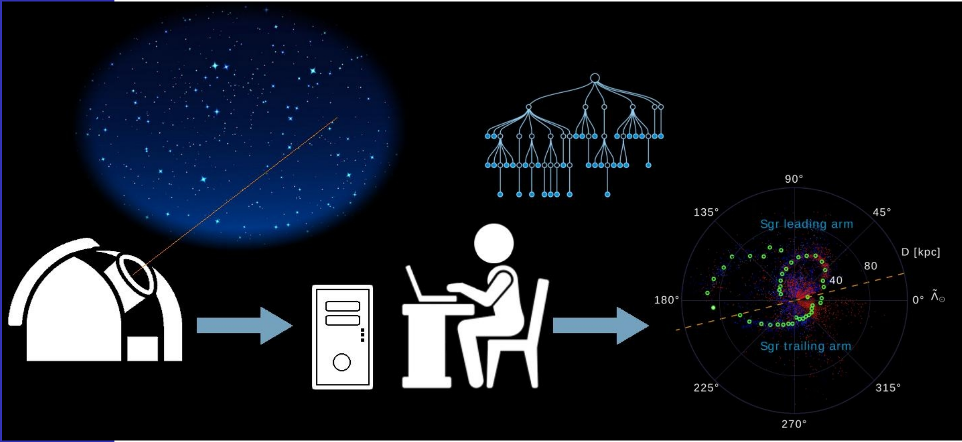
Machine Learning (Semester 1 2024)

Introduction & Statistical Inference

Nina Hernitschek
Centro de Astronomía CITEVA
Universidad de Antofagasta

April 23, 2024

Motivation



90°
45°
135°
180°
0°
225°
270°
315°
Sgr leading arm
Sgr trailing arm
D [kpc]
80
40
$\tilde{\Lambda}_\odot$

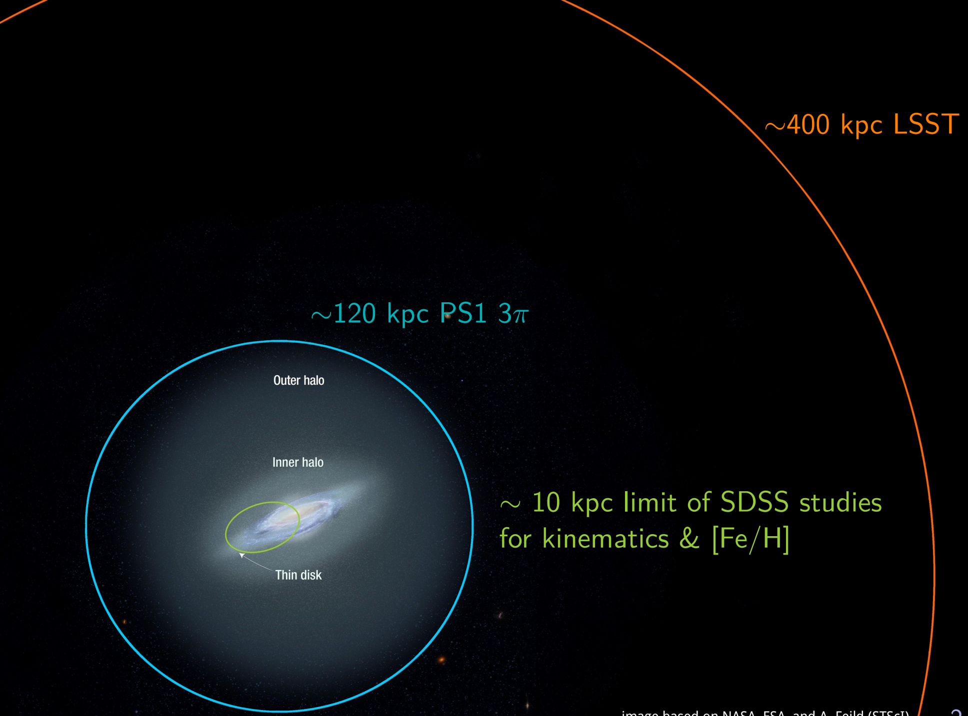
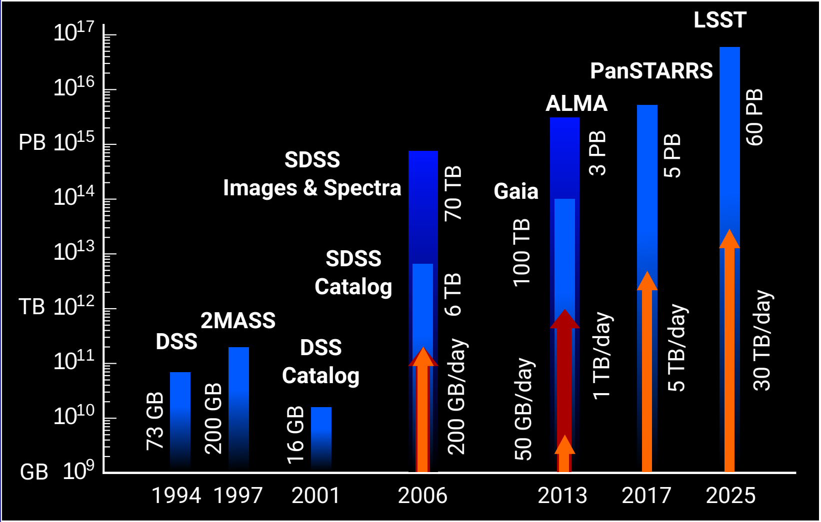


image based on NASA, ESA, and A. Feild (STScI)

Challenges in Data Handling

increasing data volume in astronomical surveys



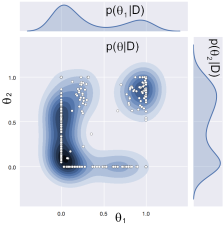
What you will learn in this class

this course will prepare you for "doing science" with current and upcoming large astronomical surveys:

What you will learn in this class

this course will prepare you for “doing science” with current and upcoming large astronomical surveys:

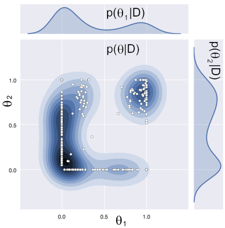
statistical methods



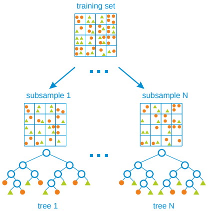
What you will learn in this class

this course will prepare you for "doing science" with current and upcoming large astronomical surveys:

statistical methods



machine learning



What do expect

1. Many statistical learning methods are relevant in a wide range of (academic and non-academic) **disciplines**. Methods presented here are not only applicable to astronomy.

What do expect

1. Many statistical learning methods are relevant in a wide range of (academic and non-academic) **disciplines**. Methods presented here are not only applicable to astronomy.

2. Statistical learning should **not be seen as "black box"** algorithms. Despite in many cases, algorithms from software/ programming libraries will be used, it is essential to understand what these algorithms do. Without understanding "what happens inside the box", it is not only impossible to select the best "box", it is also impossible to prevent common misunderstandings in interpreting data.

What do expect

1. Many statistical learning methods are relevant in a wide range of (academic and non-academic) **disciplines**. Methods presented here are not only applicable to astronomy.

2. Statistical learning should **not be seen as "black box"** algorithms. Despite in many cases, algorithms from software/ programming libraries will be used, it is essential to understand what these algorithms do. Without understanding "what happens inside the box", it is not only impossible to select the best "box", it is also impossible to prevent common misunderstandings in interpreting data.

3. This course has the objective to prepare you to apply statistical learning methods to **real-world problems**. In order to facilitate this, we have weekly practice sessions where you will write code to analyze data.

Course Logistics

course content:

- lecture: Tuesday 12 - 14h
- practice: Thursday 12 - 14h
- preparation of a paper presentation (of your choice)

grading:

- participation: 10 %
- practice: 40 %
- paper presentation: 50 %

contact and course material:

- e-mail: `nina.hernitschek@uantof.cl`
- github: `https://github.com/ninahernitschek/machine_learning_aprendizaje_de_maquinas_2024_1`

Course Logistics

April 23 Lecture 1: Introduction & Statistical Inference, **April 25** Practice 1

April 30 Lecture 2: Statistical Learning, **May 2** Practice 2

May 7 Lecture 3: Regression, **May 9** Practice 3

May 14 Lecture 4: Linear Model Selection and Regularization, **May 16** Practice 4

May 28 Lecture 5: Non-Linearity, **May 30** Practice 5

June 4 Lecture 6: Resampling Methods, **June 6** Practice 6

June 11 Lecture 7: Classification, **June 13** Practice 7

June 25 Lecture 8: Support Vector Machines, **June 27** Practice 8

July 2 Lecture 9: Tree-Based Methods, **July 4** Practice 9

July 9 Lecture 10: Dim. Reduction & Unsupervised Learning, **July 11** Paper Presentation

Rules for Coding and Presentations

coding: If you have a question when something doesn't work, summarize what you tried - often this will even lead to the solution.

paper presentation:

- LaTeX
- figures: properly cite and describe figures from the paper
- data of project shown in the paper: data description (incl. citation)
- implementation: medium-level implementation description with libraries/ software frameworks (incl. citation), figures related to the implementation (such as flow charts) from the paper
- discussion and summary: reflect the approach (strengths, weaknesses, limitations), lessions learned
- bibliography: bibtex/ref mechanism, ADS/Bibtex information

Textbooks

Basic Bibliography:

An Introduction to Statistical Learning: with Applications in Python.
James Witten, Hastie, Tibshirani; Springer.
available for free as PDF: `https://www.statlearning.com/`

Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow.
A. Géron; O'Reilly.

