COMPSCI 4ML3 Tutorial 4: Review of Probability Theory

Slides by Alireza Fathollah Pour

McMaster University

Winter 2024

Basic Elements I

- **Sample space** Ω: The set of all possible outcomes.
- **Event space** $\mathcal{F}$: The set containing all possible subsets of outcomes. i.e., A collection of possible outcomes
- **Event** A: Any element of the event space. $\forall A \in \mathcal{F},\ A \subseteq \Omega$

For the event of rolling a dice:

- $\Omega = \{1, 2, 3, 4, 5, 6\}$
- $\mathcal{F} = \{\{1\}, \ldots, \{6\}, \{1, 2\}, \ldots, \{5, 6\}, \{1, 2, 3\}, \ldots, \{1, 2, 3, 4, 5, 6\}\}$
- An example of an event is $A = \{2, 3, 6\}$

Basic Elements II

- **Probability measure** P: A funtion $P : \mathcal{F} \rightarrow \mathcal{R}$ that satisfies the following properties:
- $P(A) \geq 0, \forall A \in \mathcal{F}$
- $P(\Omega) = 1$
- For a collection of disjoint events A_i i.e., $(\forall i \neq j, A_i \cap A_j = \emptyset)$ we have

$$P(\bigcup_i A_i) = \sum_i P(A_i)$$

Probability Measure: Properties

- If $A \subseteq B$, $P(A) \leq P(B)$
- $P(A \cup B) \leq P(A) + P(B)$, which is called *Union Bound*
- $P(A \cap B) \leq \min(P(A), P(B))$
- $P(A^c) = 1 - P(A)$
- For disjoint events $A_1, \ldots, A_k$ such that $\cup_{i=1}^{k} A_i = \Omega$

$$\sum_{i=1}^{k} P(A_i) = 1,$$

which is also called the *law of total probability*.

Conditional Probability and Independence

- The **conditional probability** $P(A|B)$ is the probability of observing event A after the occurrence of B

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- Two events A and B are **independent** iff $P(A \cap B) = P(A)P(B)$. i.e, observing B does not give any information about occurrence of A and $P(A|B) = P(A)$

Conditional Probability and Independence

Example: Probability of a person's weight being y, given that her height is x.

$$P(\text{weight} = y|\text{height} = x)$$

These two features are correlated.

$$P(\text{weight} = 200lb \quad | \quad \text{height} = 190cm) = 0.2$$
$$P(\text{weight} = 200lb \quad | \quad \text{height} = 140cm) = 0.01$$

Bayes' Rule

- For two events A and B

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$
$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

- This implies that

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$
$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Chain Rule and Law of Total Probability

- For events $A_1, \ldots, A_n$, **chain rule** states that

$$P(A_n \cap \ldots \cap A_1) = P(A_n | A_{n-1} \cap \ldots \cap A_1) P(A_{n-1} \cap \ldots \cap A_1) = P(A_1) \prod_{i=2}^{n} (A_i | \bigcap_{k=1}^{i-1} A_k)$$

- If $B_1, \ldots, B_n$ are finite partition of the sample space (i.e., $\forall i \neq j, B_i \cap B_j = \emptyset$ and $\cup_{i=1}^{n} B_i = \Omega$), the **law of total probability** states that for an event A

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A|B_i) P(B_i)$$

Random Variables

A real-valued random variable X is a mapping from sample space to real values, i.e., $X : \Omega \to \mathbb{R}$, which assigns to each element $\omega \in \Omega$ a real value $X(w)$

A random variable helps us describe some functions of observed events

- We usually denote random variables with capital letters $X(\omega)$ and simply denote it with X
- We usually use small letters for the value that a random variable may take. i.e., we write $X = x$ instead of $X(\omega) = x$

Random Variables: Example

Example: We toss coin for 20 times. What is the probability that we observe 6 heads?

