DTU

Model-based Machine Learning

Probability and Statistics review

Outline

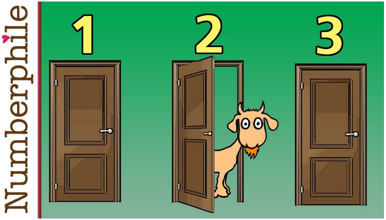
(Based on David MacKay, David Blei,
https://www.cs.princeton.edu/courses/archive/spring12/cos424/pdf/lecture02.pdf)

Teaser

Consider the "Monty Hall problem"

- There are three doors:
    - One has a car (picture)
    - Two have a goat (picture)

    1. Participant chooses one door
    2. Host (*Monty Hall*) opens another door
    3. Host's opened door is always a goat
- Should the participant change his/her choice?



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

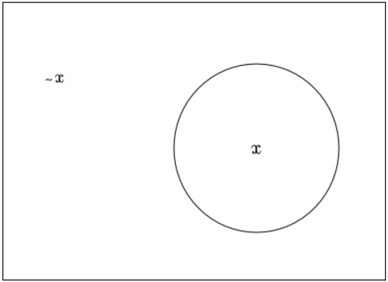
$$\sum_{x} p(X = x) = 1$$

- 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:

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

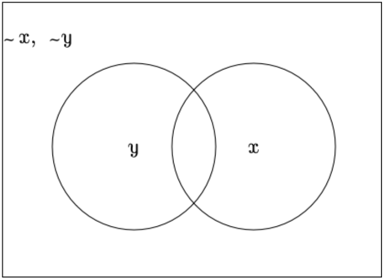
Random variable, atom, and event

- The figure below is helpful to understand these concepts well



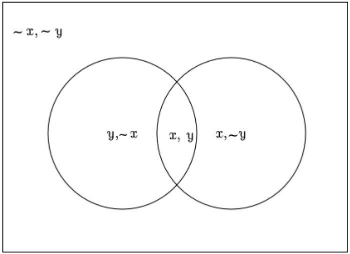
- An *atom* is a point in the box. All atoms together form the *sample space*
- An *event* is a subset of atoms. Two events in the picture are x and $\sim x$
- The probability of an event is the sum of the probabilities of its atoms

- In practice, we often combine many variables/events at the same time



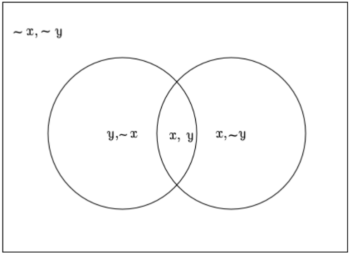
- The **joint distribution** is a distribution over the configuration of all the random variables in the ensemble
  - For the figure, the function $p(X, Y)$ gives the probability of all possible combinations of X and Y
  - Notice that $X \in \{x, \sim x\}$ and $Y \in \{y, \sim y\}$
  - Therefore $X,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

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!**

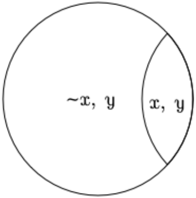
Joint distribution



- 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!**
  - Notice that you can have more variables on the conditioning side:
    $p(X|Z, \ldots) = \sum_Y p(X, Y|Z, \ldots)$
  - Or marginalize over more variables: $p(X) = \sum_Y \sum_Z \ldots p(X, Y, Z, \ldots)$

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 $Y = y$

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

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:

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

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:

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

- Notice that you can also have more variables on the conditioning side:

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

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:

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

- Notice that you can also have more variables on the conditioning side:

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

- 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***]:

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

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***]:

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

- Notice that you can also have more variables on the conditioning side:

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

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

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

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

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

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

- A few examples:
  - Two lottery numbers that two (unacquainted) people chose. Are these two numbers independent?
  - 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. Are these two, independent events?
  - The speeds in adjacent road sections x

Independence

- Example: two coins, C_1, C_2 with $p(H|C_1) = 0.6, p(H|C_2) = 0.2$
  1. Suppose that I randomly choose a number $Z \in \{1, 2\}$ (with equal probability), 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*

Teaser - Monty Hall

- We can now solve the Monty Hall problem

  X = true location of the car
  Y = door that host opened
  Z = choice of participant