Supplementary bibliography:

Statistics, Data Mining and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data
Ž. Ivezić, A. J. Connolly, J. T. VanderPlas, A. Gray

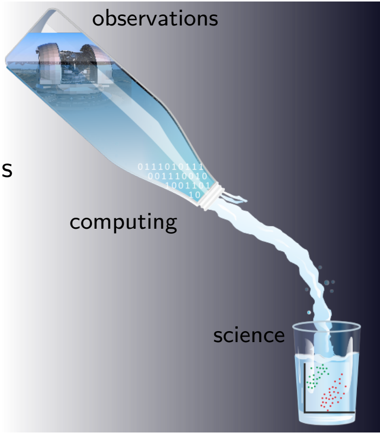
Machine Learning: An Algorithmic Perspective
Second Edition, S. Marsland; (Chapman & Hall/CRC Machine Learning & Pattern Recognition)

Challenges in Data Handling

astronomy is largely determined by computational capacity

$\Rightarrow$ telescopes & instruments as front-ends for data processing systems

$\Rightarrow$ **challenge and chance:** understanding complex phenomena requires complex data

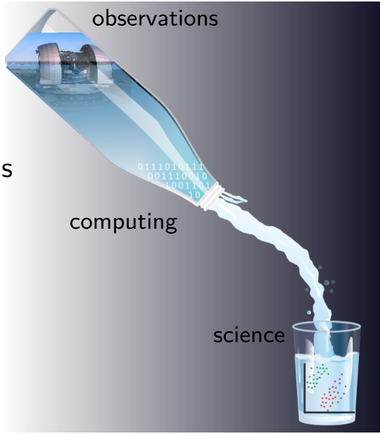


Challenges in Data Handling

astronomy is largely determined by computational capacity

$\Rightarrow$ telescopes & instruments as front-ends for data processing systems

$\Rightarrow$ **challenge and chance:** understanding complex phenomena requires complex data



Big Data is transforming how and which discoveries are made

Big Data

Laney et al. 2001: data growth challenge is **three-dimensional**

Big Data

Laney et al. 2001: data growth challenge is **three-dimensional**

Big Data is data with at least one big dimension:

- volume
- velocity: bandwidth, response speed
- variety: number and size of individual assets

Big Data

Laney et al. 2001: data growth challenge is **three-dimensional**

Big Data is data with at least one big dimension:

- volume
- velocity: bandwidth, response speed
- variety: number and size of individual assets

shifting use cases:
As data becomes big data, finding the *right* data has become more important.

Big Data

Laney et al. 2001: data growth challenge is **three-dimensional**

Big Data is data with at least one big dimension:

- volume
- velocity: bandwidth, response speed
- variety: number and size of individual assets

shifting use cases:
As data becomes big data, finding the *right* data has become more important.

$\Rightarrow$ powerful astrostatistical & machine-learning tools are needed to derive scientific insights

Big Data

shifting use cases:

As data become more plentiful, finding the *right* data has become more important.

$\Rightarrow$ powerful astrostatistical & machine-learning tools are needed to derive scientific insights

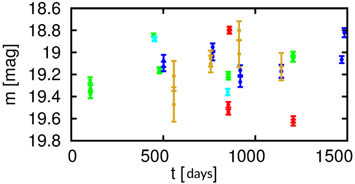
Individual measurements giving way to **statistics, clustering, patterns** in the data.

Data processing needs to be **highly automatized**.
Analysis growing more exploratory rather than pre-defined/scripted.

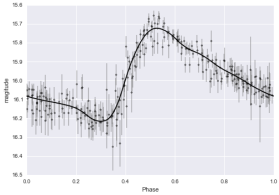
Big Data

one **example** for finding the *right* data :

Pan-STARRS1 3π survey with about 10^9 light curves like that:



goal: finding RR Lyrae* stars whose light curves look like that (if better sampled):



*less than 1 % of the light curves are expected to be from that type

Statistical Data Analysis

Data-driven methods like statistical methods can reliable **quantify information** embedded in scientific data **without the biases of physical models**.

Requirements:

- find the right method(s): modern statistics is vast in its scope and methodology
- scientific inferences should not depend on arbitrary choices in methodology and variable scale
- correct interpretation of the meaning of a statistical result w.r.t. the scientific goal: (astro-)statistics and machine learning are only tools!

(Astro-)Statistics

a lot is possible:

galaxy clustering	spatial point processes, clustering
galaxy morphology	regression, mixture models
weak lensing morphology	geostatistics, density estimation
strong lensing morphology	shape statistics
faint source detection	false discovery rate
variable source preclassification	structure functions + classifier

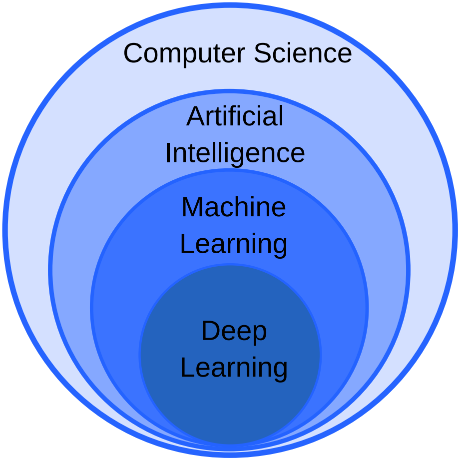
(Astro-)Statistics

a lot is possible:

galaxy clustering	$\Longleftrightarrow$	spatial point processes, clustering
galaxy morphology		regression, mixture models
weak lensing morphology		geostatistics, density estimation
strong lensing morphology		shape statistics
faint source detection		false discovery rate
variable source preclassification		structure functions + classifier

$\Rightarrow$ **fitting models**

Machine Learning - Terminology



Machine Learning

... is the sub-field of computer science that gives computers the ability to learn without being explicitly programmed
(Arthur Samuel, 1959)

Machine Learning

... is the sub-field of computer science that gives computers the ability to learn without being explicitly programmed
(Arthur Samuel, 1959)

$\Rightarrow$ allows to **uncover hidden correlation patterns** through iterative learning by sample data

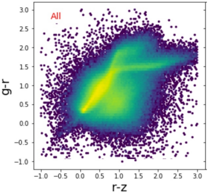
Machine Learning

... is the sub-field of computer science that gives computers the ability to learn without being explicitly programmed
(Arthur Samuel, 1959)

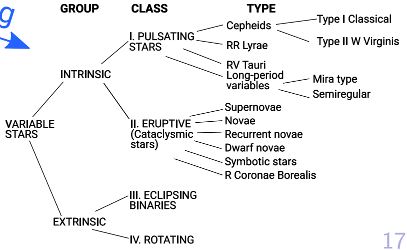
$\Rightarrow$ allows to **uncover hidden correlation patterns** through iterative learning by sample data

$\Rightarrow$ **in astronomy**:

parameter space of measurements



parameter space of astrophysical objects



Machine Learning

... is the sub-field of computer science that gives computers the ability to learn without being explicitly programmed
(Arthur Samuel, 1959)

$\Rightarrow$ allows to **uncover hidden correlation patterns** through iterative learning by sample data

$\Rightarrow$ **in astronomy**: allows **to model a survey**:

- describing data quality $\rightarrow$ outlier
- describing light curve characteristics $\rightarrow$ "features"
- classifying sources $\rightarrow$ catalogs
- finding substructure $\rightarrow$ clumps, overdensities, ...

Functions of Machine Learning Systems

Descriptive the system uses the data to explain data properties; tools: simple statistical tools such as averages, percentages

Functions of Machine Learning Systems

Descriptive	the system uses the data to explain data properties; tools: simple statistical tools such as averages, percentages
Predictive	focuses on predicting and understanding future behavior

Functions of Machine Learning Systems

Descriptive — the system uses the data to explain data properties; tools: simple statistical tools such as averages, percentages

Predictive — focuses on predicting and understanding future behavior

Prescriptive — the system uses data to make suggestions about actions to take based on the insights gained

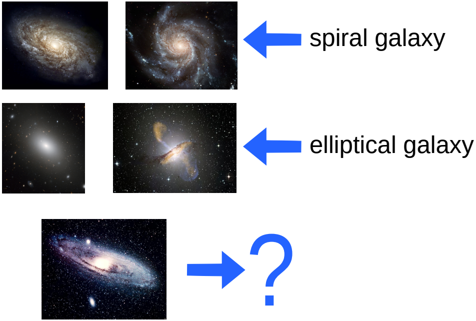
Types of Machine Learning Algorithms

there are different types of machine learning algorithms that differ mostly by **how they use data**

Supervised Machine Learning

labeled data (training data, training set) enable the supervised machine learning algorithm to understand the connection between **features** and **labels**

new observations (target data) are assigned to a group or class



applications: classification problems, regression problems

Supervised Machine Learning

The objective of a supervised learning model is to predict the correct label for newly presented input data.

Supervised Machine Learning

The objective of a supervised learning model is to predict the correct label for newly presented input data.