- Sample space Ω can be defined as the sequences of heads and tails with length 20
- Random variable X is a function that assigns to each sequence $\omega \in \Omega$ the number of heads in that sequence. i.e., $X(\omega) =$ number of heads in ω
- We are interested in finding $P(X(\omega) = 6)$ or simply $P(X = 6)$

Random Variables

- A random variable that only takes finite number of values is called a **discrete random variable**
  - The probability that a random variable X takes value x is

$$P(X = x) := P(\{\omega \in \Omega : X(\omega) = x\})$$

- A random variable that can take infinite number of values is called a **continuous random variable**
  - The probability that a random variable X takes values between a and b is

$$P(a \leq X \leq b) := P(\{\omega \in \Omega : a \leq X(\omega) \leq b\})$$

Cumulative Distribution Function

For a random variable X, we can define $P(X \leq x)$ as a function of x:

- The **Cumulative Distribution Function (CDF)** is a function $F_X(x) : \mathbb{R} \rightarrow [0, 1]$ that is defined as

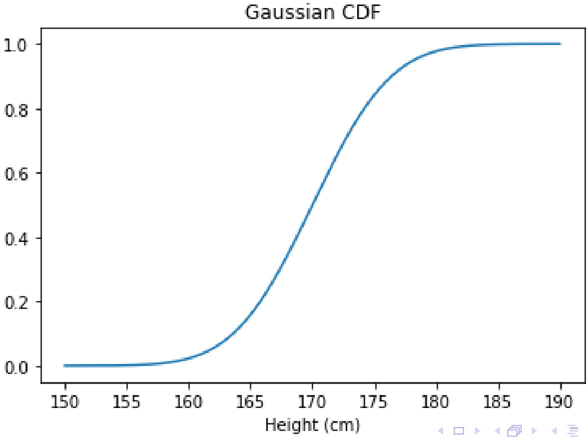
$$F_X(x) := P(X \leq x)$$

Properties:

- $0 \leq F_X(x) \leq 1$
- $P(a \leq X \leq b) = F_X(b) - F_X(a)$

Cumulative Distribution Function

Example:



Probability Mass Function

For a <u>discrete</u> random variable, the **Probability Density Function(PMF)** $p_X(x) : \mathbb{R} \to [0, 1]$ is a function that returns the probability of a random variable taking a specific value

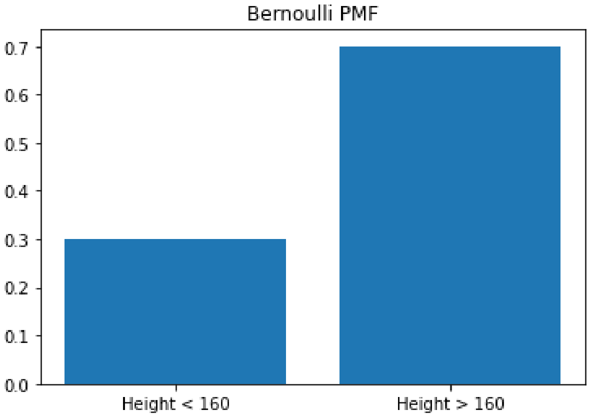
$$p_X(x) := P(X = x)$$

Properties:

- $0 \leq p_X(x) \leq 1$
- $\sum_{x \in \mathbb{D}} p_X(x) = 1$, where $\mathbb{D}$ is the set of all possible values that X can take.
- $P(X \in A) = P(\{\omega : X(\omega) \in A\}) = \sum_{x \in A} p_X(x)$

Probability Mass Function

Example:



Probability Density Function

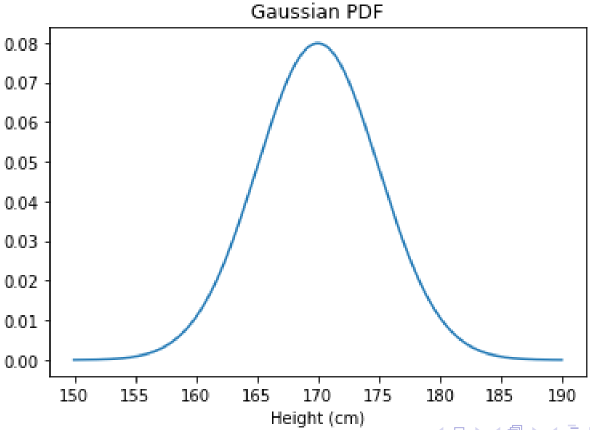
For a continuous random variable, we are interested in $P(x \leq X \leq x + \Delta x)$ when $\Delta \rightarrow 0$.
If $F_X(x)$ is differentiable everywhere, the **Probability Density Function (PDF)** $f_X(x)$ is the derivative of the CDF function

$$f_X(x) := \frac{dF_X(x)}{dx}$$

- $P(x \leq X \leq x + \Delta x) \approx f_X(x)\Delta x$
- Unlike PMF, $f_X(x)$ is not the probability that the random variable X takes a value x. i.e., $f_X(x) \neq P(X = x)$. In fact, for a continuous distribution, the probability that the random variable takes a specific value is zero. i.e, P(X=x)=0

