# Probability and Statistics review

Francisco Pereira

Filipe Rodrigues

DTU Management Engineering
Department of Management Engineering

# Outline

- Introduction
- Random variable, atom, and event
- Joint distribution
- Conditional probability
- Bayes theorem
- Independence
- Expectation
- Continuous random variables

(Based on David MacKay, David Blei,
https://www.cs.princeton.edu/courses/archive/spring12/cos424/pdf/lecture02.pdf)

# Introduction

Consider the “card problem”

- There are three cards:
  - Red/Red
  - Red/Blue
  - Blue/Blue
- I go through the following process
  1. Close my eyes and pick a card
  2. Pick a side at random
  3. Show you that side
- I show you Red. What’s the probability the other side is Red too?

# Random variable, atom, and event

- In Algebra a variable, $x$, is an unknown value
  - E.g. $2x = 4$
  - It can take at most one value at a time
- A **random variable** represents simultaneously a set of values
- Necessary in contexts where we *cannot* determine a unique value
  - Of course, theoretically, it also corresponds to one value...
  - But we can only determine its distribution
  - E.g. $p(5 < X < 10) = 0.5$
- It can be a single value, a vector, a matrix...

# Random variable, atom, and event

- Random variables take on values in a sample space
- They can be discrete or continuous
- For example:
  - Coin flip: $\{H, T\}$
  - Height: Positive values $(0, \infty)$
  - Temperature: real values $(-\infty, \infty)$
  - Number of words in a document: Positive integers $\{1, 2, ..., \infty\}$
- We call the values of random variables *atoms*

# Random variable, atom, and event

- A *discrete probability distribution* assigns probability to every atom in the sample space
- For example, if $X$ is an (unfair) coin, then
  - $p(X = H) = 0.7$
  - $p(X = T) = 0.3$
- The sum of probabilities of *any* distribution is 1

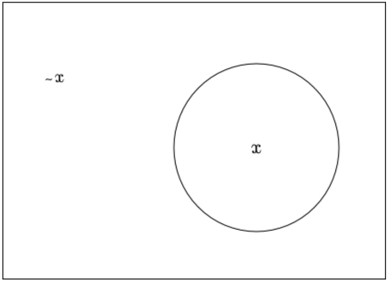
$$\sum_x p(X = x) = 1$$

- And all probabilities have to be greater or equal to 0
- Probabilities of disjunctions are sums over part of the space.
  E.g., the probability that a die is bigger than 3:

$$p(X > 3) = p(X = 4) + p(X = 5) + p(X = 6)$$

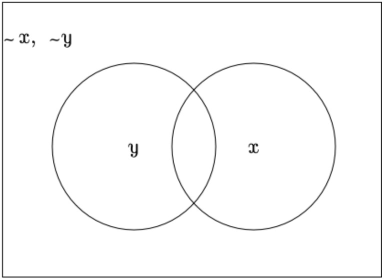
# Random variable, atom, and event

- The figure below is helpful to understand these concepts well



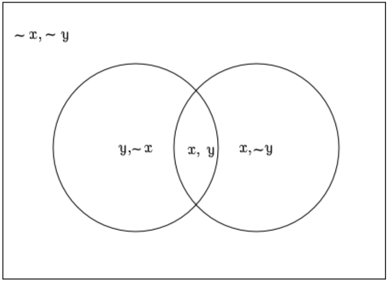

- An *atom* is a point in the box. All atoms together form the *sample space*
- An *event* is a subset of atoms. Two events in the picture are $x$ and $\sim x$
- The probability of an event is the sum of the probabilities of its atoms

- In practice, we need to consider many variables at the same time
- An event would then combine atoms from multiple variables




- The **joint distribution** is a distribution over the configuration of all the random variables in the ensemble
  - For the figure, the function $p(X, Y)$ gives the probability of all possible combinations of $X$ and $Y$
  - Notice that $X \in \{x, \sim x\}$ and $Y \in \{y, \sim y\}$
  - Therefore $X, Y \in \{(x, y), (x, \sim y), (\sim x, y), (\sim x, \sim y)\}$

# Joint distribution



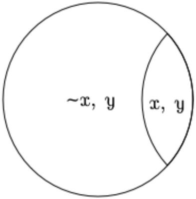

- Some useful properties:
  - Union: $p(X \cup Y) = p(X) + p(Y) - p(X, Y)$
  - **Marginalization**: $p(X) = \sum_Y p(X, Y)$
    This property is referred to as the **sum rule of probability!**

# Joint distribution

- We can make a joint distribution of *two consecutive events*
- With the cards example:
  - $X$=first draw $\in$ {Red/Blue, Red/Red, Blue/Blue}
  - $Y$=second draw $\in$ {Red/Blue, Red/Red, Blue/Blue}
- $X, Y \in$ {(Red/Blue, Red/Blue), (Red/Blue, Red/Red), (Red/Blue, Blue/Blue), (Red/Blue, Red/Red), (Red/Red, Red/Red), (Red/Red, Blue/Blue), (Red/Blue, Blue/Blue), (Blue/Blue, Blue/Blue)}
- How to calculate the join distribution, $p(X, Y)$?

