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Beta distribution



Definition for $k = 2$ discrete variables

Given two discrete r.v. X, Y, their joint distribution is

$$p(x, y) = P(X = x, Y = y)$$

The following properties hold:

1. $0 \leq p(x, y) \leq 1$
2. $\sum_{x \in V_X} \sum_{y \in V_Y} p(x, y) = 1$

Definition for $k = 2$ variables

Given two continuous r.v. X, Y, their cumulative joint distribution is defined as

$$F(x, y) = P(X \leq x, Y \leq y)$$

The following properties hold:

1. $0 \leq F(x, y) \leq 1$
2. $\lim\limits_{x,y \to \infty} F(x, y) = 1$
3. $\lim\limits_{x,y \to -\infty} F(x, y) = 0$

If $F(x, y)$ is derivable everywhere w.r.t. both x and y, joint probability density is

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}$$

The following property derives

$$\int\int_{(x,y) \in A} f(x, y) dx dy = P((X, Y) \in A)$$

Definition

$$Cov[X,Y] = E[(X - E[X])(Y - E[Y])]$$

As for the variance, we may derive

$$\begin{aligned} Cov[X,Y] &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY - XE[Y] - YE[X] + E[X]E[Y]] \\ &= E[XY] - E[X]E[Y] - E[Y]E[X] + E[E[X]E[Y]] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

Moreover, the following properties hold:

1. $Var[X+Y] = Var[X] + Var[Y] + 2Cov[X,Y]$
2. If $X \perp\!\!\!\perp Y$ then $Cov[X,Y] = 0$

Definition

Let $X_1, X_2, \ldots, X_n$ be a set of r.v.: we may then define a random vector as

$$\mathbf{x} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

Definition

Let $g : \mathbb{R}^n \mapsto \mathbb{R}^m$ be any function. It may be considered as a vector of functions

$$g(\mathbf{x}) = \begin{bmatrix} g_1(\mathbf{x})) \\ g_2(\mathbf{x})) \\ \vdots \\ g_m(\mathbf{x}) \end{bmatrix}$$

where $\mathbf{x} \in \mathbb{R}^n$.

The expectation of g is the vector of the expectations of all functions g_i,

$$E[g(\mathbf{x})] = \begin{bmatrix} E[g_1(\mathbf{x})] \\ E[g_2(\mathbf{x})] \\ \vdots \\ E[g_m(\mathbf{x})] \end{bmatrix}$$

Definition

Let $\mathbf{x} \in \mathbb{R}^n$ be a random vector: its covariance matrix $\boldsymbol{\Sigma}$ is a matrix $n \times n$ such that, for each $1 \leq i, j \leq n$, $\Sigma_{ij} = \mathit{Cov}[X_i, X_j] = \mathit{E}[(X_i - \mu_i)(X_j - \mu_j)]$, where $\mu_i = \mathit{E}[X_i]$, $\mu_j = \mathit{E}[X_j]$.

Hence,

$$
\begin{aligned}
\boldsymbol{\Sigma} &= \begin{bmatrix} \mathit{Cov}[X_1, X_1] & \mathit{Cov}[X_1, X_2] & \cdots & \mathit{Cov}[X_1, X_n] \\ \mathit{Cov}[X_2, X_1] & \mathit{Cov}[X_2, X_2] & \cdots & \mathit{Cov}[X_2, X_n] \\ \vdots & \vdots & \ddots & \vdots \\ \mathit{Cov}[X_n, X_1] & \mathit{Cov}[X_n, X_2] & \cdots & \mathit{Cov}[X_n, X_n] \end{bmatrix} \\
&= \begin{bmatrix} \mathit{Var}[X_1] & \cdots & \mathit{Cov}[X_1, X_n] \\ \vdots & \ddots & \vdots \\ \mathit{Cov}[X_n, X_1] & \cdots & \mathit{Var}[X_n] \end{bmatrix}
\end{aligned}
$$

