Crash Course on ML & AI for Economists:

What Mathematics Offers for Economics & Business

Iker Caballero Bragagnini[1]

[1] In representation of Virus Matemático

16/10/2024

Outline

Motivation

This series of talks is intended to give young economics and business students a good sense and understanding of the statistical foundations of Machine Learning and Artificial Intelligence techniques applied in economics and business. We hope this can be a good introduction to new techniques that can be helpful for students' future careers and research.

The main goal of Virus Matemático is to raise awareness of the importance of mathematics and its usefulness in daily life.

Virus Matemático

Virus Matemático is a student-led initiative in Spain to promote interest and mathematics education through playful and creative activities. The idea is to "infect" participants with a passion for mathematics by using innovative and collaborative strategies.

Figure: Members and collaborators in a street event with children.

About Me

- Education:
  - MSc in Computational and Mathematical Engineering
  - MSc in Statistics and Operations Research
  - BSc in Economics with Minor in Applied Maths
- Work Experience:
  - Quantitative Researcher & Developer - Metron Trading
  - Quantitative Credit Risk Analyst - Deloitte
  - Research Intern - Centre de Recerca Matemàtica
  - Research Intern - RISKcenter (Universitat de Barcelona)
- Interests:
  - Financial Mathematics & Statistics
  - Education
  - Academic Research
- Contact:
  - LinkedIn
  - Twitter

Summary of this Crash Course

1. Foundations of Machine Learning and Artificial Intelligence
    - What is ML & AI?
    - What is their relationship to Mathematics and Statistics?
    - Why are they important for economics and business?
    - Why are they important to you?
    - Multivariate Statistics
    - Statistical Learning Theory
2. ML & AI for Macroeconomics
    - Linear Models for Regression and Classification
    - Decision Trees and Related Models
    - Neural Networks
    - How and why ML and AI are applied to Microeconomics?

Summary of this Crash Course

3. ML & AI for Microeconomics
    - Discrete Choice Models
    - Causality
    - Simulation and Reinforcement Learning
    - How and why ML and AI are applied to Microeconomics?
4. ML & AI for Finance
    - Time Series
    - Statistical Finance
    - Advances in Financial Machine Learning
    - How and why ML and AI are applied to Finance?

Outline of this section

1. What is ML & AI?
2. What is their relationship to Mathematics and Statistics?
3. Why are they important for economics and business?
4. Why are they important to you?
5. Multivariate Statistics

What is ML & AI?

You might commonly hear people or even experts use Machine Learning (ML) and Artificial Intelligence (AI) interchangeably, but even though both fields are closely related, they differ in some aspects. It is important to understand first what we are dealing with.

- **Artificial Intelligence:** Simulation (?) of human intelligence by machines, enabling them to perform tasks that would require human intelligence such as reasoning, learning, problem-solving, or even more advanced tasks given the latest advances.
- **Machine Learning:** Subset of AI that enables machines to learn from data and improve their performance over time without being explicitly programmed. These algorithms identify patterns, make predictions, and adapt based on experience (new data).

What is ML & AI?

Some similarities

- Both aim to create intelligent systems capable of performing tasks autonomously.
- Both rely on datasets to identify patterns and improve performance.

Some differences

- AI is a broad field encompassing ML, robotics, neuroscience, and more, while ML specifically focuses on learning from the data.
- AI aims to replicate cognitive functions, while ML is more about pattern recognition and optimization based on data (no need to replicate human intelligence).

What is its relation to Mathematics and Statistics?

Mathematics plays a crucial role in the development of different techniques and models in ML and AI, encompassing areas such as:

- **Linear Algebra:** Matrices and vector operations.
- **Calculus:** Gradients and numerical integration.
- **Probability and Combinatorics**: Distributions and counting.
- **Optimization**: Optimization algorithms and loss functions.

Moreover, statistics is a key tool that uses mathematical tools and results from areas such as:

- **Inference:** Allows to conclude and predict based on data.
- **Probability**: Allows to model randomness for different tasks.
- **Many more...**

Why is it important for economics and business?

ML & AI enhance the ability to handle large datasets, provide more accurate predictions, automate complex tasks, and uncover insights.