Probability Density Function

Example:



PDF: Properties

- $f_X(x) \geq 0$
- $\int_{-\infty}^{\infty} f_X(x) = 1$
- $F_X(x) = \int_{-\infty}^{x} f_X(x)dx$

Expectation

- For a *discrete* random variable with PMF $p_X(x)$ and a function $g(x) : \mathbb{R} \to \mathbb{R}$, $g(X)$ can be considered as a random variable and the **expectation** or **expected value** of $g(X)$ is defined as

$$\mathbb{E}[g(X)] = \sum_{x \in \mathbb{D}} g(x) p_X(x)$$

- For a *continuous* random variable with PDF $f_X(x)$, the **expectation** or **expected value** of $g(X)$ is defined as

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

Mean and Variance

- Setting $g(x) = x$, the **mean** of a random variable X is defined as
$$\mu = \mathbb{E}[X] = \int_{-\infty}^{\infty} x\, f_X(x) dx$$
- The **variance** of a random variable X is a measure of how concentrated the random variable is around its mean
$$\begin{aligned}\sigma^2 = \text{Var} = \mathbb{E}[(X - \mathbb{E}[X])^2] &= \mathbb{E}[X^2 + (\mathbb{E}[X])^2 - 2X\mathbb{E}[X]] \\ = \mathbb{E}[X^2] + (\mathbb{E}[X])^2 - 2\mathbb{E}[X\mathbb{E}[X]] &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2\end{aligned}$$

Mean and Variance: Example I

Example Find the mean and variance of rolling a dice with equal probability for each face

$$\mu = \mathbb{E}[X] = \sum_{i=1}^{6} iP(X = i) = \sum_{i=1}^{6} i\frac{1}{6} = \frac{21}{6} = 3.5$$

$$\sigma^2 = \mathbb{E}[(X - \mu)^2] = \sum_{i=1}^{6} (i - 3.5)^2 P(X = i)$$

$$= \sum_{i=1}^{6} (i - 3.5)^2 \frac{1}{6} = \frac{35}{12} \approx 2.92$$

Mean and Variance: Example II

Example Find the mean and variance of a random variable with PDF $f_X(x) = 3x^2$, $0 \leq x \leq 1$

$$\mu = \mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx = \int_0^1 x f_X(x) dx = \int_0^1 3x^3 dx = \frac{3x^4}{4}\Big|_0^1 = \frac{3}{4}$$

$$\sigma^2 = \mathbb{E}[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x) dx = \int_0^1 (x - \frac{3}{4})^2 3x^2 dx$$

$$= \int_0^1 (x - \frac{3}{4})^2 3x^2 dx = \frac{3}{16}$$

Expectation: Properties

- $\mathbb{E}[c] = c, \forall c \in \mathbb{R}$
- $\mathbb{E}[cg(X)] = c\mathbb{E}[g(X)], \forall c \in \mathbb{R}$
- $\mathbb{E}[f(X) + g(X)] = \mathbb{E}[f(X)] + \mathbb{E}[g(X)]$

Variance: Properties

- $\text{Var}(c) = 0, \forall c \in \mathbb{R}$
- $\text{Var}(f(X) + c) = \text{Var}(f(X)), \forall c \in \mathbb{R}$
- $\text{Var}(cf(X)) = c^2\text{Var}(f(X)), \forall c \in \mathbb{R}$