By definition of covariance,

$$\begin{aligned}\boldsymbol{\Sigma} &= \begin{bmatrix} E[X_1^2] - E[X_1]^2 & \cdots & E[X_1 X_n] - E[X_1]E[X_n] \\ \vdots & \ddots & \vdots \\ E[X_n X_1] - E[X_n]E[X_1] & \cdots & E[X_n^2] - E[X_n]E[X_n] \end{bmatrix} \\ &= E[\mathbf{X}\mathbf{X}^T] - \boldsymbol{\mu}\boldsymbol{\mu}^T\end{aligned}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ is the vector of expectations of the random variables $X_1, \ldots, X_n$.

Properties

The covariance matrix is necessarily:

- semidefinite positive: that is, $\mathbf{z}^T \boldsymbol{\Sigma} \mathbf{z} \geq 0$ for any $\mathbf{z} \in \mathbb{R}^n$
- symmetric: $Cov[X_i, X_j] = Cov[X_j, X_i]$ for $1 \leq i, j \leq n$

Correlation

For any pair of r.v. X, Y, the Pearson correlation coefficient is defined as

$$\rho_{X,Y} = \frac{\mathit{Cov}[X, Y]}{\sqrt{\mathit{Var}[X]\mathit{Var}[Y]}}$$

Note that, if $Y = aX + b$ for some pair a, b, then

$$\mathit{Cov}[X, Y] = E[(X - \mu)(aX + b - a\mu - b)] = E[a(X - \mu)^2] = a\mathit{Var}[X]$$

and, since

$$\mathit{Var}[Y] = (aX - a\mu)^2 = a^2\mathit{Var}[X]$$

it results $\rho_{X,Y} = 1$. As a corollary, $\rho_{X,X} = 1$.

Observe that if X and Y are independent, $p(X, Y) = p(X)p(Y)$: as a consequence, $\mathit{Cov}[X, Y] = 0$ and $\rho_{X,Y} = 0$. That is, independent variables have null covariance and correlation.

The contrary is not true: null correlation does not imply indepedence: see for example X uniform in $[-1, 1]$ and $Y = X^2$.

Correlation matrix

The correlation matrix of $(X_1, \ldots, X_n)^T$ is defined as

$$
\begin{aligned}
\mathbf{\Sigma} &= \begin{bmatrix} \rho_{X_1,X_1} & \rho_{X_1,X_2} & \cdots & \rho_{X_1,X_n} \\ \vdots & \ddots & \vdots & \\ \rho_{X_n,X_1} & \rho_{X_n,X_2} & \cdots & \rho_{X_n,X_n} \end{bmatrix} \\
&= \begin{bmatrix} 1 & \rho_{X_1,X_2} & \cdots & \rho_{X_1,X_n} \\ \vdots & \ddots & \vdots & \\ \rho_{X_n,X_1} & \rho_{X_n,X_2} & \cdots & 1 \end{bmatrix}
\end{aligned}
$$

Definition

Let $x_i \in \mathbb{N}$ for $i = 1, \ldots, k$, then $(x_1, \ldots, x_k) \sim \mathit{Mult}(n, p_1, \ldots, p_k)$ with $0 \leq p \leq 1$, if

$$p(x_1, \ldots, x_k) = \frac{n!}{x_1! \ldots x_k!} \prod_{i=1}^{k} p_i^{x_i} \qquad \text{con} \sum_{i=1}^{k} x_i = n$$

Generalization of the binomial distribution to $k \geq 2$ possible toss results $t_1, \ldots, t_k$ with probabilities $p_1, \ldots, p_k$ ($\sum_{i=1}^{k} p_i = 1$).

Probability that in a sequence of n independent tosses $p_1, \ldots, p_k$, exactly x_i tosses have result t_i ($i = 1, \ldots, k$).