AI & ML

- Handle massive datasets and real-time information processing.
- Uses complex algorithms to capture non-linear patterns to forecast
- Greater personalization based on user data.
- Scalable across datasets and complex environments.
- Automation of complex tasks.

Classical methods

- Manual or rule-based systems that use smaller datasets.
- Rely upon econometric models and regressions.
- Broadly segmented based on limited data.
- Difficult to scale as data grows
- Automation of simple repetitive tasks.

Why are they important to you?

New job opportunities for economics/business students:

- Data Scientists
- Quantitative Analyst
- Policy Analysts
- Business Intelligence
- Economic Forecasters
- Supply Chain Analysts
- Marketing Analysts
- ...

All of these opportunities are fresh out of the oven for you, and most of them now use ML & AI techniques. Therefore, **you need to understand how the techniques and models used in these roles work.**

Translation: You need to understand the maths and the statistics behind them.

Why are they important to you?

A **common mistake** of economics and business students (and surely other students as well) who want to get these quantitative roles is that they hustle their way to the interview phase just to get demolished by **technical** (but most times **conceptual**) questions about **statistics** and **mathematics**.

In this series, we tackle this problem, so that each of the students can conceptually understand the mathematics and statistics used in this role and that create the building blocks of ML & AI. **Stay tuned**.

Multivariate Statistics

Before delving into the specific techniques and models in ML & AI, we first need to review some basic statistics that will help us throughout the discussions.

Because we normally deal with various variables/attributes, we will start with multivariate statistics, where we will understand:

- Multivariate Data
- Random Vectors
- Multivariate Statistics
- Distances
- Multivariate Distributions
- The Curse of Dimensionality

Multivariate Statistics: Multivariate Data

The databases you normally deal with are multivariate by nature, which means that you will always have many observations of different characteristics or attributes. These are normally represented in a table like the following (is this the only way?):

	Variable 1	Variable 2	$\cdots$	Variable k	$\cdots$	Variable p
Item 1:	x_{11}	x_{12}	$\cdots$	x_{1k}	$\cdots$	x_{1p}
Item 2:	x_{21}	x_{22}	$\cdots$	x_{2k}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
Item j:	x_{j1}	x_{j2}	$\cdots$	x_{jk}	$\cdots$	x_{jp}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
Item n:	x_{n1}	x_{n2}	$\cdots$	x_{nk}	$\cdots$	x_{np}

Figure: Table of multiple measurements for n items and p attributes.

Multivariate Statistics: Multivariate Data

We can take this table representation of the data as a two-dimensional $n \times p$ array, which we can think about as a matrix:

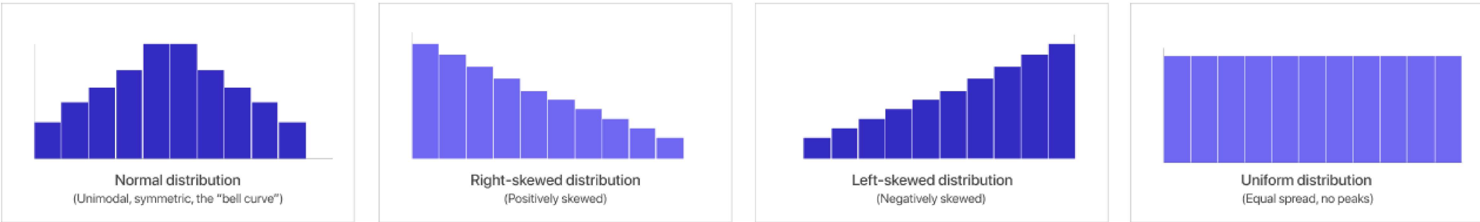
$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix}$$

where x_{jk} is the measurement of the kth attribute for the jth item or individual. Given this representation, we can use **vectors** and **matrices** to easily compute different results.

Multivariate Statistics: Multivariate Data

However, the data does not tell us anything by putting it into a matrix or a table, we need to use **probability** and **statistics** to obtain useful insights. The following logic motivates their usage:

- We have n instances for each of the p attributes. This means we have n examples of each characteristic, which will be distributed o organized in some way (depending on the nature of the process generating the data).