Discrete Random Variables: Bernoulli

- $X \sim \text{Bernoulli}(p)$, where $0 \leq p \leq 1$

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

  - $\mathbb{E}[X] = p$
  - $\text{Var}(X) = p(1 - p)$

Discrete Random Variables: Binomial

- $X \sim \text{Binomial}(n, p)$, where $0 \leq p \leq 1$

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

  - $\mathbb{E}[X] = np$
  - $\text{Var}(X) = np(1-p)$

Discrete Random Variables: Poisson

- $X \sim \text{Possion}(\lambda)$, where $\lambda > 0$

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

  - $\mathbb{E}[X] = \lambda$
  - $\text{Var}(X) = \lambda$

Continuous Random Variables: Uniform

- $X \sim U_{[a,b]}$, where $a \leq b$

$$f_X(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

  - $\mathbb{E}[X] = \dfrac{b+a}{2}$
  - $\text{Var}(X) = \dfrac{(b-a)^2}{12}$

Continuous Random Variables: Exponential

- $X \sim \text{Exponential}(\lambda)$, where $\lambda > 0$

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

  - $\mathbb{E}[X] = \frac{1}{\lambda}$
  - $\text{Var}(X) = \frac{1}{\lambda^2}$

Continuous Random Variables: Gaussian/Normal

- $X \sim \mathcal{N}(\mu, \sigma^2)$

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

  - $\mathbb{E}[X] = \mu$
  - $\text{Var}(X) = \sigma^2$

Joint Distribution

Example:

X = image :

Y = label (1 if image contains cat and 0 otherwise)

What is the probability of an image contains a cat?

$$P(X = \text{image}, Y = 0) = ?$$
$$P(X = \text{image}, Y = 1) = ?$$

Joint Cumulative Distributions

It happens that we need to consider two random variables X and Y together and discuss X and Y at the same time during a random experiment.

- The **joint cumulative distribution** function for random variables X and Y is defined as

$$F_{X,Y}(x,y) = P(X \leq x,\ Y \leq y)$$

- The marginal CDFs can be found by

$$F_X(x) = \lim_{y\to\infty} F_{X,Y}(x,y)$$
$$F_Y(y) = \lim_{x\to\infty} F_{X,Y}(x,y)$$

Joint CDF: Properties

- $0 \leq F_{X,Y}(x, y) \leq 1$
- $\lim_{x,y \to \infty} F_{X,Y}(x, y) = 1$
- $\lim_{x,y \to -\infty} F_{X,Y}(x, y) = 0$

Joint Probability Mass Function

- The **joint probability mass function** for *discrete* random variables X and Y is defined as

$$p_{X,Y}(x, y) = P(X = x,\ Y = y)$$

- The marginal PMFs can be found by

$$p_X(x) = \sum_{y \in \mathbb{D}_y} p_{X,Y}(x, y)$$

$$p_Y(y) = \sum_{x \in \mathbb{D}_y} p_{X,Y}(x, y)$$

Joint PMF: Properties

- $0 \leq p_{X,Y}(x,y) \leq 1$
- $\sum_{x \in \mathbb{D}_x, y \in \mathbb{D}_y} p_{X,Y}(x,y) = 1$

Joint Probability Density Function

- If the joint CDF is differentiable everywhere in x and y, the **joint probability density function** for *continuous* random variables X and Y is defined as

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}$$

- The marginal PDFs can be found by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$$
$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$$

Joint PDF: Properties

- $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy = 1$
- $\int\int_{A} f_{X,Y}(x,y)dxdy = P((X,Y) \in A)$

Conditional Distributions

- **Conditional PMF** refers to the probability distribution over X when we know that Y has taken a certain value (if $p_Y(y) \neq 0$)

$$p_{X|Y}(x|y) = P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{p_{X,Y}(x, y)}{p_Y(y)}$$

- **Conditional PDF** is defined as (if $f_Y(y) \neq 0$)

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

Independent Random Variables

Two random variables X and Y are independent iff

$$F_{X,Y}(x|y) = F_X(x)F_Y(y),\ \forall x, y$$

- If two *discrete random variables* X and Y are independent

$$p_{X,Y}(x,y) = p_X(x)p_Y(y),\ \forall x, y$$
$$p_{X|Y}(x|y) = p_X(x),\ \forall x, y \text{ such that } p_Y(y) \neq 0$$

- If two *continuous random variables* X and Y are independent

$$f_{X,Y}(x,y) = f_X(x)f_Y(y),\ \forall x, y$$
$$f_{X|Y}(x|y) = f_X(x),\ \forall x, y \text{ such that } f_Y(y) \neq 0$$

