

IFT6390

Fondements de l'apprentissage machine

Probability distributions

The multivariate Gaussian

Professor: Ioannis Mitliagkas

Slides: Pascal Vincent

Distributions

Reminder: https://en.wikipedia.org/wiki/Random_variable

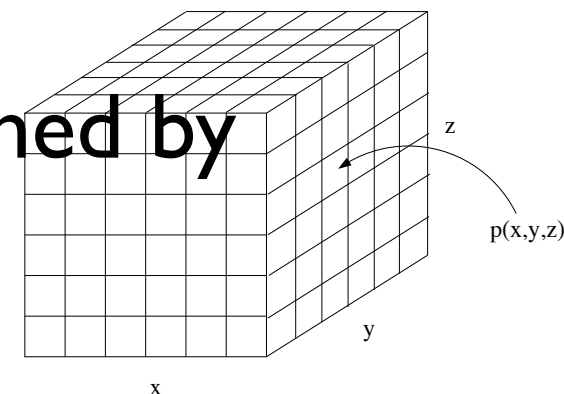
- A **probability distribution** over a random variable is a function that provides the probabilities of occurrence of different possible values of the variable.

- A distribution over a random variable can be given by its **cumulative distribution function (c.d.f.)**:

$$F_X(x) = P(X \leq x) = P(X_1 \leq x_1, \dots, X_n \leq x_n)$$

- The distribution of a **discrete variable** is determined by **the probability of each value** it can take.

=> probability table (must sum to one...).



- The distribution of a **continuous variable** can be given by its **probability density function (p.d.f.)** which is the derivative of the c.d.f.** *The probability that a draw falls within some region is equal to the **integral** of the p.d.f. over this region.*

Operations on distributions

Given a distribution, we may want to

- **Generate data**, i.e. draw samples from this distribution.
- **Compute the probability/likelihood of a configuration** (e.g. knowing the value of some of the variables, after marginalizing the unknown variables).
- **Inference**: *infer* the most likely value or the expectation of some variables given the values of other variables.
- **Learn** the parameters of a distribution **given a data set** (such that the likelihood of the data being generated by this distribution with these parameters is maximized: *maximum likelihood estimation*).

Ex. Discrete variable X

Probability table:

x	$P(X=x)$
1	0.10
2	0.80
3	0.10

Generate



Data set:

2
2
2
1
2
2
2
2
3
2
2
⋮

Learn



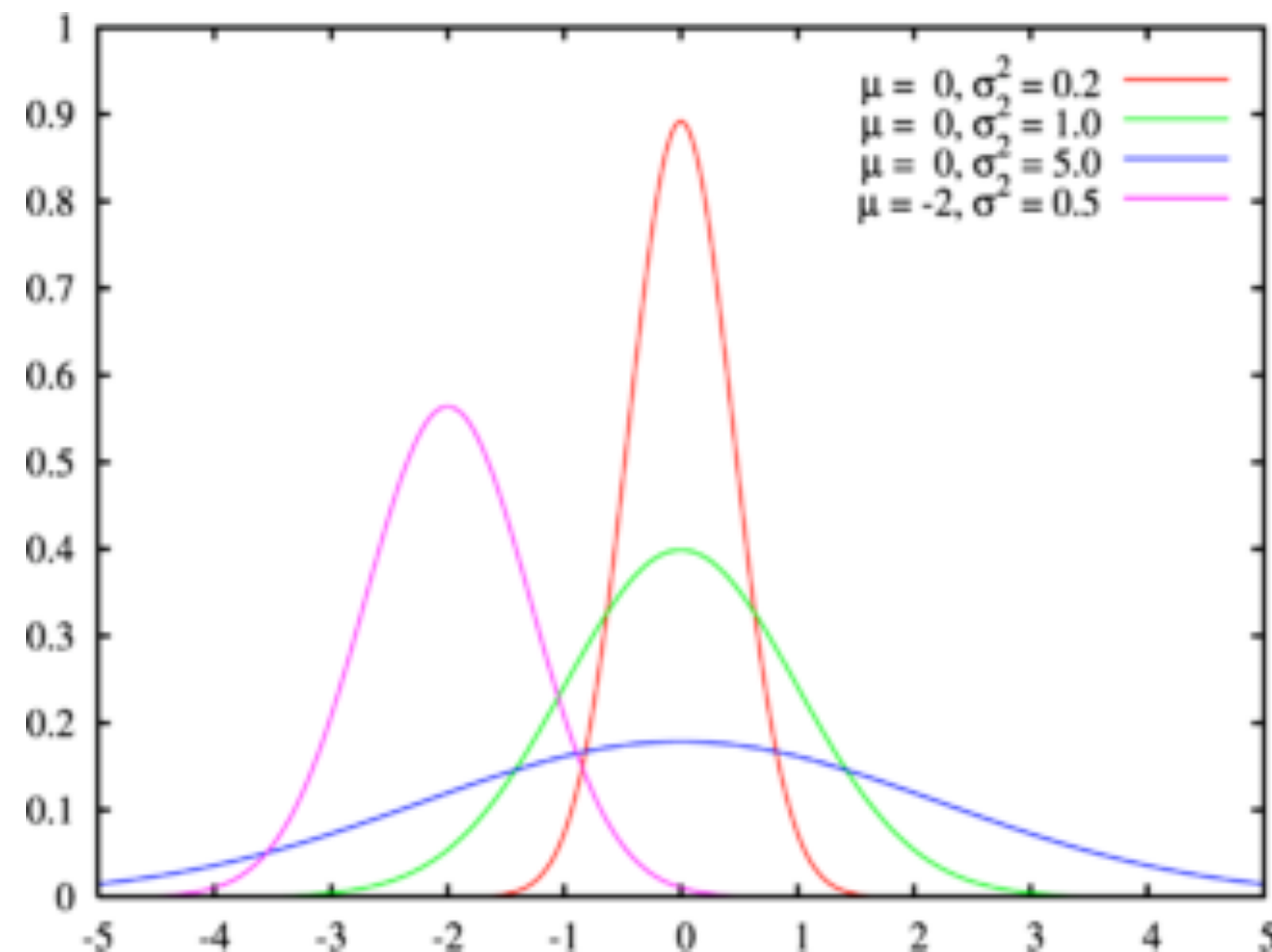
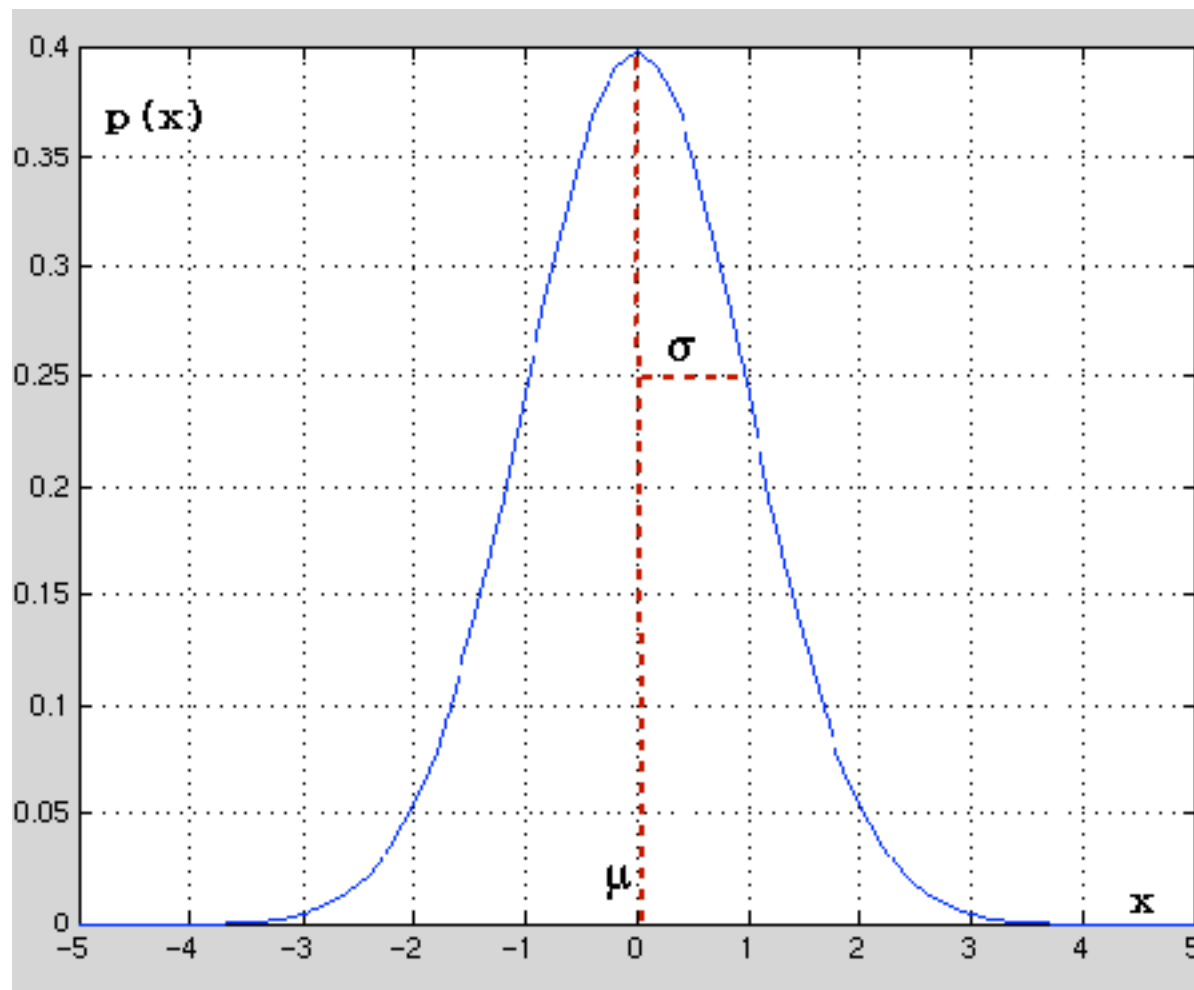
Ex. continuous (scalar) variable x