# Conditional probability

- What about when we have observed one event, but want to know the probability of another one?
- The **conditional probability** of $X$ given $Y$ is the probability of event $X$ when event $Y$ is known




- So, we only concentrate on the subset of events where the specific value of $Y$ occurs
- In the above figure, we focus on when $Y = y$

$$p(X|Y = y) = \frac{p(X, Y = y)}{p(Y = y)}$$

# Conditional probability

- We can now solve the card problem
- Let's have two events, $X_1$ for observed side of the card, and $X_2$ for the side we want to guess
- We need to calculate $p(X_2 = \text{Red}|X_1 = \text{Red})$

$p(X_1 = \text{Red}) = \frac{1}{2}$
$p(X_1 = \text{Red}, X_2 = \text{Red}) = \frac{1}{3}$

therefore

$$p(X_2 = \text{Red}|X_1 = \text{Red}) = \frac{p(X_2 = \text{Red}, X_1 = \text{Red})}{p(X_1 = \text{Red})} = \frac{1/3}{1/2} = \frac{2}{3}$$

# The chain rule (or product rule)

- Consider the conditional probability rule

$$p(X|Y) = \frac{p(X,Y)}{p(Y)}$$

- It allows us to derive the chain rule, which defines the joint distribution as a product of conditionals:

$$\begin{aligned} p(X,Y) &= p(X,Y)\,\frac{p(Y)}{p(Y)} \\ &= p(X|Y)\,p(Y) \end{aligned}$$

- In general, for any set of variables

$$p(X_1, X_2, ..., X_N) = \prod_{n=1}^{N} p(X_n|X_1, X_2, ..., X_{n-1})$$

- For example:

$$p(X,Y,Z) = p(X)\,p(Y|X)\,p(Z|Y,X)$$

# Bayes theorem

- Using the chain rule, we can trivially say:

$$p(X|Y)\,p(Y) = p(Y|X)\,p(X)$$

which means that [***Bayes theorem***]:

$$p(X|Y) = \frac{p(Y|X)\,p(X)}{p(Y)}$$

- The Bayes theorem is an important foundation for Bayesian statistics, and particularly for Probabilistic Graphical Models!

# Playtime!

- Open "1.Probability_Review.ipynb" in Jupyter
- Do Part 1, estimated duration 20 min

# Independence

- Random variables are *independent* if knowing about $X$ tells us nothing about $Y$

$$p(Y|X) = p(Y)$$

- This means that their joint distribution is

$$p(X, Y) = p(X)\,p(Y)$$

- Why?
- A couple of examples:
  - Two persons, A, and B, start their trip in different parts of town. The transport mode for A is $X$ and for B, it is $Y$. Are these two choices independent?
  - It's a rainy day. Two accidents happen on different roads of the city, far from each other. Are these two, independent events?

# Independence

- Are these independent?
  - the speeds in adjacent road sections
  - the flow of pedestrians and the flow of cars in the same road
  - whether it is raining and the number of taxi cabs
  - whether it is raining and the amount of time it takes me to hail a cab
  - the departure time and arrival time of a trip

# Independence

- Example: two coins, $C_1, C_2$ with $p(H|C_1) = 0.5, p(H|C_2) = 0.3$

  1. Suppose that I randomly choose a number $Z \in \{1, 2\}$, and take coin $C_Z$
  2. I flip it twice, with results $(X_1, X_2)$

  Are $X_1$ and $X_2$ independent? What about if I know $Z$?

# Conditional independence

- $X$ and $Y$ are *conditionally independent* given $Z$

$$p(X|Y,Z) = p(X|Z)$$

- So, we can say that

$$X \perp\!\!\!\perp Y|Z \implies p(X,Y|Z) = p(X|Z)\,p(Y|Z)$$

# Conditional independence

- $X$ and $Y$ are *conditionally independent* given $Z$

$$p(X|Y,Z) = p(X|Z)$$

- So, we can say that

$$X \perp\!\!\!\perp Y|Z \implies p(X,Y|Z) = p(X|Z)\,p(Y|Z)$$

- *If we know Z, then knowing about Y tells us nothing about X*

# Playtime!

- Open "1.Probability_Review.ipynb" in Jupyter
- Do Part 2, estimated duration 30 min