Bayes' Rule for Joint Probability Distribution

- For two *discrete* random variables X and Y

$$p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)p_X(x)}{p_Y(y)}$$

- For two *continuous* random variables X and Y

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)}$$

Expectation of Joint Distributions

- For *discrete* random variables X and Y with joint PMF $p_{X,Y}(x,y)$ and a function $g(x,y) : \mathbb{R}^2 \to \mathbb{R}$, $g(X,Y)$ can be considered as a random variable and the **expectation** or **expected value** of $g(X,Y)$ is defined as

$$\mathbb{E}[g(X,Y)] = \sum_{x \in \mathbb{D}_x} \sum_{y \in \mathbb{D}_y} g(x,y) p_{X,Y}(x,y)$$

- For *continuous* random variables X and Y with joint PDF $f_{X,Y}(x,y)$, the **expectation** or **expected value** of $g(X,Y)$ is defined as

$$\mathbb{E}[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y) dxdy$$

Covariance of Joint Distributions

- The covariance of two random variables X and Y is defined as

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

- If $\mathrm{Cov}(X, Y) = 0$, two random real-valued variables are called **uncorrelated**.

Expectation and Covariance: Properties

- $\mathbb{E}[f(X,Y)+g(X,Y)] = \mathbb{E}[f(X,Y)] + \mathbb{E}[g(X,Y)]$
- $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y)$
- If X and Y are independent, $\text{Cov}(X,Y) = 0$
- If X and Y are independent, $\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)]$

Generalized Joint Distribution for n Variables

For n random variables $X_1, \ldots, X_n$

- Joint CDF is defined as

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = P(X_1 \le x_1, \ldots, X_n \le x_n)$$

- For *discrete* random variables joint PMF is defined as

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n)$$

- For *continuous* random variables joint PDF is defined as

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \frac{\partial^n F_{X_1,\ldots,X_n}(x_1, \ldots, x_n)}{\partial x_1 \ldots \partial x_n}$$

Joint Distribution for n Variables

For n randaom variable $X_1, \ldots, X_n$

- Marginal PDFs can be derived by

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) dx_2 \ldots dx_n$$

-

$$P((X_1, \ldots, X_n) \in A) = \int \cdots \int_A f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) dx_1 \ldots dx_n$$

- $X_1, \ldots, X_n$ are mutually independent iff

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

Random Vectors

When dealing with n random variables, we can consider them as a random vector $X = [X_1, \ldots, X_n]^T$

- For the random vector X, the expectation is in the form of a vector. For a function $g : \mathbb{R}^n \to \mathbb{R}^m$

$$\mathbb{E}[g(X)] = \begin{bmatrix} \mathbb{E}[g_1(X)] \\ \vdots \\ \mathbb{E}[g_m(X)] \end{bmatrix}$$

- The mean vector is $\boldsymbol{\mu} = \mathbb{E}[X] = [\mathbb{E}[X_1], \ldots, \mathbb{E}[X_n]]^T$

Covariance Matrix

For a random vector $X \in \mathbb{R}^n$, its covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$ is a symmetric positive semidefinite matrix, where $\Sigma_{i,j} = \text{Cov}(X_i, X_j)$

$$\Sigma = \begin{bmatrix} \text{Cov}(X_1, X_1) & \text{Cov}(X_1, X_2) & \dots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \text{Cov}(X_2, X_2) & \dots & \text{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \text{Cov}(X_n, X_2) & \dots & \text{Cov}(X_n, X_n) \end{bmatrix}$$
$$= \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T] = \mathbb{E}[XX^T] - \mathbb{E}[X]\mathbb{E}[X]^T$$

Note $\text{Cov}(X_i, X_i) = \text{Var}(X_i)$

Multivariate Gaussian Distribution

A multivariate Gaussian random variable $X \sim \mathcal{N}(\mu, \Sigma)$ can be defined as

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \frac{1}{\sqrt{(2\pi)^n}|\Sigma|^{1/2}} \exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$$

If variables $X_1,\ldots,X_n$ are uncorrelated, the covariance matrix Σ will become a diagonal matrix with variances of individual variables in its main diagonal. In this case,

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp(-\frac{(x_i-\mu_i)^2}{2\sigma_i^2})$$