Mean and variance

$$E[x_i] = np_i \qquad \mathit{Var}[x_i] = np_i(1 - p_i) \qquad i = 1, \ldots, k$$

Dirichlet distribution

Definition

Let $x_i \in [0,1]$ for $i = 1, \ldots, k$, then $(x_1, \ldots, x_k) \sim$ *Dirichlet*$(\alpha_1, \alpha_2, \ldots, \alpha_k)$ if

$$f(x_1, \ldots, x_k) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k x_i^{\alpha_i - 1} = \frac{1}{\Delta(\alpha_1, \ldots, \alpha_k)} \prod_{i=1}^k x_i^{\alpha_i - 1}$$

with $\sum_{i=1}^k x_i = 1$.

Generalization of the Beta distribution to the multinomial case $k \geq 2$.

A random variable $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_K)$ with Dirichlet distribution takes values on the $K-1$ dimensional simplex (set of points $\mathbf{x} \in \mathbb{R}^K$ such that $x_i \geq 0$ for $i = 1, \ldots, K$ and $\sum_{i=1}^K x_i = 1$)

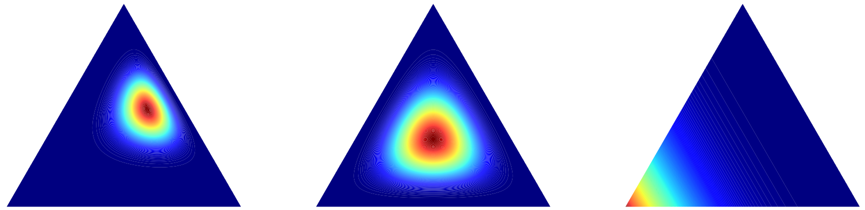
Mean and variance

$$E[x_i] = \frac{\alpha_i}{\alpha_0} \qquad Var[x_i] = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)} \qquad i = 1, \ldots, k$$

with $\alpha_0 = \sum_{j=1}^k \alpha_j$

Dirichlet distribution

Examples of Dirichlet distributions with $k = 3$



Symmetric Dirichlet distribution

Particular case, where $\alpha_i = \alpha$ for $i = 1, \ldots, K$

$$p(\phi_1, \ldots, \phi_K | \alpha, K) = \mathrm{Dir}(\boldsymbol{\phi} | \alpha, K) = \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \prod_{i=1}^{K} \phi_i^{\alpha-1} = \frac{1}{\Delta_K(\alpha)} \prod_{i=1}^{K} \phi_i^{\alpha-1}$$

Mean and variance

In this case,

$$E[x_i] = \frac{1}{K} \qquad Var[x_i] = \frac{K-1}{K^2(\alpha+1)} \qquad i = 1, \ldots, K$$

Gaussian distribution

- Properties
  - Analytically tractable
  - Completely specified by the first two moments
  - A number of processes are asintotically gaussian (theorem of the Central Limit)
  - Linear transformation of gaussians result in a gaussian

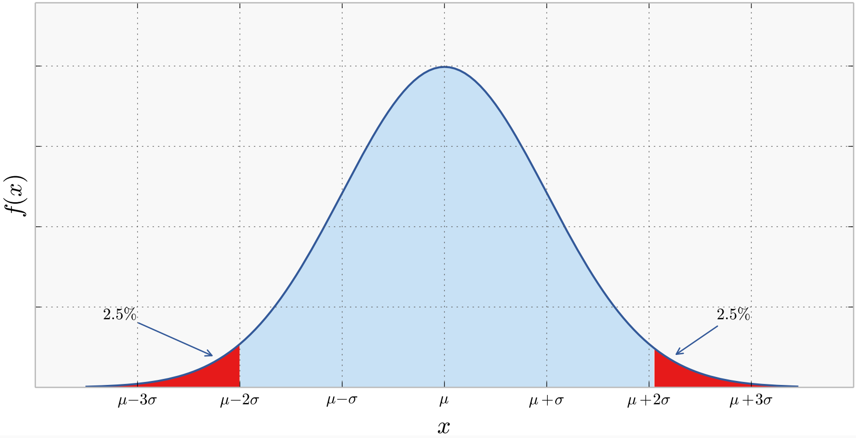
For $x \in \mathbb{R}$:

$$
\begin{aligned}
p(x) &= \mathcal{N}(\mu, \sigma^2) \\
&= \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]
\end{aligned}
$$

with

$$
\begin{aligned}
\mu = E[x] &= \int_{-\infty}^{\infty} x p(x) dx \\
\sigma^2 = E[(x-\mu)^2] &= \int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx
\end{aligned}
$$

Univariate gaussian



A univariate gaussian distribution has about 95% of its probability in the interval $|x - \mu| \geq 2\sigma$.