Univariate Gaussian/
Normal distribution

Gaussian density

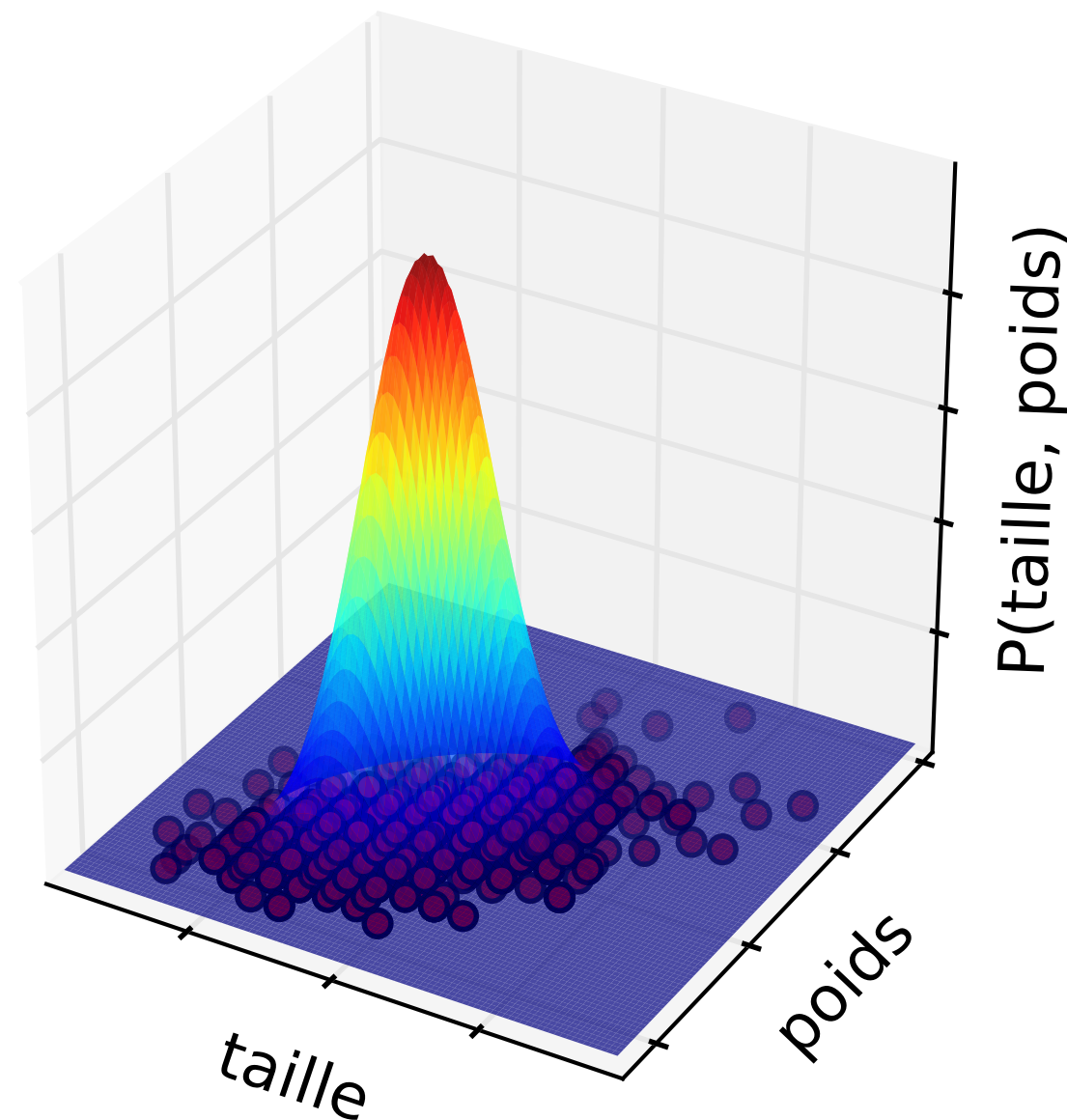
Univariate (i.e. one dimensional) Gaussian density with mean μ and variance σ^2 (standard deviation σ).