- Hence, we can model an attribute k as **random variable** with some distribution, for which an observation j is a draw of this underlying distribution.

$$x_k \sim Dist_k(\theta) \quad \text{for} \quad k = 1, 2, ..., p \;\Rightarrow\; x_{1k}, x_{2k}, ..., x_{nk}$$

Multivariate Statistics: Random Vectors

Most of you will have notions of univariate and bivariate statistics, but we commonly need to deal with $p > 2$ attributes in ML and AI. Therefore, we go one step further and abstract to **random vectors**, which is simply a p-dimensional vector with p random variables as entries.

$$\mathbf{x}_j = \begin{bmatrix} x_{j1} & x_{j2} & \dots & x_{jk} & \dots & x_{jp} \end{bmatrix} \quad \text{for} \quad j = 1, 2, ..., n$$

If we take into account that in our database we would have n draws of the p variables (results are realized), we can obtain the matrix $\mathbf{X}$ of data again.

Now, this allows us to operate with multiple random variables at the same time using **linear algebra**. Most ML and AI techniques are explained using vectors and matrices because it is easier to understand.

Multivariate Statistics: Multivariate Statistics

We can compute descriptive statistics in more than 1 dimension (not just a number):

- Sample Mean:

$$\bar{\mathbf{x}} = \frac{1}{n}\begin{bmatrix}\mathbf{1}^T\mathbf{x}_1 & \mathbf{1}^T\mathbf{x}_2 & \dots & \mathbf{1}^T\mathbf{x}_p\end{bmatrix}$$

$$\text{where} \quad \mathbf{1}^T\mathbf{x}_k = \frac{\sum_{j=1}^{n} x_{jk}}{n} \quad \text{for} \quad k = 1, 2, ..., p \quad \& \quad \mathbf{1} = \begin{bmatrix}1 & 1 & \dots & 1\end{bmatrix}$$

Multivariate Statistics: Multivariate Statistics

- Sample Covariance:

$$\widehat{Cov}(\mathbf{x}_i, \mathbf{x}_k) = \sigma_{i,k} = \frac{1}{n-1}\tilde{\mathbf{x}}_i^T \tilde{\mathbf{x}}_k = \frac{\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)}{n-1}$$

$$\text{where} \quad \tilde{\mathbf{x}}_k = \begin{bmatrix} x_{1k} - \bar{x}_k & \dots & x_{nk} - \bar{x}_k \end{bmatrix} \quad \text{for} \quad k = 1, 2, ..., p$$

$$\textbf{recall:} \quad Cov(\mathbf{x}_k, \mathbf{x}_k) = Var(\mathbf{x}_k)$$

- Sample Covariance Matrix:

$$\tilde{\mathbf{S}} = \left[\sigma_{ik} = \frac{1}{n-1}\tilde{\mathbf{x}}_i^T \tilde{\mathbf{x}}_k\right] = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \dots & \sigma_{1,k} & \dots & \sigma_{1,p} \\ \sigma_{1,2} & \sigma_2^2 & \dots & \sigma_{2,k} & \dots & \sigma_{2,p} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \sigma_{1,k} & \sigma_{2,k} & \dots & \sigma_k^2 & \dots & \sigma_{k,p} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,p} & \sigma_{2,p} & \dots & \sigma_{k,p} & \dots & \sigma_p^2 \end{bmatrix}$$

- Sample Correlation Matrix:

$$\mathbf{R} = \left[r_{ik} = \frac{\sigma_{ik}}{\sqrt{\sigma_i^2}\sqrt{\sigma_k^2}} \right] = \begin{bmatrix} 1 & r_{1,2} & \cdots & r_{1,k} & \cdots & r_{1,p} \\ r_{1,2} & 1 & \cdots & r_{2,k} & \cdots & r_{2,p} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ r_{1,k} & r_{2,k} & \cdots & 1 & \cdots & r_{k,p} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{1,p} & r_{2,p} & \cdots & r_{k,p} & \cdots & 1 \end{bmatrix}$$