When training a supervised learning algorithm, the **training set** will consist of inputs paired with the correct outputs. Inputs in the training set should represent the **target set** which we have to classify: composition of the data, data quality.

Supervised Machine Learning

The objective of a supervised learning model is to predict the correct label for newly presented input data.

When training a supervised learning algorithm, the **training set** will consist of inputs paired with the correct outputs. Inputs in the training set should represent the **target set** which we have to classify: composition of the data, data quality.

During **training**, the algorithm will search for patterns in the data that correlate with the desired outputs.

Supervised Machine Learning

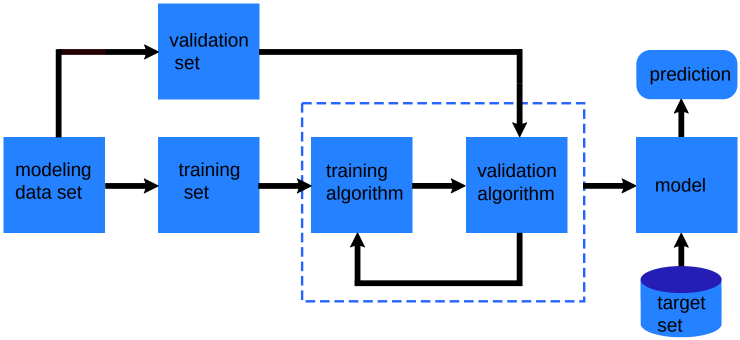
The objective of a supervised learning model is to predict the correct label for newly presented input data.

When training a supervised learning algorithm, the **training set** will consist of inputs paired with the correct outputs. Inputs in the training set should represent the **target set** which we have to classify: composition of the data, data quality.

During **training**, the algorithm will search for patterns in the data that correlate with the desired outputs.

After training, a supervised learning algorithm will take in new unseen inputs and will determine which label the new inputs will be classified based on prior training data.

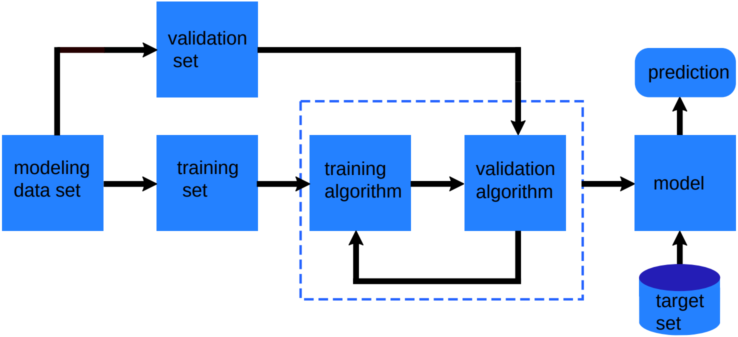
Supervised Machine Learning



split modeling set into training set and validation set:
the validation set (also: test set) is used for the testing the model after the model has been trained on the training set - it is extremely important to test the model on data not being part of the training set

A **fundamental assumption** of supervised machine learning is that the distribution of training examples is identical to the distribution of validation examples and future unseen examples (the target set).

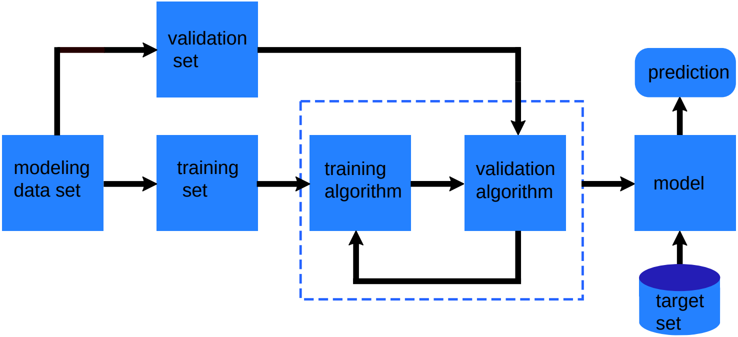
Supervised Machine Learning



Training:
given a training set of labeled examples $\{(x_1, y_1), ..., (x_n, y_n)\}$, estimate the prediction function f and parameters θ which minimizes the prediction error on the training set

Validation:
apply f to validation set x, output predicted value $y = f(x)$
from this we generate performance measures, also called accuracy measures

Supervised Machine Learning

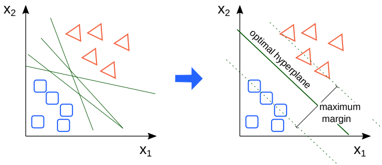


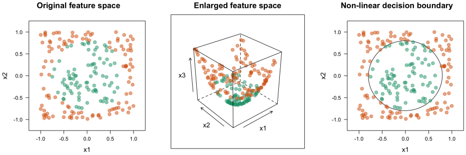
Application:
Run the model on the target set.

Supervised Machine Learning

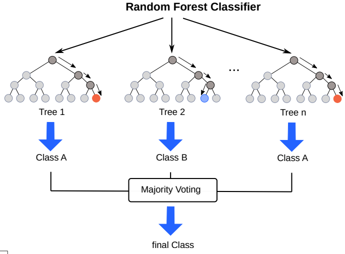
state-of-the-art (before Deep Learning):

Support Vector Machines
binary classification



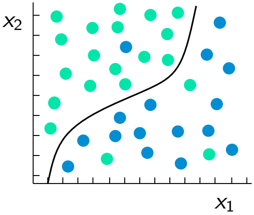


Random Forest Classifiers
multiclass classification

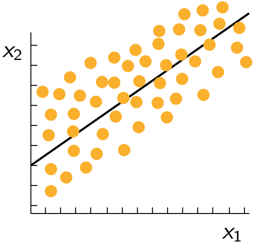


two main areas of supervised machine learning:
classification problems and **regression** problems

mapping input value(s) to a discrete value, the class
example: predicting whether an object is a star or a galaxy



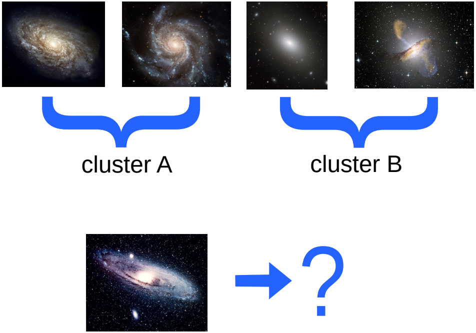
mapping input value(s) to continuous data
example: predicting the surface temperature of a star



Unsupervised Machine Learning

unlabeled data enable the unsupervised machine learning algorithm to **understand the data** and **find patterns in data themselves**

data is clustered, new data is assigned to clusters



applications: data exploration, data clustering, anomaly

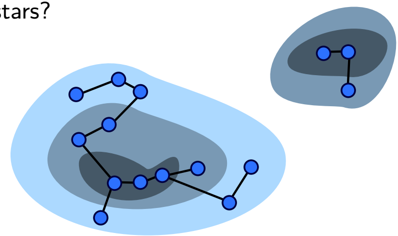
Unsupervised Machine Learning

three main areas of unsupervised machine learning:
clustering, **association** and **dimensionality reduction**

find hidden patterns in the data based on similarities or differences
example: are there subtypes within a given type of stars?

find the probability of co-occurrence of items in a collection
example: which stars likely host exoplanets?

reduce the dimensions of the data
example: feature rextraction to reduce the number of random variables



Types of Machine Learning - Overview

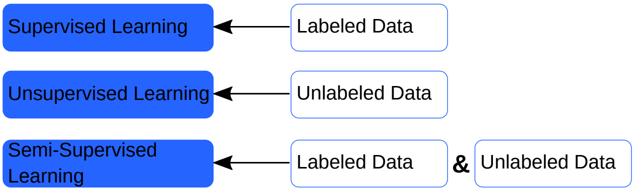
Machine Learning

- **Supervised Learning**
  - Classification — table, time series classifiction, image classification
  - Regression — predictions, continuous variables
- **Unsupervised Learning**
  - Clustering — anomaly detection, data exploration
  - Association — co-occurrence
  - Dimensionality Reduction — data visualization, feature extraction, structure discovery
- **Semi-supervised Learning** — using small labeled data set and large unlabeled
- **Reinforcement Learning** — robot navigation, survey strategy planning

The Role of Data in Training Process

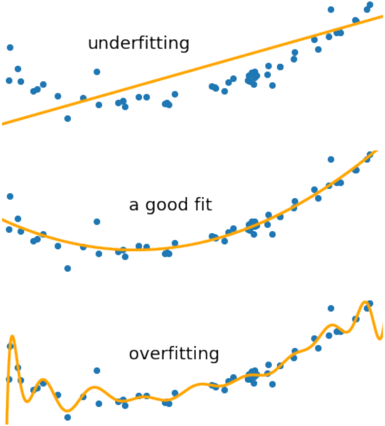
Supervised Learning learns from labeled training set data by iteratively making predictions on the data and adjusting for the correct answer. This makes supervised Learning models **more accurate** than unsupervised learning models.

Unsupervised Learning models work on their own to discover the inherent structure of unlabeled data. The unsupervised learning algorithm works with unlabeled data, in which the output is based solely on the collection of perceptions. This makes unsupervised methods **more flexible** to deal with new data.