  We want to know $p(X|Y,Z)$

Teaser - Monty Hall

- We can now solve the Monty Hall problem

  X = true location of the car
  Y = door that host opened
  Z = choice of participant

  We want to know $p(X|Y,Z)$, from Bayes rule we have:

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

Teaser - Monty Hall

- We can now solve the Monty Hall problem

  $$\begin{aligned} \text{X} &= \text{true location of the car} \\ \text{Y} &= \text{door that host opened} \\ \text{Z} &= \text{choice of participant} \end{aligned}$$

  We want to know $p(X|Y,Z)$, from Bayes rule we have:

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

Teaser - Monty Hall

- We can now solve the Monty Hall problem

  X = true location of the car
  Y = door that host opened
  Z = choice of participant

  We want to know $p(X|Y,Z)$, from Bayes rule we have:

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

  we can also say:

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

Teaser - Monty Hall

- We can now solve the Monty Hall problem

  X = true location of the car
  Y = door that host opened
  Z = choice of participant

  We want to know $p(X|Y,Z)$, from Bayes rule we have:

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

  we can also say:

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

  We know that, $p(X)$, the prior probability of the location is:

$$p(X) = \frac{1}{3}$$

  But what about $p(Y|X,Z)$?

Teaser - Monty Hall

But what about $p(Y|X, Z)$?
It's the *likelihood* of host choosing location Y, *given that* he knows X and Z

Teaser - Monty Hall

But what about $p(Y|X, Z)$?
It's the *likelihood* of host choosing location Y, *given that* he knows X and Z

	$Y=1$	$Y=2$	$Y=3$
$X=1, Z=1$	0	0.5	0.5
$X=1, Z=2$	0	0	1
$X=1, Z=3$	0	1	0
$X=2, Z=1$	0	0	1
$X=2, Z=2$	0.5	0	0.5
$X=2, Z=3$	1	0	0
$X=3, Z=1$	0	1	0
$X=3, Z=2$	1	0	0
$X=3, Z=3$	0.5	0.5	0

Table: $p(Y|X, Z)$

Teaser - Monty Hall

But what about $p(Y|X,Z)$?
It's the *likelihood* of host choosing location Y, *given that* he knows X and Z

	$Y = 1$	$Y = 2$	$Y = 3$
$X = 1, Z = 1$	0	0.5	0.5
$X = 1, Z = 2$	0	0	1
$X = 1, Z = 3$	0	1	0
$X = 2, Z = 1$	0	0	1
$X = 2, Z = 2$	0.5	0	0.5
$X = 2, Z = 3$	1	0	0
$X = 3, Z = 1$	0	1	0
$X = 3, Z = 2$	1	0	0
$X = 3, Z = 3$	0.5	0.5	0

Table: $p(Y|X,Z)$

Ok, let's try a scenario. Let's assume that the participant chose door 3 and the host opened door 2. Then our table becomes:

	$Y = 2$
$X = 1, Z = 3$	1
$X = 2, Z = 3$	0
$X = 3, Z = 3$	0.5

Table: $p(Y = 2|X, Z = 3)$

Teaser - Monty Hall

In this scenario, we want to calculate

$$p(X|Y=2,Z=3) \propto p(Y=2|X,Z=3)P(X)$$

and we have

	$Y=2$
$X=1, Z=3$	1
$X=2, Z=3$	0
$X=3, Z=3$	0.5

Table: $p(Y=2|X,Z=3)$

$$P(X)=\frac{1}{3}$$

- Let's just calculate for the two possible cases (X is either in door 1 or 3!):

$$p(Y=2|X=1,Z=3)\times\frac{1}{3}=\frac{1}{3}$$

$$p(Y=2|X=3,Z=3)\times\frac{1}{3}=\frac{1}{2}*\frac{1}{3}=\frac{1}{6}$$

- So we can get our normalizing quantity: $\sum_X p(Y|X,Z)p(X)=\frac{1}{3}+\frac{1}{6}=\frac{1}{2}$

Teaser - Monty Hall

Remember: to calculate the distribution of X (i.e. "where the car probably is"), we need to calculate