Multivariate Statistics: Multivariate Statistics

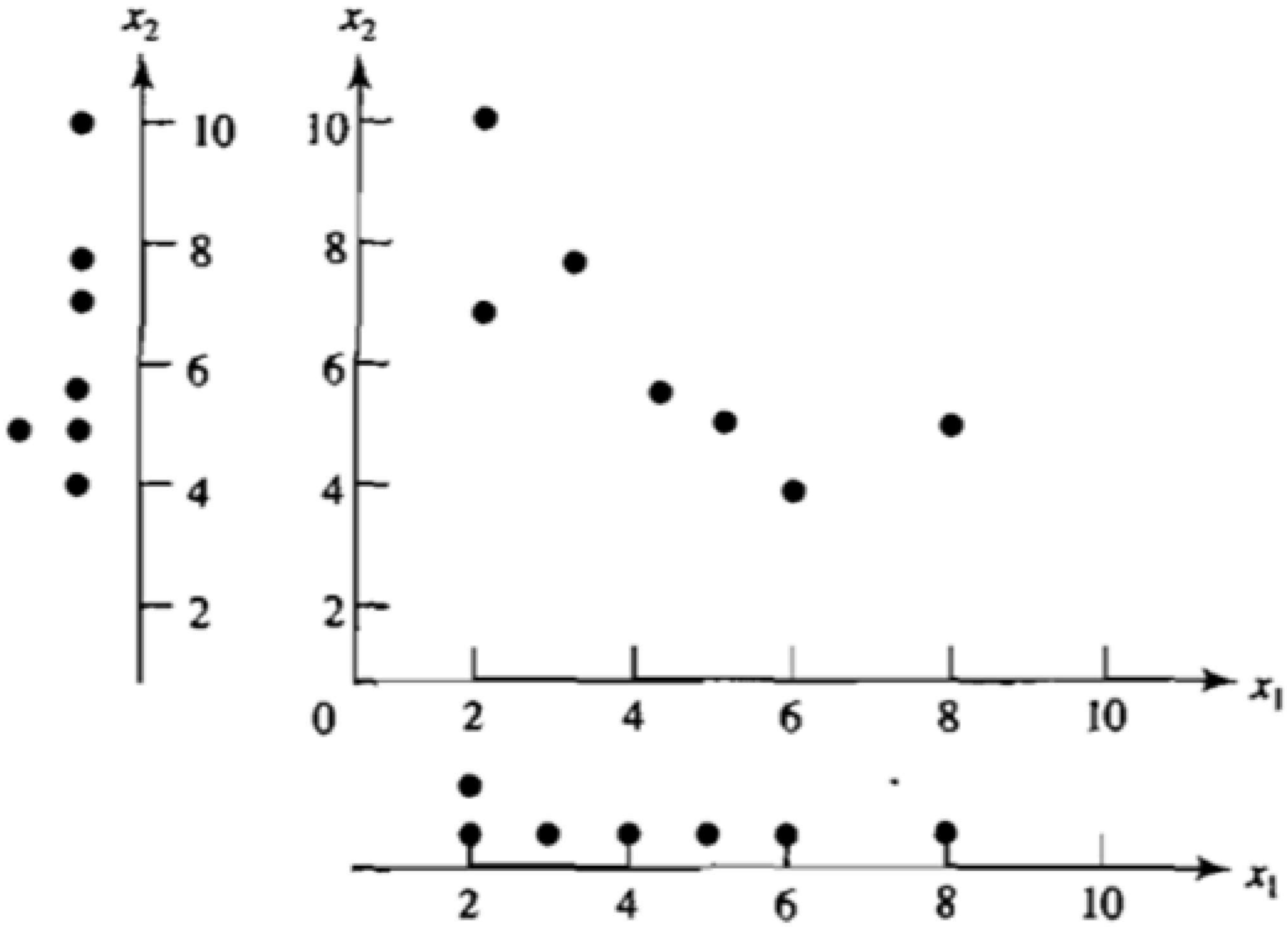


Figure: Scatter plot 2D

Multivariate Statistics: Multivariate Statistics

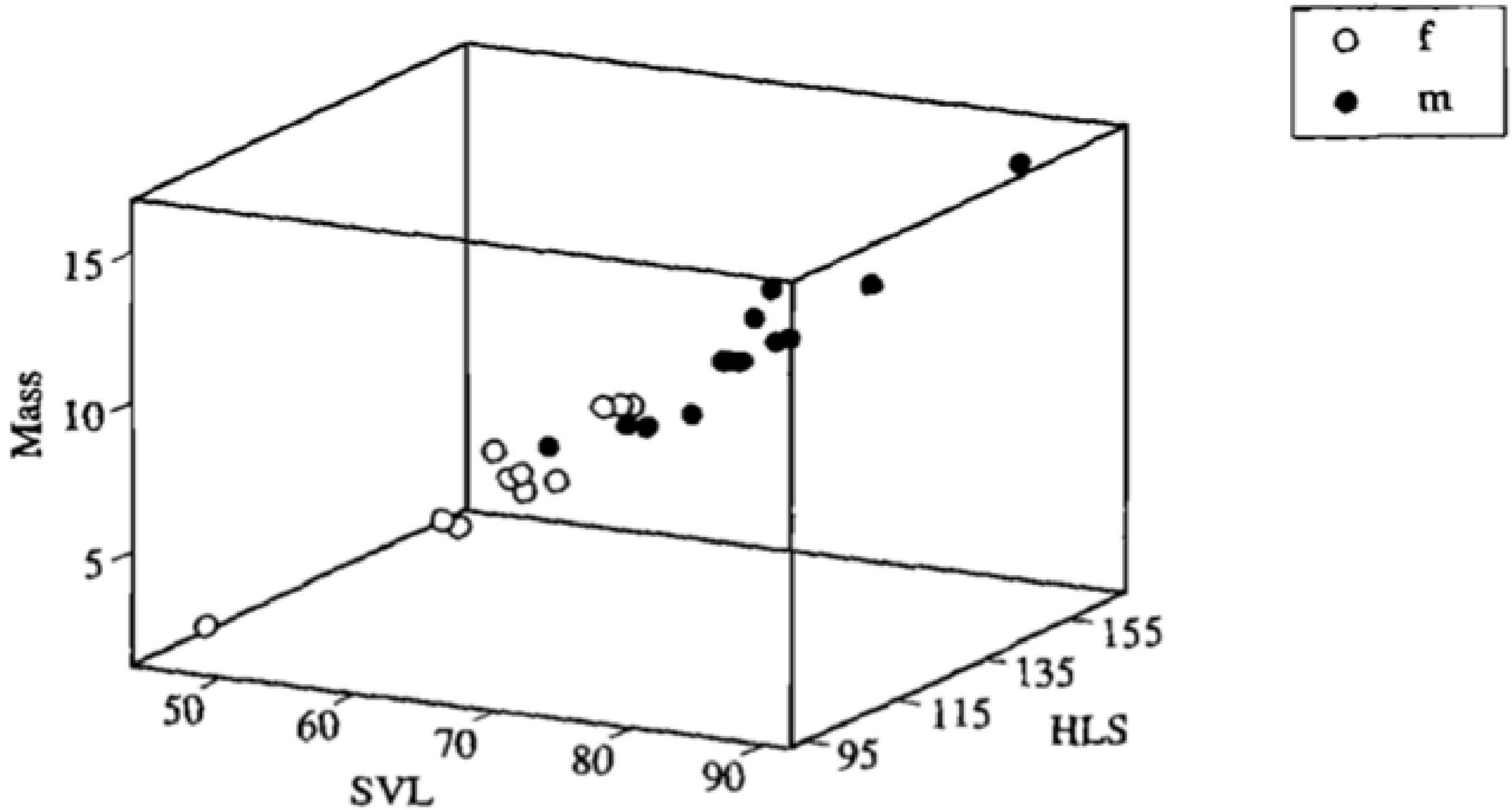


Figure: Scatter plot 3D

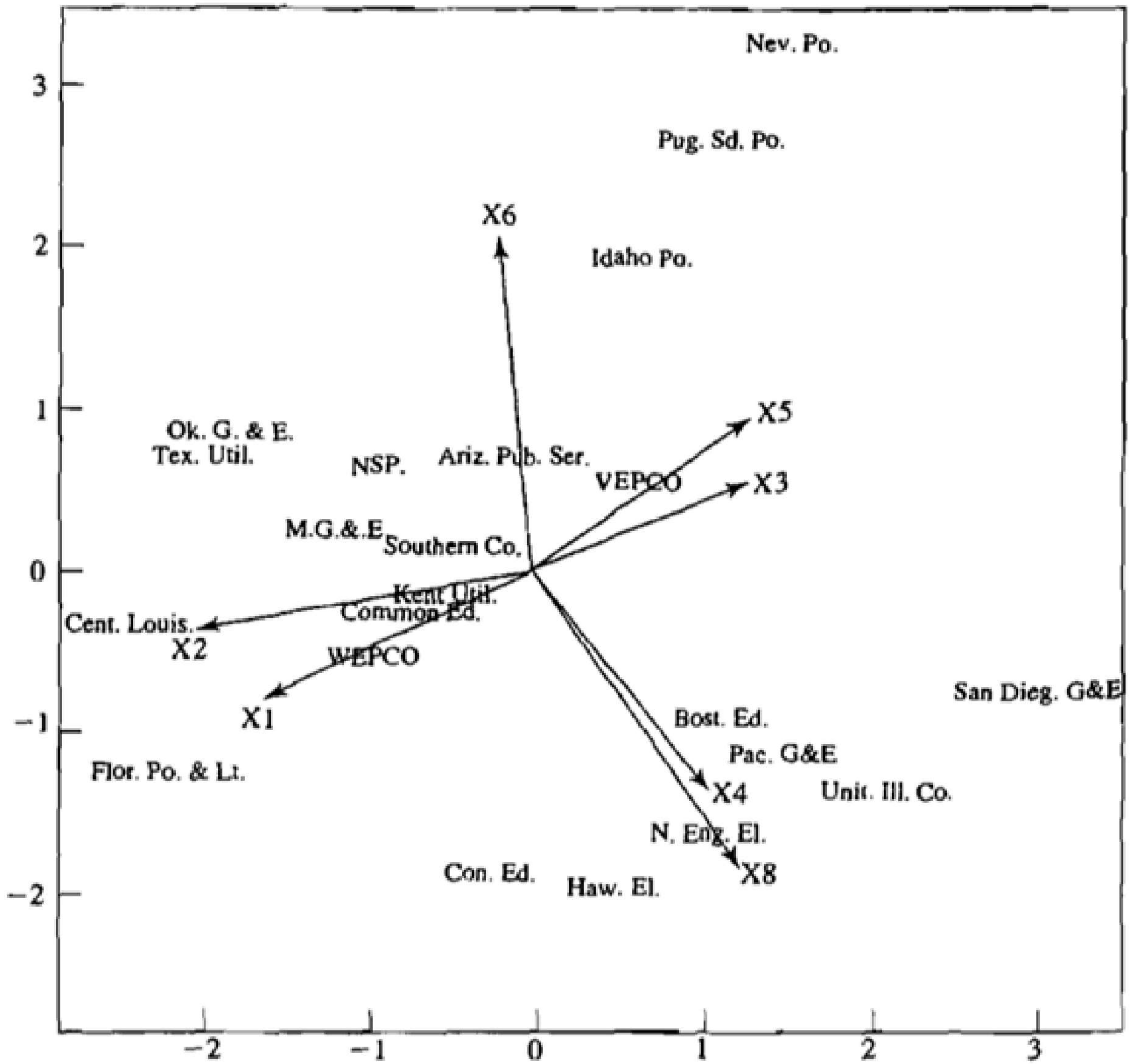


Figure: Biplot

Multivariate Statistics: Multivariate Statistics

Important things to think about with these graphs:

- You can see why the dimensions of the results of usual statistics like the mean, the variance, etc. increase when we use more variables.
- Matrices and vectors are important for manipulating a lot of numbers at the same time without doing various computations.
- From two variables with their own distribution, we can observe some joint behavior when we analyze them together and we can measure and study these relations.
- We can think about methods for solving some problems such as prediction or classification that include some or all of the attributes in the data.
- Because is useful for visualizing data, we might want to find methods to reduce dimensionality and represent multivariate data.