# Expectation

- The *expected value* of a random variable is the probability-weighted average of all possible values
- In other words, it is the *mean* of the distribution of this random variable

$$\mathbb{E}[X] = \sum_x x \, p(X = x)$$

- More generically (remember the $f(x)$ can be itself a random variable)

$$\mathbb{E}[f(X)] = \sum_x f(x) \, p(X = x)$$

# Playtime!

- Open "1.Probability_Review.ipynb" in Jupyter
- Do Part 3, estimated duration 10 min

# Continuous random variables

- We've only used discrete random variables so far (e.g., dice, cards)
- Random variables can be continuous
- We need a density function $p(x)$, which integrates to one.

$$\int_{-\infty}^{\infty} p(x)\, dx = 1$$

- Probabilities are integrals over $p(x)$
- An *event* is thus defined by an interval of possible values of the random variable

$$P(a \leq X \leq b) = \int_{a}^{b} p(x)\, dx$$

- Notice that we use $X$, $x$, $P$, and $p$!...

# Some distributions - Gaussian

- By far, the most common one...
- Two parameters:
  - Mean, $\mu$
  - Standard deviation, $\sigma$ (or, variance, $\sigma^2$)
- $p(x)$ is defined as

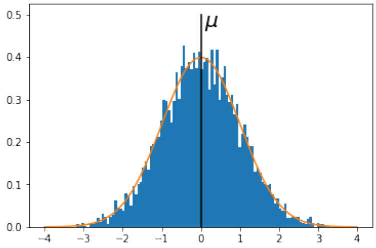
$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

- Often represented as:

$$p(x) \sim \mathcal{N}(\mu, \sigma^2)$$

# Some distributions - Gaussian

- Support is $]-\infty, \infty[$
- Symmetrical



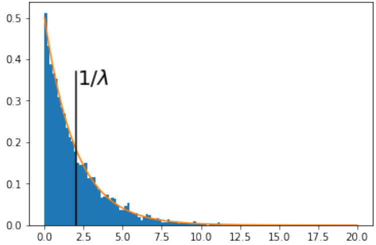

- The Central limit theorem (CLT) establishes that *the distribution of the sampling means approaches a normal distribution as the sample size gets larger, no matter what the shape of the population distribution.*

# Some distributions - Exponential

- Exponential distribution, with *rate* $\lambda$

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

- Support is $[0, \infty[$

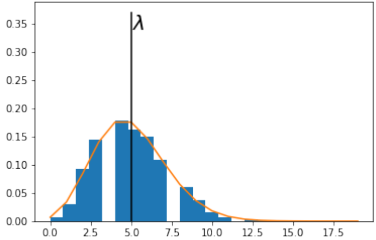


## Some distributions - Poisson

- Poisson distribution, with *rate* $\lambda$

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

- for $k = 0, 1, 2...$
- Pretty common in transportation (e.g. arrival rates)




1

[1] In fact, this distribution relates to a discrete random variable, so we include it to emphasize that not only continuous variables can be parameterized as a probability distribution.

# Independent and identically distributed random variables (iid)

- Independent
- Identically distributed

# Independent and identically distributed random variables (iid)

- Independent
- Identically distributed

If we repeatedly flip the same coin $N$ times and record the outcome, then $X_1, ..., X_N$ are **iid**

- The iid assumption can be extremely useful in data analysis

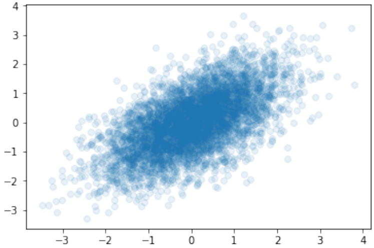
# Multivariate distributions

- So far, we've been working with single variable distributions
- Multivariate means it's the same as above, but with more variables at the same time!
- In practice, joint distribution of variables that share a common structure
- In some cases (e.g. Poisson), it is not a trivial problem
- In others (e.g. Gaussian), it is well studied, and extensively applied

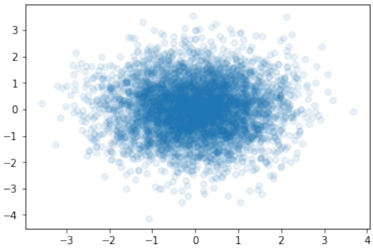
$$p(\mathbf{x}) = \frac{1}{\sqrt{2\pi}|\boldsymbol{\Sigma}|}\, e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

# Multivariate distributions

- Bivariate Gaussian




$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix}$$




$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

# Playtime!

- Open "2. Probability_Review.ipynb"
- Do part 1. Est. time is 15 min

# A note on notation

- So far we have been using a rather standard statistics notation
  - $X$ is a random variable and $x$ is atom/event
  - We write e.g. $p(X = x)$
- In the machine learning literature, this notation is typically simplified
  - Lowercase letters, such as $x$, represent random variables
  - We simply write $p(x)$. Everything else should be clear from the context!
- This allows us to have
  - Bold letters denote vectors (e.g. $\mathbf{x}$, where the $i^{th}$ element is referred as $x_i$)
  - Matrices are represented by bold uppercase letters such as $\mathbf{X}$
  - Roman letters, such as $N$, denote constants
- This is the notation that we will adopt from now on!