Multivariate gaussian

For $\mathbf{x} \in \mathbb{R}^d$:

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \end{aligned}$$

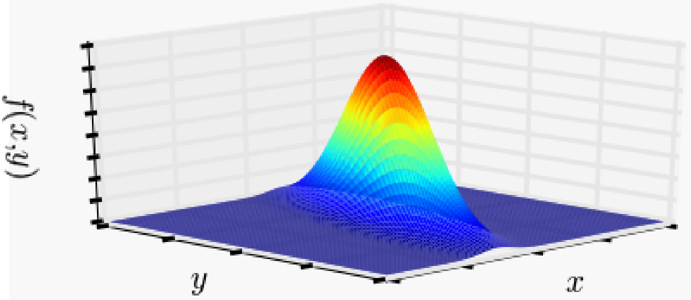
where

$$\boldsymbol{\mu} = E[\mathbf{x}] = \int \mathbf{x} p(\mathbf{x}) d\mathbf{x}$$

$$\boldsymbol{\Sigma} = E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T] = \int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T p(\mathbf{x}) d\mathbf{x}$$

Multivariate gaussian

- $\boldsymbol{\mu}$: expectation (vector of size d)
- $\boldsymbol{\Sigma}$: matrix $d \times d$ of covariance. $\sigma_{ij} = E[(X_i - \mu_i)(X_j - \mu_j)]$

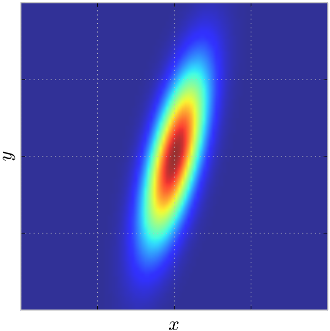


Mahalanobis distance

- Probability is a function of $\mathbf{x}$ through the quadratic form

$$\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

- Δ is the Mahalanobis distance from $\boldsymbol{\mu}$ to $\mathbf{x}$: it reduces to the euclidean distance if $\boldsymbol{\Sigma} = \mathbf{I}$.
- Constant probability on the curves (ellipsis) at constant Δ.



In general,

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = (\mathbf{x}^T\mathbf{A}\mathbf{x})^T = \mathbf{x}^T\mathbf{A}^T\mathbf{x}$$

this implies that

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \frac{1}{2}\mathbf{x}^T\mathbf{A}^T\mathbf{x} = \mathbf{x}^T\left(\frac{1}{2}\mathbf{A} + \frac{1}{2}\mathbf{A}^T\right)\mathbf{x}$$

- $\mathbf{A} + \mathbf{A}^T$ is necessarily symmetric, as a consequence, $\mathbf{\Sigma}$ is symmetric
- as a consequence, its inverse $\mathbf{\Sigma}^{-1}$ does exist.