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



Ex. continuous vector variable x

Multivariate Gaussian/normal distribution

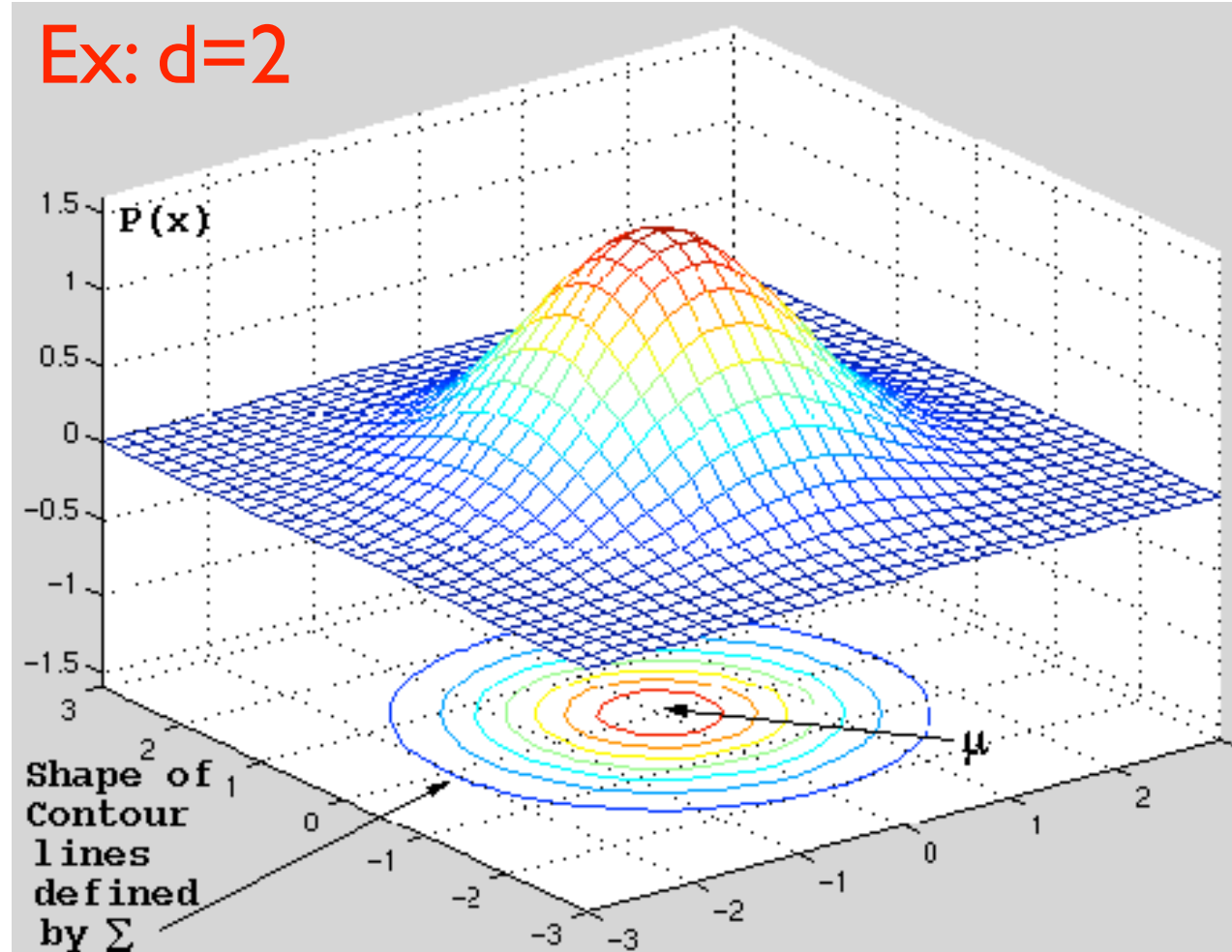


Multivariate Gaussian

Isotropic (“spherical”) Gaussian in d dimensions:

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

Gaussian “hill” “**centered**” in μ with “**width**” σ , (same width in all directions)



General Gaussian distribution in d dimensions, with mean μ and covariance matrix Σ .