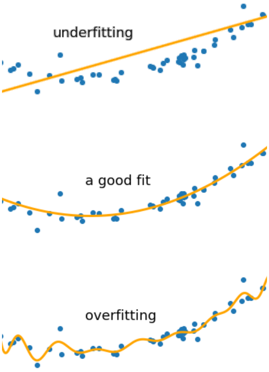
Challenges and Limitations

In most scenarios, the cause of the poor performance of any machine learning algorithm is due to **underfitting or overfitting**.



Challenges and Limitations

In most scenarios, the cause of the poor performance of any machine learning algorithm is due to **underfitting or overfitting**.



Underfitting is a scenario where the machine learning model can neither learn the relationship between variables in the data nor predict a new data point correctly. In other words, the machine learning system hasn't found a correlation between data.

Overfitting occurs when the machine learning model learns from the training data a little too much, attempting to fit every point on the curve and, as a result, memorizes the data patterns. In other words, it narrowed its focus too much on the examples given, making it unable to see the bigger picture and fails to predict new data points.

Challenges and Limitations

Underfitting can occur when:

- The model was trained using the wrong features.
- The model is too simple and can't remember enough features.
- The target data is too varied or complex - the training set doesn't represent the target data's distribution realistically.

Challenges and Limitations

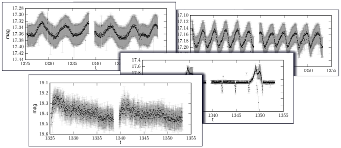
Underfitting can occur when:

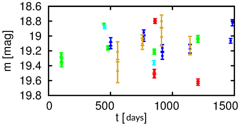
- The model was trained using the wrong features.
- The model is too simple and can't remember enough features.
- The target data is too varied or complex - the training set doesn't represent the target data's distribution realistically.

Overfitting can occur when:

- The model was trained using the wrong parameters and over-observed the training data.
- The model complexity is too high for the presented data variability.
- The training data's labels are too restrictive.

example: Training on 'nice' (high cadence, long baseline, good S/N) light curves - applied to worse. Don't do that!





Key Takeaways: Machine Learning Basics

- Machine learning is a concept that allows computers to learn and improve from experience without being explicitly programmed.
- Machine learning works by the approach of *find the pattern, apply the pattern*.
- Machine Learning consists of Supervised, Unsupervised, Reinforcement, and Semi-Supervised Learning.
- Supervised learning is useful when dealing with purely labeled datasets for training and knowing how the output should look like.
- Unsupervised Learning is useful for finding the hidden patterns.
- A machine learning model is underfitted when it fails to capture the relationship between the input and output.
- If a machine learning model performs better on the training set than on the test set, then it is likely overfitting: it memorizes the data it was trained on without being able to generalize.
- Machine learning is part of many nowadays everyday applications such as Google Maps, Alexa, Youtube...
- It is increasingly important for astronomy for such as source classification, anomaly detection, survey strategy planning...

Recap: Probability

$p(A) =$ the probability of A (or the probability density at A)

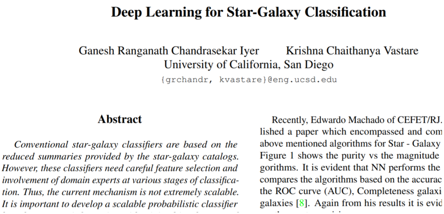
example: the probability that an observed object is a galaxy

Recap: Probability

$p(A)$ = the probability of A (or the probability density at A)

example: the probability that an observed object is a galaxy

The probability reflects our current state of knowledge of the object, and our belief that it is a galaxy.



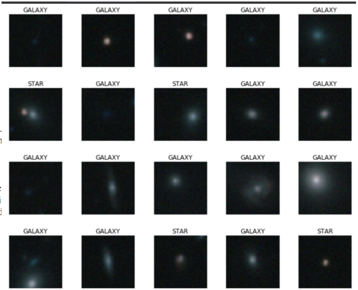
Deep Learning for Star-Galaxy Classification

Ganesh Ranganath Chandrasekar Iyer   Krishna Chaithanya Vastare
University of California, San Diego
{grchandr, kvastare}@eng.ucsd.edu

Abstract

Conventional star-galaxy classifiers are based on the reduced summaries provided by the star-galaxy catalogs. However, these classifiers need careful feature selection and involvement of domain experts at various stages of classification. Thus, the current mechanism is not extremely scalable. It is important to develop a scalable probabilistic classifier

Recently, Edwardo Machado of CEFET/RJ. lished a paper which encompassed and com above mentioned algorithms for Star - Galaxy Figure 1 shows the purity vs the magnitude gorithms. It is evident that NN performs the compares the algorithms based on the accurac the ROC curve (AUC), Completeness galaxi galaxies [8]. Again from his results it is evid



Recap: Probability

$p(A)$ = the probability of A (or the probability density at A)

Probabilities are normalized: $\sum_i p(A_i) = 1$

example: the probability that an observed object is a galaxy is 0.37, that it is a star is 0.61, and that it is somthing else is 0.02.

Notation

$A \cup B$ is the **union** of sets A and B. Read as A OR B.

Notation

$A \cup B$ is the **union** of sets A and B. Read as A OR B.

$A \cap B$ is the **intersection** of sets A and B. Read as A AND B.
Different notations $p(A \cap B) = p(A, B)$

We will use the comma notation throughout.

Kolmogorov Axioms

first axiom:
The probability of an event is a non-negative real number:

$$p(A) \geq 0 \; \forall A$$

Kolmogorov Axioms

first axiom:
The probability of an event is a non-negative real number:

$$p(A) \geq 0 \; \forall A$$

second axiom:
This is the assumption of unit measure: The probability that at least one of the elementary events in the entire sample space will occur is 1.
$p(\Omega) = 1$ where Ω is the set of all possible outcomes, i.e. the sum/ integral of all possible outcomes is 1.

Kolmogorov Axioms

first axiom:
The probability of an event is a non-negative real number:

$$p(A) \geq 0 \; \forall A$$

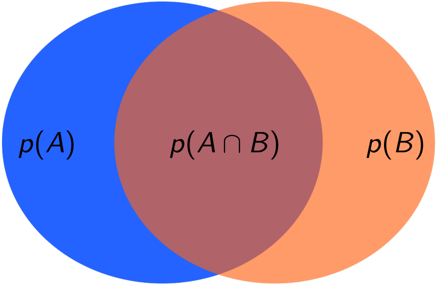
second axiom:
This is the assumption of unit measure: The probability that at least one of the elementary events in the entire sample space will occur is 1.
$p(\Omega) = 1$ where Ω is the set of all possible outcomes, i.e. the sum/ integral of all possible outcomes is 1.

third axiom:
Any countable sequence of disjoint sets (synonymous with mutually exclusive events) $A_1, A_2, \ldots$ satisfies

$$p\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} p(A_i)$$

Consequences of Kolmogorov Axioms

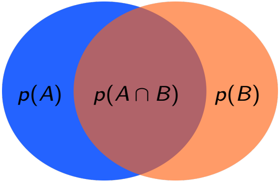
If we have two events, A and B, the possible combinations are illustrated by the following figure:



$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

Consequences of Kolmogorov Axioms

If we have two events, A and B, the possible combinations are illustrated by the following figure:



$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

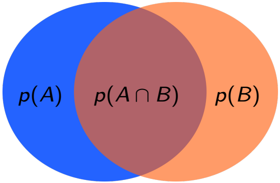
The probability that either A or B will happen (which could include both) is the union, given by

$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

The term $-p(A \cap B)$ is necessary as A and B overlap, thus we would count it twice.

Consequences of Kolmogorov Axioms

If we have two events, A and B, the possible combinations are illustrated by the following figure:



$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

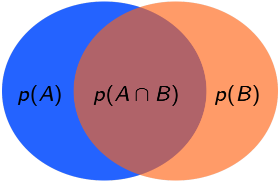
The probability that both A and B will happen, $p(A \cap B)$, is

$$p(A \cap B) = p(A|B)p(B) = p(B|A)p(A)$$

where $p(A|B)$ is the probability of A given that B is true and is called the **conditional probability** (so the $|$ is short for "given").

Consequences of Kolmogorov Axioms

If we have two events, A and B, the possible combinations are illustrated by the following figure:



$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

The law of the total probability says that (for independent A_i, B_i),

$$p(A) = \sum_i p(A|B_i)p(B_i)$$

Consequences of Kolmogorov Axioms

It is important to realize that the following is always true:
$p(A, B) = p(A|B)p(B) = p(B|A)p(A)$

However, if A and B are independent, then $p(A|B) = p(A)$ and $p(B|A) = p(B)$ and $p(A, B) = p(A)p(B)$.

Consequences of Kolmogorov Axioms

It is important to realize that the following is always true:
$p(A, B) = p(A|B)p(B) = p(B|A)p(A)$

However, if A and B are independent, then $p(A|B) = p(A)$ and $p(B|A) = p(B)$ and $p(A, B) = p(A)p(B)$.

example: classic marbles in bag scenario

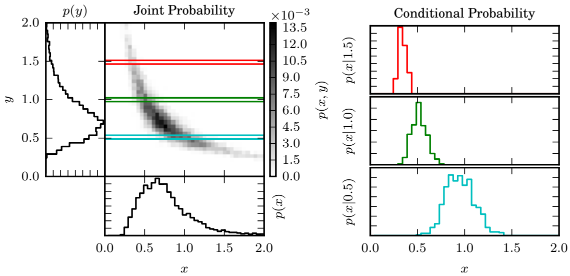
If you have a bag with 5 marbles (3 yellow and 2 blue) and you want to know the probability of picking 2 yellow marbles in a row, that would be $p(Y_1, Y_2) = p(Y_1)p(Y_2|Y_1)$ with $p(Y_1) = \frac{3}{5}$ since you have an equally likely chance of drawing any of the 5 marbles.