Assume a diagonal covariance matrix:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

then, $|\mathbf{\Sigma}| = \sigma_1^2 \sigma_n^2 \dots \sigma_n^2$ and

$$\mathbf{\Sigma}^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_n^2} \end{bmatrix}$$

Diagonal covariance matrix

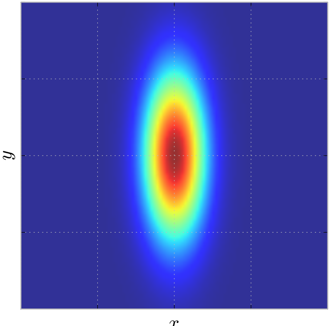
Easy to verify that

$$(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \sum_{i=1}^{n}\frac{(x_i-\mu_i)^2}{\sigma_i^2}$$

and

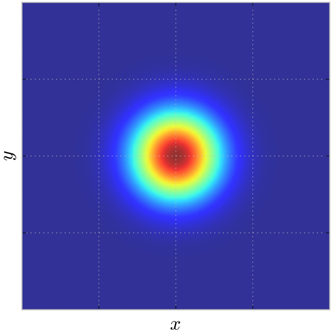
$$f(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{1}{2}\frac{(x_i-\mu_i)^2}{\sigma_i^2}\right)$$

The multivariate distribution turns out to be the product of d univariate gaussians, one for each coordinate x_i.



Identity covariance matrix

The distribution is the product of d ``copies" of the same univariate gaussian, one copy for each coordinate x_i.



Spectral properties of Σ

$\mathbf{\Sigma}$ is real and symmetric: then,

1. all its eigenvalues λ_i are in $\mathbb{R}$
2. there exists a corresponding set of orthonormal eigenvectors $\mathbf{u}_i$ (i.e. such that ($\mathbf{u}_i^T\mathbf{u}_j = 1$ if $i = j$ and 0 otherwise)

Let us define the $d \times d$ matrix $\mathbf{U}$ whose columns correspond to the orthonormal eigenvectors

$$\mathbf{U} = \begin{bmatrix} | & | & & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_d \\ | & | & & | \end{bmatrix}$$

and the diagonal $d \times d$ matrix $\mathbf{\Lambda}$ with eigenvalues on the diagonal

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & & & & \\ & \lambda_2 & & \text{0} & \\ & & \lambda_3 & & \\ & \text{0} & & \ddots & \\ & & & & \lambda_d \end{bmatrix}$$

Decomposition of $\mathbf{\Sigma}$

By the definition of $\mathbf{U}$ and $\mathbf{\Lambda}$, and since $\mathbf{\Sigma}\mathbf{u}_i = \mathbf{u}_i\lambda_i$ for all $i = 1, \ldots, d$, we may write

$$\mathbf{\Sigma}\mathbf{U} = \mathbf{U}\mathbf{\Lambda}$$

Since the eigenvectors u_i are orthonormal, $\mathbf{U}^{-1} = \mathbf{U}^T$ by the properties of orthonormal matrices: as a consequence ,

$$\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T = \sum_{i=1}^{d} \lambda_i \mathbf{u}_i \mathbf{u}_i^T$$

Then, its inverse matrix is a diagonal matrix itself

$$\mathbf{\Sigma}^{-1} = \sum_{i=1}^{d} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^T$$

Density as a function of eigenvalues and eigenvectors

As shown before,

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

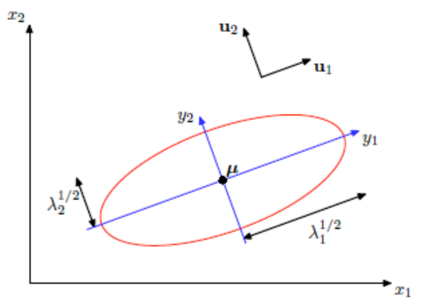
Let $y_i = \mathbf{u}_i^T(\mathbf{x}-\boldsymbol{\mu})$: then

$$(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \sum_{i=1}^n \frac{y_i^2}{\lambda_i}$$

and

$$f(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\lambda_i}}\exp\left(-\frac{1}{2}\frac{y_i^2}{\lambda_i}\right)$$

y_i is the scalar product of $\mathbf{x} - \boldsymbol{\mu}$ and the i-th eigenvector $\mathbf{u}_i$, that is the length of the projection of $\mathbf{x} - \boldsymbol{\mu}$ along the direction of the eigenvector. Since eigenvectors are orthonormal, they are the basis of a new space, and for each vector $\mathbf{x} = (x_1, \ldots, x_d)$, the values $(y_1, \ldots, y_d)$ are the coordinates of $\mathbf{x}$ in the eigenvector space.