Multivariate Statistics: Distances

Various techniques in statistics and ML & AI use the concept of **distance** between different points P and Q (understood as data vectors or observations). There are different distance measures $d(P, Q)$, but let's delve into the Euclidean.

$$d(P, Q) = \sqrt{(P - Q)^T (P - Q)} = \sqrt{(x_{1k} - x_{1i})^2 + ... + (x_{nk} - x_{ni})^2}$$

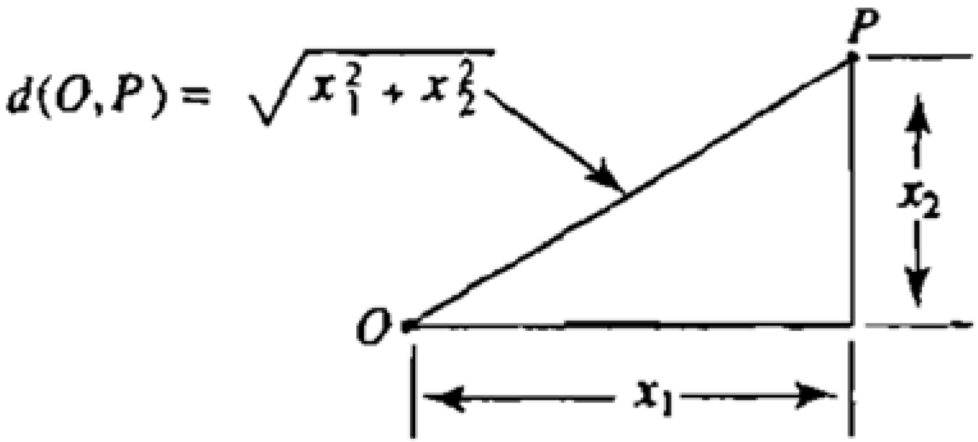


Figure: Euclidean Distance

Multivariate Statistics: Distances

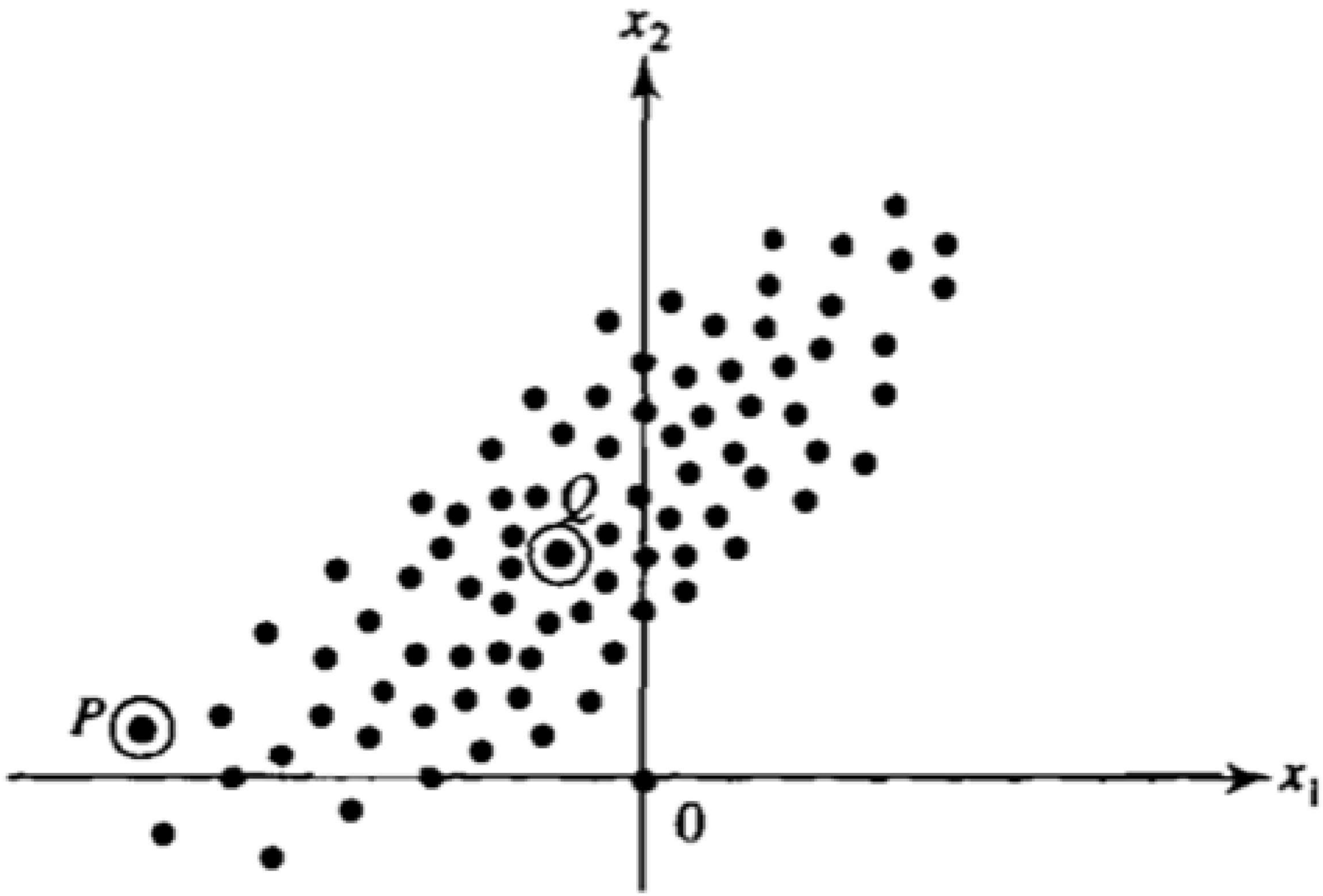


Figure: Cluster of points

Multivariate Statistics: Distances