If you did not put the first marble back in the bag after drawing it (**sampling without replacement**), then the probability is $p(Y_2|Y_1) = \frac{2}{4}$, so that $p(Y_1, Y_2) = \frac{3}{5} \times \frac{2}{4} = \frac{3}{10}$.

But if you put the first marble back (**sampling with replacement**), then $p(Y_2|Y_1) = \frac{3}{5} = p(Y_2)$, so that $p(Y_1, Y_2) = \frac{3}{5} \times \frac{3}{5} = \frac{9}{25}$.
In the first case Y_1 and Y_2 are not independend, but in the second they are.

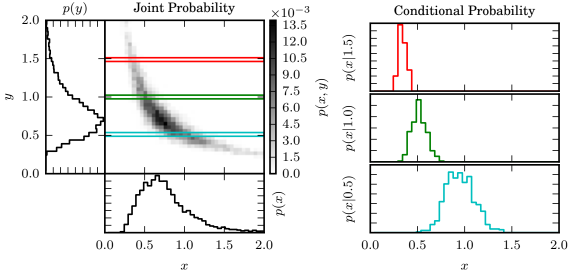
Bayes' Theorem

In this 2-D distribution in $x - y$ parameter space, x and y are not independent as, once you pick a y, your values of x are constrained.



Bayes' Theorem

In this 2-D distribution in $x - y$ parameter space, x and y are not independent as, once you pick a y, your values of x are constrained.



We have

$$p(x, y) = p(x|y)p(y) = p(y|x)p(x)$$

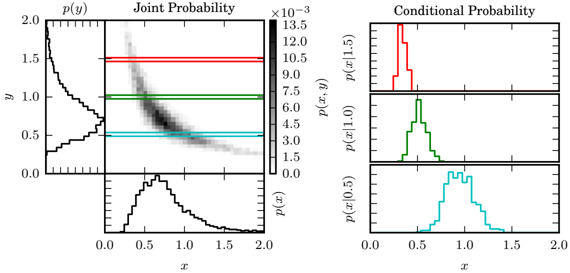
We can define the **marginal probability** as

$$p(x) = \int p(x, y)dy$$

It is shown on the left and bottom of the left panel.

Bayes' Theorem

In this 2-D distribution in $x - y$ parameter space, x and y are not independent as, once you pick a y, your values of x are constrained.

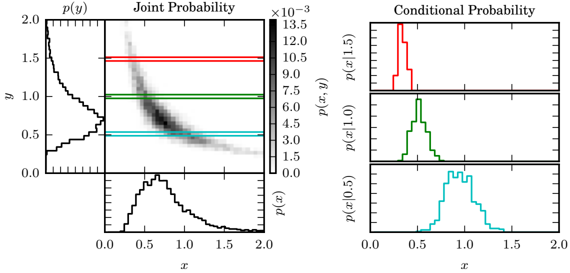


The three panels on the right show the **conditional probability** (of x) $p(x|y = y_0)$ for three values of y_0.

These are just normalized "slices" through the 2-D distribution. The marginal probability of x can be re-written using conditional probabilities as

$$p(x) = \int p(x|y)p(y)dy$$

Bayes' Theorem



But since $p(x|y)p(y) = p(y|x)p(x)$, we can write

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\int p(x|y)p(y)dy}$$

which in words says that

the (conditional) probability of y given x is the (conditional) probability of x given y times the (marginal) probability of y divided by the (marginal) probability of x, where the latter is just the integral of the numerator.

Example: Monty Hall Problem (or "Deal Or No Deal")

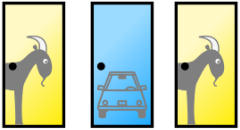
You are in a game show and are shown 2 doors. Behind one door is a car, behind the other one is a goat. What are your chances of picking the door with the car?

Bayes' Theorem

Example: Monty Hall Problem (or "Deal Or No Deal")

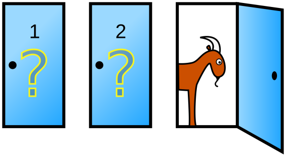
You are in a game show and are shown 2 doors. Behind one door is a car, behind the other one is a goat. What are your chances of picking the door with the car?

You are in a game show and are shown 3 doors. Behind one door is a car, behind the other two are goats. What are your chances of picking the door with the car?



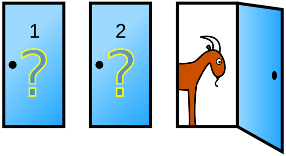
Example: Monty Hall Problem (or "Deal Or No Deal")

You are in a game show and are shown 3 doors. The game show host asks you to pick a door, but not to open it yet. Then the host opens one of the other two doors (that you did not pick), making sure to select one with a goat. The host offers you the opportunity to **switch** doors. Do you?



Bayes' Theorem

Example: Monty Hall Problem (or "Deal Or No Deal")

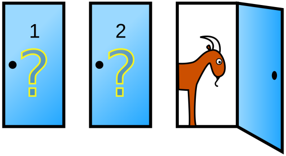


What we know about the **probabilities**:

Probability of car behind Door 1: $1/3$
Probability of car behind Doors 2 or 3: $2/3$

Bayes' Theorem

Example: Monty Hall Problem (or "Deal Or No Deal")



What we know about the **probabilities**:

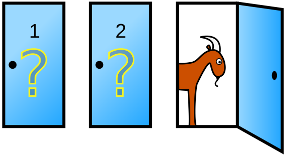
Probability of car behind Door 1: $1/3$
Probability of car behind Doors 2 or 3: $2/3$

Probability of you had picked the car: $1/3$
$\Rightarrow$ host can open either of the other doors

Probability of you had picked a goat (one of the two goats): $2/3$
$\Rightarrow$ host opens door that is also a goat, remaining door has the car

Bayes' Theorem

Example: Monty Hall Problem (or “Deal Or No Deal”)



What we know about the **probabilities**:

Probability of car behind Door 1: $1/3$
Probability of car behind Doors 2 or 3: $2/3$

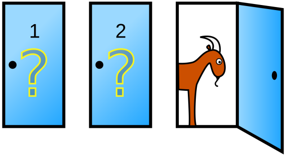
Probability of you had picked the car: $1/3$
$\Rightarrow$ host can open either of the other doors

Probability of you had picked a goat (one of the two goats): $2/3$
$\Rightarrow$ host opens door that is also a goat, remaining door has the car

The **advice** is to switch to the door you hadn't chosen and that's not open yet - precisely, switching doubles your chances.

Bayes' Theorem

Example: Monty Hall Problem (or "Deal Or No Deal")



This is an example of the use of **conditional probability**, where we have $p(A|B) \neq p(A)$.

Notation

In most textbook and here:

x is a scalar quantity that is measured N times to form a dataset

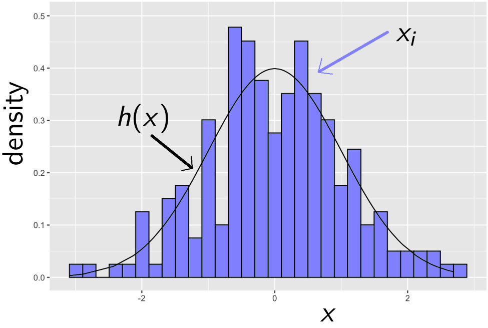
x_i is a single measurement with $i = 1, ..., N$

$\{x_i\}$ refers to the set of all N measurements comprising the dataset

measurements (data) can be real numbers, discrete labels (strings or numbers), or even “missing values” (we sometimes pad our datasets with `NaN` in this case)

idea:

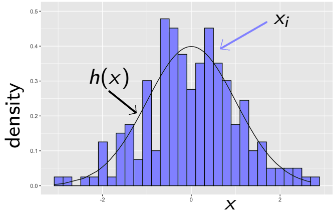
- measurements are drawn from an underlying probability distribution function (pdf) $h(x)$
- we can only observe the measurements x_i, not the underlying pdf



Goal of Statistical Inference

idea:

- measurements are drawn from an underlying probability distribution function (pdf) $h(x)$
- we can only observe the measurements x_i, not the underlying pdf

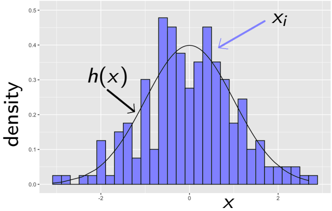


using measurements x_i, we are trying to estimate the probability density (distribution) function or the *pdf* $h(x)$ from which the individual x_i are drawn

Goal of Statistical Inference

idea:

- measurements are drawn from an underlying probability distribution function (pdf) $h(x)$
- we can only observe the measurements x_i, not the underlying pdf

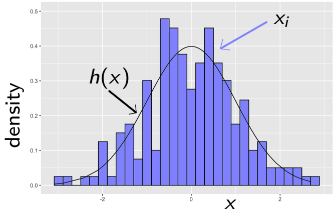


Question: What is the probability of a value lying between x and $x + \mathrm{d}x$, where $\mathrm{d}x$ is infinitesimal?