Eigenvectors of $\boldsymbol{\Sigma}$ correspond to the axes of the distribution; each eigenvalue is a scale factor along the axis of the corresponding eigenvector.

Linear transformations

Let $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{A} \in \mathbb{R}^{d \times k}$, $\mathbf{y} = \mathbf{A}^T \mathbf{x} \in \mathbb{R}^k$: then, if $\mathbf{x}$ is normally distributed, so is $\mathbf{y}$.

In particular, if the distribution of $\mathbf{x}$ has mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, the distribution of $\mathbf{y}$ has mean $\mathbf{A}^T \boldsymbol{\mu}$ and covariance matrix $\mathbf{A}^T \boldsymbol{\Sigma} \mathbf{A}$.

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \implies \mathbf{y} \sim \mathcal{N}(\mathbf{A}^T \boldsymbol{\mu}, \mathbf{A}^T \boldsymbol{\Sigma} \mathbf{A})$$

Marginal and conditional of a joint gaussian

Let $\mathbf{x}_1 \in \mathbb{R}^h$, $\mathbf{x}_2 \in \mathbb{R}^k$ be such that $\left[\begin{array}{c} \mathbf{x}_1 \\ \hline \mathbf{x}_2 \end{array}\right] \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ and let

- $\boldsymbol{\mu} = \left[\begin{array}{c} \boldsymbol{\mu}_1 \\ \hline \boldsymbol{\mu}_2 \end{array}\right]$ with $\boldsymbol{\mu}_1 \in \mathbb{R}^h, \boldsymbol{\mu}_2 \in \mathbb{R}^k$
- $\boldsymbol{\Sigma} = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right]$ with $\boldsymbol{\Sigma}_{11} \in \mathbb{R}^{h \times h}$, $\boldsymbol{\Sigma}_{12} \in \mathbb{R}^{h \times k}$, $\boldsymbol{\Sigma}_{21} \in \mathbb{R}^{k \times h}$, $\boldsymbol{\Sigma}_{22} \in \mathbb{R}^{k \times k}$

then

- the marginal distribution of $\mathbf{x}_1$ is $\mathbf{x}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$
- the conditional distribution of $\mathbf{x}_1$ given $\mathbf{x}_2$ is $\mathbf{x}_1 | \mathbf{x}_2 \sim \mathcal{N}(\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2})$ with

$$\begin{aligned} \boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\ \boldsymbol{\Sigma}_{1|2} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \end{aligned}$$

Let $\mathbf{x}, \mathbf{y}$ be such that

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1) \qquad \text{and} \qquad \mathbf{y}|\mathbf{x} \sim \mathcal{N}(\mathbf{A}\mathbf{x} + \mathbf{b}, \boldsymbol{\Sigma}_2)$$

That is, the marginal distribution of $\mathbf{x}$ (the prior) is a gaussian and the conditional distribution of $\mathbf{y}$ w.r.t. $\mathbf{x}$ (the likelihood) is also a gaussian with (conditional) mean given by a linear combination on $\mathbf{x}$. Then, both the the conditional distribution of $\mathbf{x}$ w.r.t. $\mathbf{y}$ (the posterior) and the marginal distribution of y (the evidence) are gaussian.

$$\begin{aligned}
\mathbf{y} &\sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \boldsymbol{\Sigma}_2 + \mathbf{A}\boldsymbol{\Sigma}_1\mathbf{A}^T) \\
\mathbf{x}|\mathbf{y} &\sim \mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})
\end{aligned}$$

where

$$\begin{aligned}
\hat{\boldsymbol{\mu}} &= (\boldsymbol{\Sigma}_1^{-1} + \mathbf{A}^T\boldsymbol{\Sigma}_2^{-1}\mathbf{A})^{-1}(\mathbf{A}^T\boldsymbol{\Sigma}_2^{-1}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}) \\
\hat{\boldsymbol{\Sigma}} &= (\boldsymbol{\Sigma}_1^{-1} + \mathbf{A}^T\boldsymbol{\Sigma}_2^{-1}\mathbf{A})^{-1}
\end{aligned}$$

Maximum likelihood and gaussians

Given a d-dimensional dataset $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, we estimate by maximum likelihood the parameters of a gaussian distribution modeling $\mathbf{X}$.