# The likelihood function

- Imagine you have the data. For example:
  - $N$ readings of traffic counts at a certain time, each one called $x_i$, $i = 1...N$
- You assume it follows some parametric distribution (e.g. Gaussian)
- How do you determine its parameters, $\Theta$?

# The likelihood function

- Imagine you have the data. For example:
  - $N$ readings of traffic counts at a certain time, each one called $x_i$, $i = 1...N$
- You assume it follows some parametric distribution (e.g. Gaussian)
- How do you determine its parameters, $\Theta$?
- The likelihood function, $L(\Theta)$, should be:

$$L(\Theta) = \prod_{i}^{N} p(x_i|\Theta)$$

# The likelihood function

- Imagine you have the data. For example:
  - $N$ readings of traffic counts at a certain time, each one called $x_i$, $i = 1 \dots N$
- You assume it follows some parametric distribution (e.g. Gaussian)
- How do you determine its parameters, $\Theta$?
- The likelihood function, $L(\Theta)$, should be:

$$L(\Theta) = \prod_{i}^{N} p(x_i|\Theta)$$

- Notice that this is the joint distribution of all **independent** data points!

## The likelihood function

- Imagine you have the data. For example:
  - $N$ readings of traffic counts at a certain time, each one called $x_i$, $i = 1...N$
- You assume it follows some parametric distribution (e.g. Gaussian)
- How do you determine its parameters, $\Theta$?
- The likelihood function, $L(\Theta)$, should be:

$$L(\Theta) = \prod_{i}^{N} p(x_i|\Theta)$$

- Notice that this is the joint distribution of all **independent** data points!
- In the case of the Gaussian, we should have $\Theta = \{\mu, \sigma\}$
- The likelihood function, $L(\Theta)$, would be

$$L(\Theta) = \prod_{i}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)}$$

## The likelihood function

- In the case of the Gaussian, we should have $\Theta = \{\mu, \sigma\}$
- The likelihood function, $L(\Theta)$, would be

$$L(\Theta) = \prod_{i}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)}$$

# The likelihood function

- In the case of the Gaussian, we should have $\Theta = \{\mu, \sigma\}$
- The likelihood function, $L(\Theta)$, would be

$$L(\Theta) = \prod_{i}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)}$$

- If you actually *had* the true parameters, the likelihood function would have the maximum value, right?
- So, this becomes an optimization problem:
    - Find the values of $\Theta$ that maximize the function $L(\Theta)$

# The log-likelihood function

- For practical reasons, we apply a logarithmic transformation to the likelihood function
  - Less prone to numeric error (numerical stability)
  - Computationally faster

# The log-likelihood function

- For practical reasons, we apply a logarithmic transformation to the likelihood function
    - Less prone to numeric error (numerical stability)
    - Computationally faster
- In the case of the Gaussian distribution, the log likelihood becomes:

$$-\frac{N}{2}(\log(2\pi) + \log(\sigma^2)) - \frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2$$

# Maximum likelihood estimate (MLE)

- The maximum likelihood estimate is the value of the parameter that maximizes the log likelihood (equivalently, the likelihood)
- In the case of the Gaussian, the MLE corresponds to:

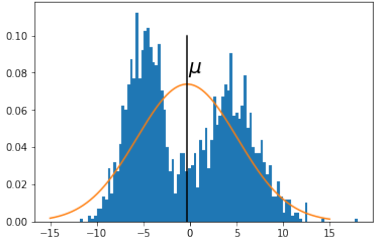
$$\hat{\mu} = \frac{\sum_{i=1}^{N} x_i}{N}, \quad \text{i.e. the } \textit{sample mean}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} (x_i - \hat{\mu})^2}{N}, \quad \text{i.e. the } \textit{sample variance}$$

# Maximum likelihood estimate (MLE)

DISCLAIMER:

- The fact that you get a MLE doesn't mean you found a good model!




- You need to know your data...

# Playtime!

- Open “2. Probability_Review.ipynb”
- Do part 2. Est. time is 30 min