Goal of Statistical Inference

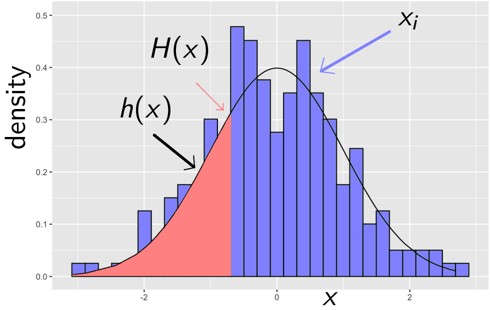
idea:

- measurements are drawn from an underlying probability distribution function (pdf) $h(x)$
- we can only observe the measurements x_i, not the underlying pdf



Question: What is the probability of a value lying between x and $x + \mathrm{d}x$, where $\mathrm{d}x$ is infinitesimal?
Answer: $h(x)\,\mathrm{d}x$

The "left to right" integral of $h(x)$ is the **cumulative distribution function** (cdf), $H(x) = \int_{-\infty}^{x} h(x')\,\mathrm{d}x'$.



The inverse function of the cdf is the **quantile function**, answering the question: Which x value has e.g. 90% of the distribution below it?

Empirical Distribution Function

Don't neglect measurement errors:

Using measurements x_i, we are trying to estimate the probability density (distribution) function or the *pdf* $h(x)$ from which the individual x_i are drawn.

Empirical Distribution Function

Don't neglect measurement errors:

Using measurements x_i, we are trying to estimate the probability density (distribution) function or the *pdf* $h(x)$ from which the individual x_i are drawn.

While $h(x)$ is the underlying pdf (also called *population pdf*), what we measure from the data is the **empirical pdf** $f(x)$.

Empirical Distribution Function

Don't neglect measurement errors:

Using measurements x_i, we are trying to estimate the probability density (distribution) function or the *pdf* $h(x)$ from which the individual x_i are drawn.

While $h(x)$ is the underlying pdf (also called *population pdf*), what we measure from the data is the **empirical pdf** $f(x)$.

So, $f(x)$ is a model of $h(x)$. In principle, with infinite data $f(x) \rightarrow h(x)$, but in reality the blurring effect of **measurement errors** keep this from being strictly true. Likewise, the empirical cdf (cumulative density function) is denoted $F(x)$.

Errors and Uncertainties

Technically, **errors** are systematic biases that we can not mitigate through collecting lots and lots of data.

Statistical uncertainties are the result of random measurement uncertainty.

But “error” will be used for both, and denoted as either statistical errors (error bars) or systematic errors (biases).

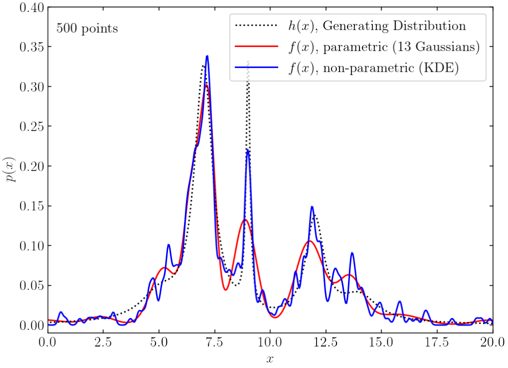
Statistical error distributions (error bars) that vary from data point to data point are called heteroscedastic errors (usually the case in astronomy). If they are the same for all points then they are homoscedastic errors.

Physical Models

idea: We can either:

- Describe the data, then the process is non-parametric, i.e. we are just trying to **describe** the data behavior in a compact practical way.
- Guess a physical model for $h(x)$, then the process is parametric.

  From a **model** we can generate new data that mimic measurements.

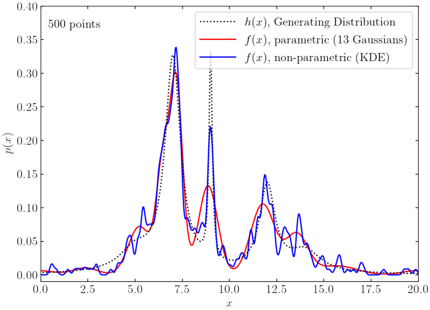


Physical Models

idea: We can either:

- Describe the data, then the process is non-parametric, i.e. we are just trying to **describe** the data behavior in a compact practical way.
- Guess a physical model for $h(x)$, then the process is parametric.

  From a **model** we can generate new data that mimic measurements.



$\Rightarrow$ We will see later on how to do that with Python!

Distributions

A **probability distribution** is the mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment. It is a mathematical description of a random phenomenon in terms of its sample space and the probabilities of events (subsets of the sample space).

Distributions

A **probability distribution** is the mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment. It is a mathematical description of a random phenomenon in terms of its sample space and the probabilities of events (subsets of the sample space).

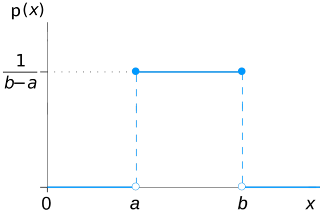
What's the point of all these distributions?

To understand the **significance** of a measurement, we want to know how likely it is that we would get that measurement in our experiment by random chance. To determine that we need to know the shape of the distribution.
If we are attempting to characterize our data in a way that is **parameterized**, then we need a functional form for a distribution.

Uniform Distribution

The probability density function of the **continuous uniform distribution** is given by:

$$p(x|a,b) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases}$$



The mean (first moment) of the distribution is:

$$E(X) = \frac{1}{2}(b+a)$$

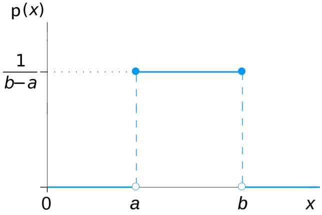
The variance (second central moment) is:

$$V(X) = \frac{1}{12}(b-a)^2$$

Uniform Distribution

The probability density function of the **continuous uniform distribution** is given by:

$$p(x|a,b) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases}$$



The mean (first moment) of the distribution is:

$$E(X) = \frac{1}{2}(b+a)$$

The variance (second central moment) is:

$$V(X) = \frac{1}{12}(b-a)^2$$

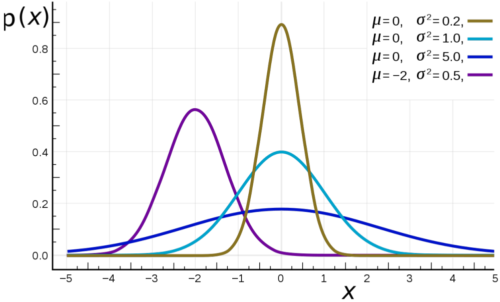
Random numbers drawn from uniform distributions are the basis of all **(pseudo-)random number generators**.

Gaussian Distribution

The probability density function of the **Gaussian distribution** is given by:

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

It is also called the **normal distribution** and can be noted by $\mathcal{N}(\mu,\sigma)$.



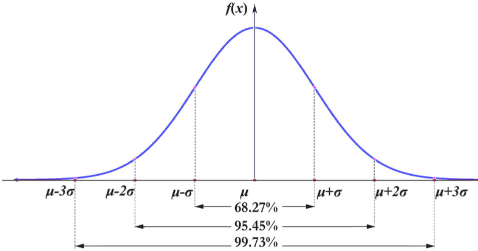
Gaussian Distribution

Gaussian confidence levels

The probability of a measurement drawn from a Gaussian distribution that is between $\mu - a$ and $\mu + b$ is

$$\int_{\mu-a}^{\mu+b} p(x|\mu,\sigma)\,\mathrm{d}x$$

examples:
for a $=$ b $= 1\sigma$, we get 68.3%
for a $=$ b $= 2\sigma$, we get 95.4%
for a $=$ b $= 3\sigma$, we get 99.7%



We refer to the ranges $\mu \pm 1\sigma$, $\mu \pm 2\sigma$, $\mu \pm 3\sigma$ as the 68%, 95% and 99% **confidence limits**, respectively. Note: These numbers are only valid for Gaussian distributions (but can be calculated for other distributions).

Gaussian confidence levels

The probability of a measurement drawn from a Gaussian distribution that is between $\mu - a$ and $\mu + b$ is

$$\int_{\mu-a}^{\mu+b} p(x|\mu,\sigma)\,\mathrm{d}x$$

confidence level vs. confidence interval:

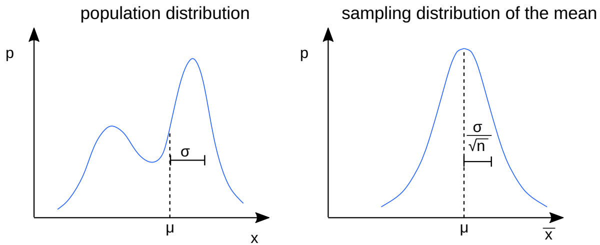
The **confidence level** is the percentage of all possible samples that are expected to include the true population parameter: the percentage of times you expect to get close to the same estimate if you repeat the experiment.

The **confidence interval** is the upper and lower bounds of the estimate you expect to find at a given level of confidence - an interval that is likely to contain an unknown population parameter.

The Central Limit Theorem

The central limit theorem states that the distribution of a normalized version of the sample mean converges to a standard normal distribution. This holds even if the original variables are not normally distributed.