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

$$\mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_d \end{bmatrix} \quad \Sigma = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1d} \\ \vdots & & \vdots \\ \sigma_{d1} & \cdots & \sigma_{dd} \end{bmatrix}$$

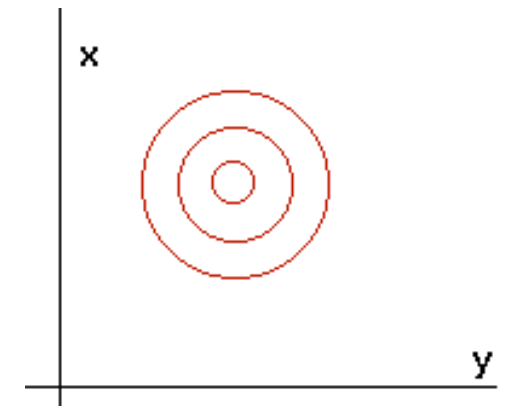
} determinant of Σ

Note: this denominator is only the normalization constant (assuring that the density integrates to 1).

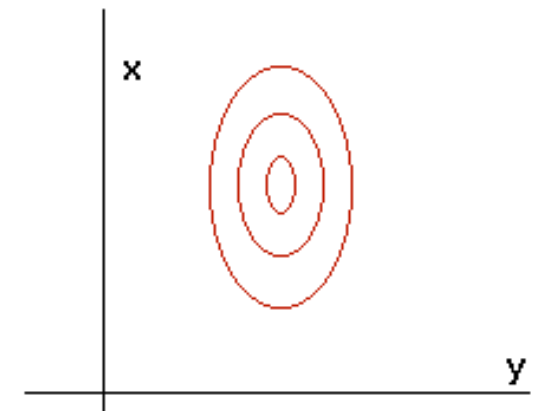
Multivariate Gaussian

example of covariance matrices

- Isotropic/spherical Gaussian: $\Sigma = \sigma^2 I$
(I is the identity matrix)



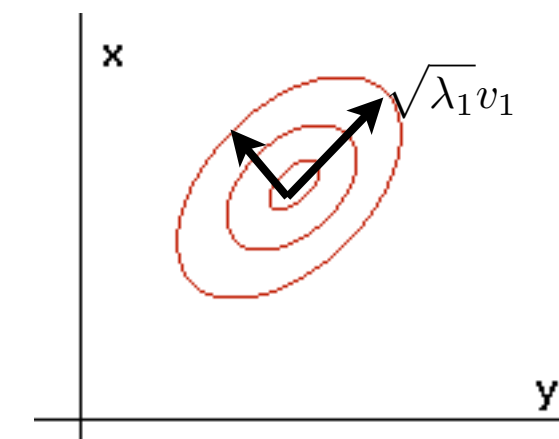
- Diagonal Gaussian: $\Sigma = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \sigma_2^2 & \\ 0 & & \ddots \\ & & & \sigma_d^2 \end{pmatrix}$



- Eigendecomposition:

$$\Sigma = \sum_{i=1}^d \lambda_i v_i v_i^T$$

The eigenvectors correspond to the ellipsoid axes, and the eigenvalues to the corresponding widths...



The determinant $|\Sigma| = \lambda_1 \lambda_2 \dots \lambda_d$ gives the "size" of the ellipsoid...

Multivariate Gaussian

Learning the parameters

- We can easily **learn the parameters of a Gaussian distribution** from a data set:
- μ is estimated by the **empirical mean** (“centroid” of the training points).

$$\mu = \frac{1}{n} \sum_{t=1}^n \mathbf{x}_t$$

- Σ is estimated by the **empirical covariance matrix**:

$$\Sigma_{ij} = \frac{1}{n} \sum_{t=1}^n (\mathbf{x}_{ti} - \mu_i)(\mathbf{x}_{tj} - \mu_j) \quad \text{or} \quad \Sigma = \frac{1}{n} \sum_{t=1}^n (\mathbf{x}_t - \mu)(\mathbf{x}_t - \mu)'$$

- We will see later in the course how to derive these formulas (*maximum likelihood principle*).