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

- If we follow our example, we get

$$p(X|Y=2,Z=3) = \frac{p(Y=2|X,Z=3) \times \frac{1}{3}}{\frac{1}{2}}$$

Teaser - Monty Hall

Remember: to calculate the distribution of X (i.e. "where the car probably is"), we need to calculate

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

- If we follow our example, we get

$$p(X|Y=2,Z=3) = \frac{p(Y=2|X,Z=3) \times \frac{1}{3}}{\frac{1}{2}}$$

- using the calculations from the previous slide, we have

$$p(X=1|Y=2,Z=3) = \frac{\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}$$

$$p(X=3|Y=2,Z=3) = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$$

- By this reasoning, we **always** have $\frac{2}{3}$ chances when we change doors, and keep $\frac{1}{3}$ if we keep it!

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 “01 - Probability theory review.ipynb” in Jupyter
- Do Part 1, estimated duration 20 min
- Do Part 2, estimated duration 30 min
- 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, μ
  - Standard deviation, σ (or, variance, σ^2)
- $p(x)$ is defined as

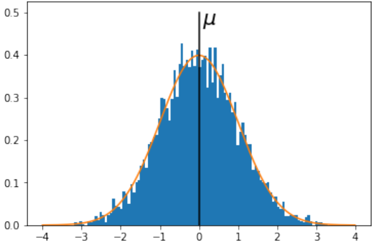
$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

- Often represented as:

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

Some distributions - Gaussian

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



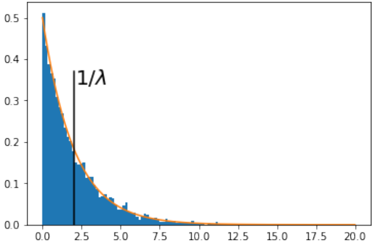
- The Central limit theorem (CLT) establishes that *the distribution of the sampling means approaches a normal distribution as the sample size gets larger, no matter what the shape of the population distribution.*

Some distributions - Exponential

- Exponential distribution, with *rate* λ

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

- Support is $[0, \infty[$

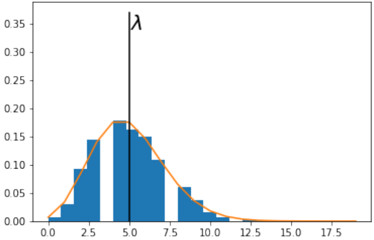


Some distributions - Poisson

- 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 $X_1, \ldots, X_N$ are **iid**

- The iid assumption can be extremely useful in data analysis

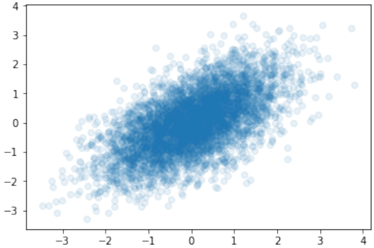
Multivariate distributions

- 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

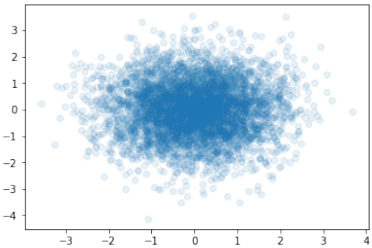
$$p(\mathbf{x}) = \frac{1}{\sqrt{2\pi}|\mathbf{\Sigma}|} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

Multivariate distributions

- Bivariate Gaussian



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



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

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, 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, $L(\Theta)$, 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, $i = 1...N$
- You assume it follows some parametric distribution (e.g. Gaussian)
- How do you determine its parameters, Θ?
- 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, Θ?
- 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 Θ 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:

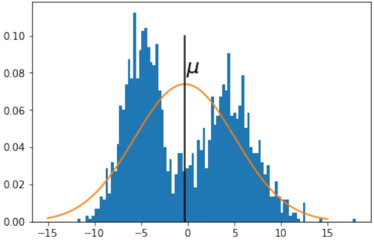
$$\hat{\mu} = \frac{\sum_{i=1}^{N} x_i}{N}, \quad \text{i.e. the } \textit{sample mean}$$

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

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 “01 - Probability theory review.ipynb”
- Do part 4. Est. time is 15 min
- Do part 5. Est. time is 30 min