Probability and Statistics review (part 2)

Francisco Pereira

Filipe Rodrigues

DTU Management Engineering
Department of Management Engineering

Outline

- Continuous random variables

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

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, μ
  - Standard deviation, σ (or, variance, σ^2)
- p(x) is defined as

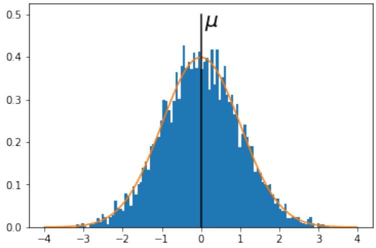
$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

- Often represented as:

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

Some distributions - Gaussian

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



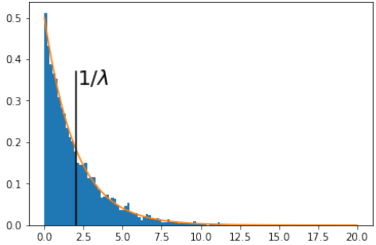
- The Central limit theorem (CLT) establishes that *the distribution of the sampling means approaches a normal distribution as the sample size gets larger, no matter what the shape of the population distribution.*

Some distributions - Exponential

- Exponential distribution, with *rate* λ

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

- Support is $[0, \infty[$

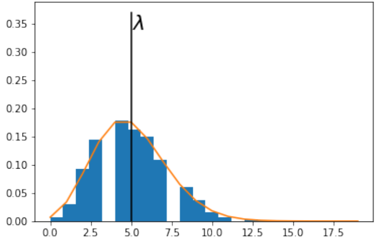


Some distributions - Poisson

- Poisson distribution, with *rate* λ

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

- 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 useful in data analysis.

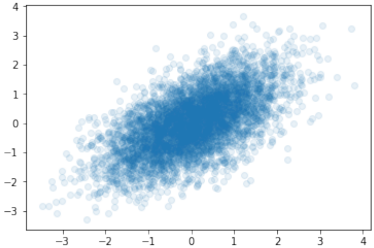
Multivariate distributions

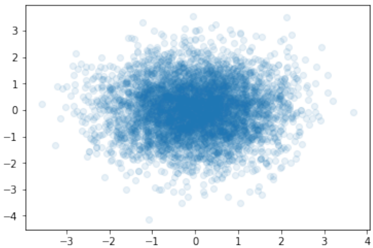
- So far, we've been working with single variable distributions
- Multivariate means it's the same as above, but with more variables at the same time!
- In practice, joint distribution of variables that share a common structure
- In some cases (e.g. Poisson), it is not a trivial problem
- In others (e.g.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

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

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

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, Θ?

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, Θ?
- The likelihood function should be:

$$\prod_i^N p(x_i|\Theta)$$

The likelihood function

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

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

- 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, Θ?
- The likelihood function should be:

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

- 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 would be

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

The likelihood function

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

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

The likelihood function

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

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

- 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 Θ that maximize the function L

The log-likelihood function

- For practical reasons, we apply a logarithmic transformation to the Likelihood function
  - Less prone to numeric error
  - Computationally faster

The log-likelihood function

- For practical reasons, we apply a logarithmic transformation to the Likelihood function
    - Less prone to numeric error
    - Computationally faster
- In the case of the Gaussian distribution, it 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:

$$\hat{\mu} = \frac{\sum_{i=1}^{N} x_i}{N}$$

, i.e. the *sample mean*

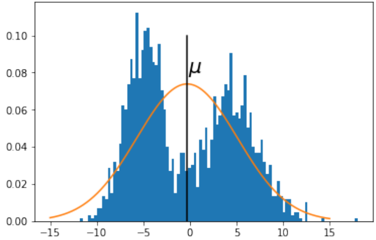
$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} (x_i - \hat{\mu})^2}{N}$$

, i.e. the *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