Let $X_1, \ldots, X_n$ denote a random sample of n independent observations from a population with overall expectation value (average) μ and finite variance σ^2, and let $\bar{X}_n$ denote the sample mean of that sample (which is itself a random variable). Then the limit as $n \to \infty$ of the distribution of $\frac{\bar{X}_n - \mu}{\sigma_{\bar{X}_n}}$, where $\sigma_{\bar{X}_n} = \frac{\sigma}{\sqrt{n}}$, is the standard normal distribution.

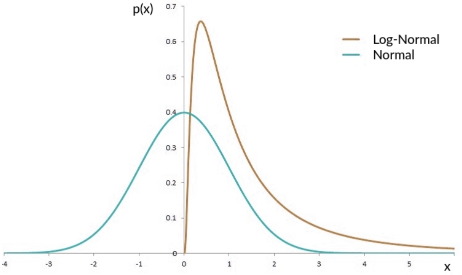


Log-Normal Distribution

If x is Gaussian distributed with $\mathcal{N}(\mu, \sigma)$, then $y = \exp(x)$ has a **log-normal distribution**, where the mean of y is $\exp(\mu + \sigma^2/2)$, the median is $\exp(\mu)$ and the mode is $\exp(\mu - \sigma^2)$.

The probability distribution function for a log-normal distribution is

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

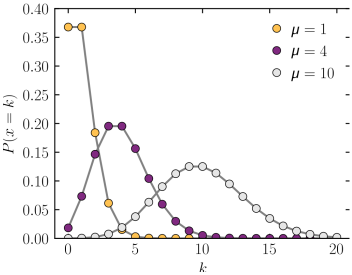


Poisson Distribution

The **Poisson distribution** is a distribution for a discrete variable, telling the probability of k events occuring within a certain time when the mean is μ:

$$p(k|\mu) = \frac{\mu^k \exp(-\mu)}{k!}$$

The mean μ completely characterizes the distribution. The mode is $(\mu - 1)$, the standard deviation is $\sqrt{\mu}$.
As μ increases, the Poisson distribution becomes more and more similar to a Gaussian with $\mathcal{N}(\mu, \sqrt{\mu})$.



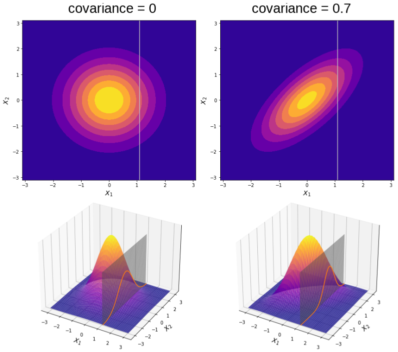
In the plot, the horizontal axis k, the number of occurrences. μ is the expected rate of occurrences. The vertical axis is the probability of k occurrences given μ. The function is defined only at integer values of k.

Higher-Dimensional Distributions

A common higher-dimensional distribution is the **Multivariate Gaussian**:

$$\mathcal{N}(\vec{x}|\vec{\mu},\mathbf{V}) = \frac{1}{(2\pi)^{d/2}|\det \mathbf{V}|}\exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T\mathbf{V}^{-1}(\vec{x}-\vec{\mu})\right)$$

e.g. in 2D: $$\vec{\mu} = \begin{pmatrix}\mu_x \\ \mu_y\end{pmatrix}, \quad \mathbf{V} = \begin{pmatrix}\sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2\end{pmatrix}$$



Population Covariance:

$$\mathrm{Cov}(x,y) = \frac{\sum_i (x_i-\mu_x)(x_i-\mu_y)}{n}$$

Marginal distributions:

integrate over x: $\mathcal{N}(y|\mu_y,\sigma_y^2)$

integrate over y: $\mathcal{N}(x|\mu_x,\sigma_x^2)$

Descriptive Statistics

Our goal is to **estimate** a probabiliy distribution function $h(x)$ given some measured data, by reconstructing the data-based distribution $f(x)$.

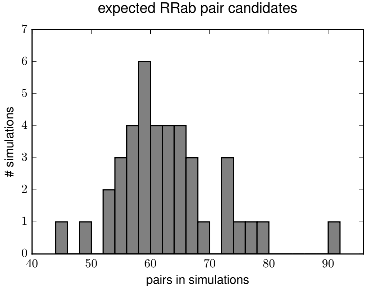
An arbitrary distribution* can be characterized by location parameters (i.e., position), scale parameters (i.e., width), and shape parameters. These parameters are called **descriptive statistics**.

Descriptive Statistics

Our goal is to **estimate** a probabiliy distribution function $h(x)$ given some measured data, by reconstructing the data-based distribution $f(x)$.

An arbitrary distribution* can be characterized by location parameters (i.e., position), scale parameters (i.e., width), and shape parameters. These parameters are called **descriptive statistics**.

*can be anything, e.g. the distribution of velocities in a globular cluster, the distribution of the number of epochs in light curves...



Descriptive Statistics

mean of a sample:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

This is the **sample arithmetic mean**

Descriptive Statistics

mean of a sample:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

This is the **sample arithmetic mean**
which is derived as the first momentum of the distribution from Monte Carlo integration as

$$\mu = E(x) = \langle x \rangle = \int_{-\infty}^{+\infty} x h(x) d(x) \approx \frac{1}{N}\sum_{i=1}^{N} x_i$$

where $\{x_i\}$ are random samples from the properly normalized $h(x)$, and $E(\cdot)$ is the **expectation value**

Descriptive Statistics

median of a sample:

The **median** is a more robust estimator than the mean.

Descriptive Statistics

median of a sample:

The **median** is a more robust estimator than the mean.
The median is defined as follows: For a distribution x with n elements, ordered from smallest to greatest,

if n is odd,

$$\mathrm{median}(x) = x_{(n+1)/2}$$

if n is even,

$$\mathrm{median}(x) = \frac{x_{(n/2)} + x_{(n/2)+1}}{2}$$

median of a sample:

The **median** is a more robust estimator than the mean.
The median is defined as follows: For a distribution x with n elements, ordered from smallest to greatest,

if n is odd,

$$\mathrm{median}(x) = x_{(n+1)/2}$$

example:
1 3 3 4 7 8 128

if n is even,

$$\mathrm{median}(x) = \frac{x_{(n/2)} + x_{(n/2)+1}}{2}$$

median of a sample:

The **median** is a more robust estimator than the mean.
The median is defined as follows: For a distribution x with n elements, ordered from smallest to greatest,

if n is odd,

$$\mathrm{median}(x) = x_{(n+1)/2}$$

example:

1 3 3 **4** 7 8 128

median $= 4$

if n is even,

$$\mathrm{median}(x) = \frac{x_{(n/2)} + x_{(n/2)+1}}{2}$$

Descriptive Statistics

median of a sample:

The **median** is a more robust estimator than the mean.
The median is defined as follows: For a distribution x with n elements, ordered from smallest to greatest,

if n is odd,

$$\mathrm{median}(x) = x_{(n+1)/2}$$

example:
1 3 3 **4** 7 8 128

median = 4

if n is even,

$$\mathrm{median}(x) = \frac{x_{(n/2)} + x_{(n/2)+1}}{2}$$

example:
1 3 3 4 7 8

median of a sample:

The **median** is a more robust estimator than the mean.
The median is defined as follows: For a distribution x with n elements, ordered from smallest to greatest,

if n is odd,

$$\text{median}(x) = x_{(n+1)/2}$$

example:

1 3 3 **4** 7 8 128

median $= 4$

if n is even,

$$\text{median}(x) = \frac{x_{(n/2)} + x_{(n/2)+1}}{2}$$

example:

1 3 **3** **4** 7 8

median $= 3.5$

Descriptive Statistics

Other descriptive statistics are related to **higher order moments** of the distribution:

"average" location value (the mean, median) discussed before
$\Rightarrow$ information about *deviations* from the average (which is related to the shape of the distribution)

Descriptive Statistics

Other descriptive statistics are related to **higher order moments** of the distribution:

"average" location value (the mean, median) discussed before $\Rightarrow$ information about *deviations* from the average (which is related to the shape of the distribution)

One could start with a deviation such as

$$d_i = x_i - \mu$$

Descriptive Statistics

Other descriptive statistics are related to **higher order moments** of the distribution:

"average" location value (the mean, median) discussed before $\Rightarrow$ information about *deviations* from the average (which is related to the shape of the distribution)

One could start with a deviation such as

$$d_i = x_i - \mu$$

However, the average deviation is zero by definition of the mean.

Descriptive Statistics

Other descriptive statistics are related to **higher order moments** of the distribution:

One can compute the **mean absolute deviation (MAD)**:

$$\frac{1}{N}\sum |x_i - \mu|$$

but the absolute values can hide the true scatter of the distribution in some cases

Descriptive Statistics

Other descriptive statistics are related to **higher order moments** of the distribution:

Better idea: square the differences

$$\sigma^2 = \frac{1}{N} \sum (|x_i - \mu|)^2$$

which is the **variance**.
The variance V is the expectation value of $(x-\mu)^2$ (and related to the 2nd moment)

$$\sigma^2 = V = E((x-\mu)^2) \int_{-\infty}^{+\infty} (x-\mu)^2 h(x)\,\mathrm{d}x$$

where σ is the **standard deviation**.