The log-likelihood of $\mathbf{X}$ is

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{nd}{2}\ln(2\pi) - \frac{n}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})$$

which is maximized for

$$\boldsymbol{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$$

and

$$\boldsymbol{\Sigma}_{ML} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu}_{ML})(\mathbf{x}_i - \boldsymbol{\mu}_{ML})^T$$

While $E[\boldsymbol{\mu}_{ML}] = \boldsymbol{\mu}$, $E[\boldsymbol{\Sigma}_{ML}] = \frac{n-1}{n}\boldsymbol{\Sigma}$.

A better (unbiased) estimator of $\boldsymbol{\Sigma}$ is

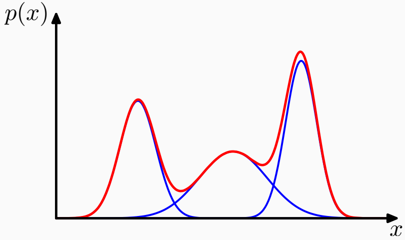
$$\tilde{\boldsymbol{\Sigma}}_{ML} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu}_{ML})(\mathbf{x}_i - \boldsymbol{\mu}_{ML})^T$$

Mixture of gaussians

Convex combination of gaussian components

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k p_k(\mathbf{x})$$

with $\pi_k \geq 0$, $\sum_{i=1}^{K} \pi_k = 1$, $\int p_k(\mathbf{x}) d\mathbf{x} = 1$



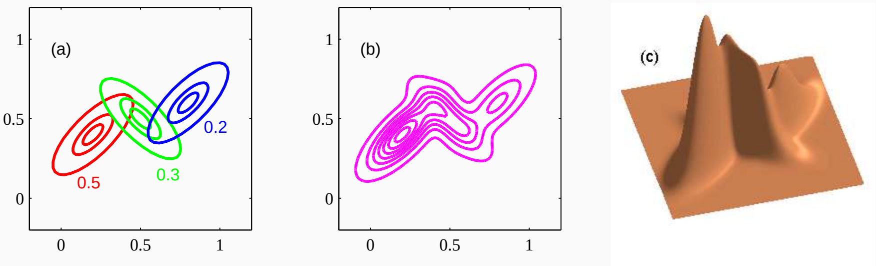
Mixture of gaussians

π_k are the mixing coefficients: they can be seen as probabilities, since $0 \leq \pi_k \leq 1$ and $\sum_{i=1}^K \pi_k = 1$.

Since, in general,

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(k)p(\mathbf{x}|k)$$

it derives $\pi_k = p(k)$: that is, π_k is the prior probability of the k-th component



Mixture of gaussians: generative interpretation

Assume a dataset is generated by random sampling from mixture of gaussian. Then, for each $\mathbf{x}$:

- a component p_k is picked with probability π_k: this is the (prior) probability that a point is generated by the k-th component
- the point is sampled with (conditional) probability $p_k(\mathbf{x})$
- the overall probability that a point $\mathbf{x}$ is sampled is the marginal $p(\mathbf{x})$
- the posterior probability of a component $p(k|\mathbf{x})$ is the probability that a point $\mathbf{x}$ has been generated by sampling the k-th component $p_k(\mathbf{x})$

By Bayes' rule,

$$p(k|\mathbf{x}) = \frac{p_k(\mathbf{x})p(k)}{p(\mathbf{x})} = \frac{p_k(\mathbf{x})p(k)}{\sum_{i=1}^{K} p_i(\mathbf{x})p(i)}$$