Descriptive Statistics

Other descriptive statistics are related to **higher order moments** of the distribution:

Better idea: square the differences

$$\sigma^2 = \frac{1}{N} \sum (|x_i - \mu|)^2$$

which is the **variance**.
The variance V is the expectation value of $(x - \mu)^2$ (and related to the 2nd moment)

$$\sigma^2 = V = E((x-\mu)^2) \int_{-\infty}^{+\infty} (x-\mu)^2 h(x)\,\mathrm{d}x$$

where σ is the **standard deviation**.

For discrete distributions, the integral gets replaced by a sum.

The **Median Absolute Deviation** (also MAD) given by

$$\mathrm{median}(|x_i - \mathrm{median}(\{x_i\})|)$$

Descriptive Statistics

In statistics and probability, **quantiles** are cut points dividing the range of a probability distribution into continuous intervals with equal probabilities, or dividing the observations in a sample in the same way.

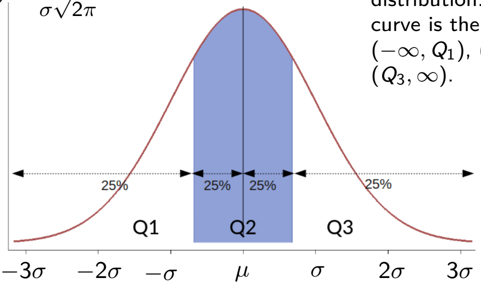
A **quartile** is a quantile which divides the distribution into four parts of equal size.

Descriptive Statistics

In statistics and probability, **quantiles** are cut points dividing the range of a probability distribution into continuous intervals with equal probabilities, or dividing the observations in a sample in the same way.

A **quartile** is a quantile which divides the distribution into four parts of equal size.

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



Probability density of a normal distribution. The area below the red curve is the same in the intervals $(-\infty, Q_1)$, (Q_1, Q_2), (Q_2, Q_3), and (Q_3, ∞).

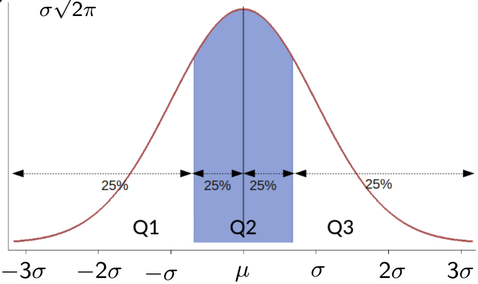
Descriptive Statistics

P % quantiles (or the p^{th} **percentile**, q_p) are computed as

$$\frac{p}{100} = H(q_p) = \int_{-\infty}^{q_p} h(x)\,\mathrm{d}x$$

The full integral from $-\infty$ to ∞ is 1 (100%). So, here you are looking for the value of x that accounts for P percent of the distribution.
For example, the 25^{th}, 50^{th}, and 75^{th} percentiles are just Q1, Q2, Q2.

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

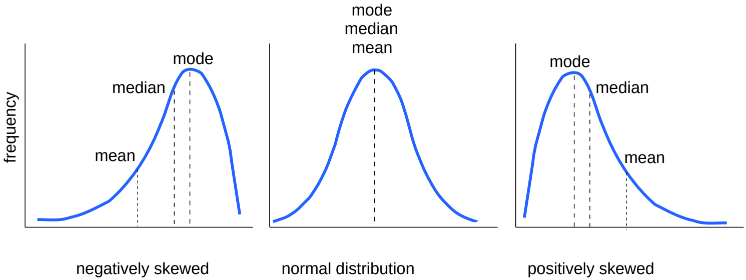


Other useful shape measures include the "higher order" moments:

Skewness:

$$\Sigma = \int_{-\infty}^{\infty} \left(\frac{x - \mu}{\sigma} \right)^3 h(x) \, \mathrm{d}x$$

The skewness measures the distribution's asymmetry. Distributions with long tails to larger x values have positive Σ.



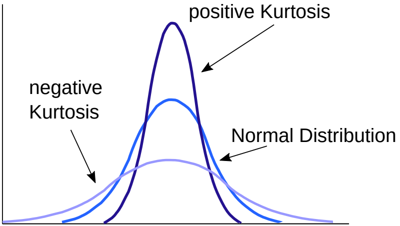
Descriptive Statistics

Other useful shape measures include the "higher order" moments:

Kurtosis:

$$K = \int_{-\infty}^{\infty} \left(\frac{x - \mu}{\sigma} \right)^4 h(x)\,\mathrm{d}x - 3$$

The kurtosis measures how peaked or flat-topped a distribution is, with strongly peaked ones being positive and flat-topped ones being negative. K is calibrated to a Gaussian distribution (hence the subtraction of 3).

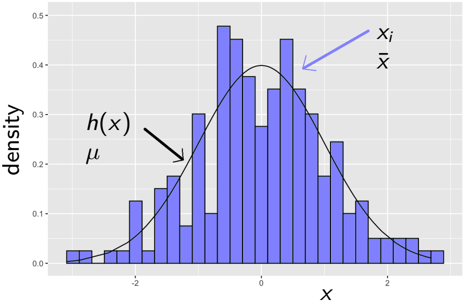


Sample versus Population Statistics

Statistics estimated from data are called **sample statistics**, compared to **population statistics** derived from knowing the functional form of the pdf.

Specifically, μ is the **population mean**, i.e. it is the expectation value of x for $h(x)$. As $h(x)$ is unknown, the **sample mean** $\bar{x}$ is an estimator for μ:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

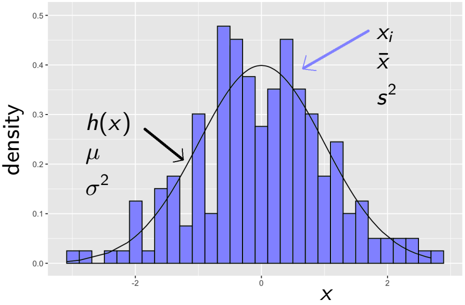


Sample versus Population Statistics

Instead of the **population variance** σ^2, we have the **sample variance** s^2, where

$$s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - x)^2$$

The denominator $N - 1$ (instead of N) accounts for the fact that we determine x from the data itself instead of using a known μ. Ideally one tries to work in a regime where N is large enough to ignore this.



Uncertainty of Sample Statistics

What are the uncertainties of our estimates $\bar{x}$ and s^2?

Uncertainty of Sample Statistics

What are the uncertainties of our estimates $\bar{x}$ and s^2?

Note that s is the width estimate of the underlying distribution; it is not the uncertainty of $\bar{x}$.

Uncertainty of Sample Statistics

What are the uncertainties of our estimates $\bar{x}$ and s^2?

Note that s is the width estimate of the underlying distribution; it is not the uncertainty of $\bar{x}$.

The uncertainty of $\bar{x}$, $\sigma_{\bar{x}}$, is

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{N}}$$

which we call the **standard error of the mean**. The uncertainty of s itself is

$$\sigma_s = \frac{s}{\sqrt{2(N-1)}} = \frac{1}{\sqrt{2}}\sqrt{\frac{N}{N-1}}\sigma_{\bar{x}}$$

Note that for large N, $\sigma_{\bar{x}} \sim \sqrt{2}\sigma_s$, and for small N, σ_s is not much smaller than s.

Transformations of Random Variables

If x is a random variable then $f(x)$ is also a random variable for any function f. To transform probability distributions when taking functions of random variables, we can simply use conservation of dimensionless probability, i.e.

$$\mathrm{Prob}(x, x + \mathrm{d}x) = \mathrm{Prob}(y, y + y)$$

$\Rightarrow p(x)\,\mathrm{d}x = p(y)dy$ with $y = f(x)$

Thus

$$p(y) = |\mathrm{d}x/\mathrm{d}y|p(x)$$

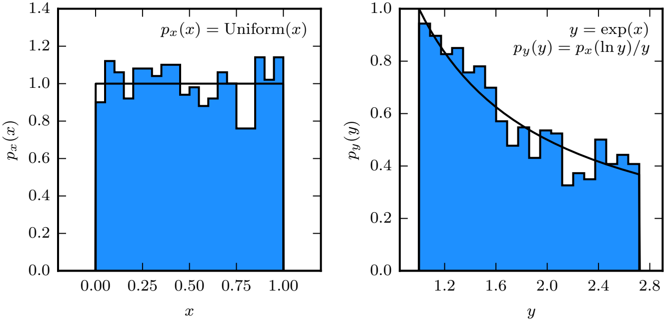
Example:
Let x be a random variable drawn from a uniform distribution between 0 and 1. So $p(x) = 1/(1-0) = 1$. Let's transform to $y = e^x$:
$p(y) = \left|\frac{\mathrm{d}y}{\mathrm{d}x}\right|^{-1} p(x) = 1/y$

Transformations of Random Variables

Example:

Let x be a random variable drawn from a uniform distribution between 0 and 1. So $p(x) = 1/(1-0) = 1$. Let's transform to $y = e^x$:

$p(y) = \left|\frac{\mathrm{d}y}{\mathrm{d}x}\right|^{-1} p(x) = 1/y$



Summary & Outlook

We have seen what **Machine Learning** means.

By reviewing the basics of **statistics** you might have seen in former courses we are now familiar with the building blocks of what we will see in the next lecture: **Statistical Learning**