When we look at the data, we might find that the distance between some observations might tell us something about the underlying distribution or process. Many ML & AI techniques, such as **clustering**, use distances to compute different elements using data observations.

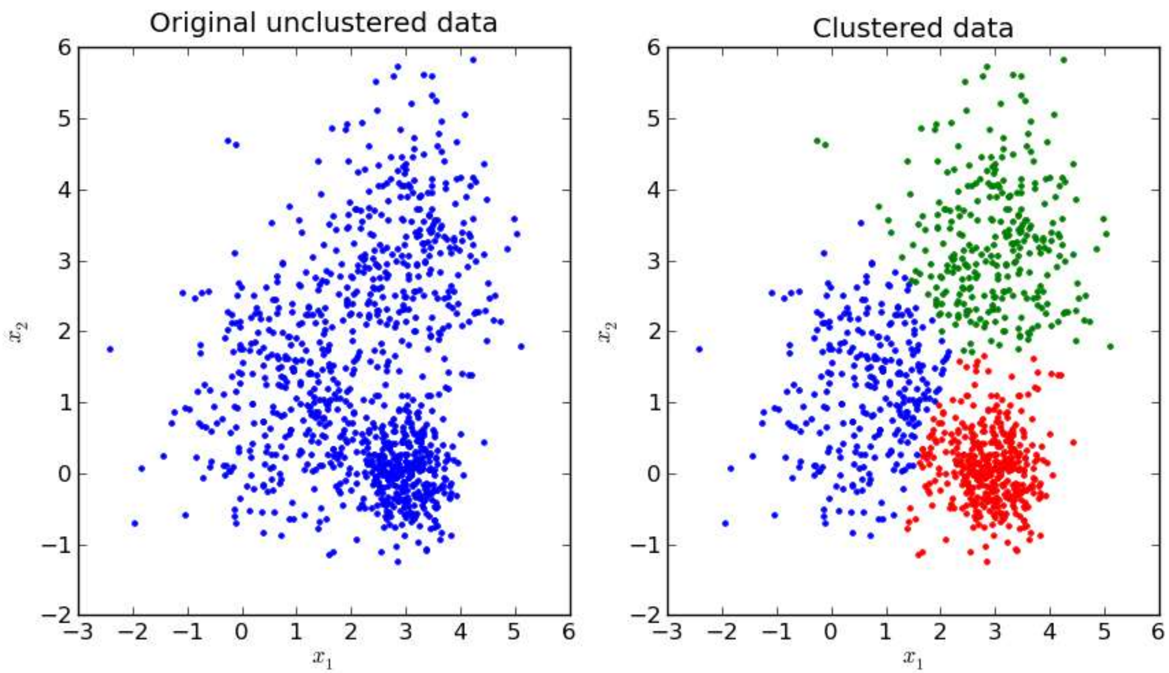


Figure: Clusters of points

Multivariate Statistics: Multivariate Distributions

Understanding random vectors as a group of random variables, it is obvious we can think about distributions with more than one dimension, leading to important distributions such as the **multivariate normal distribution**, in which various ML & AI algorithms are based.

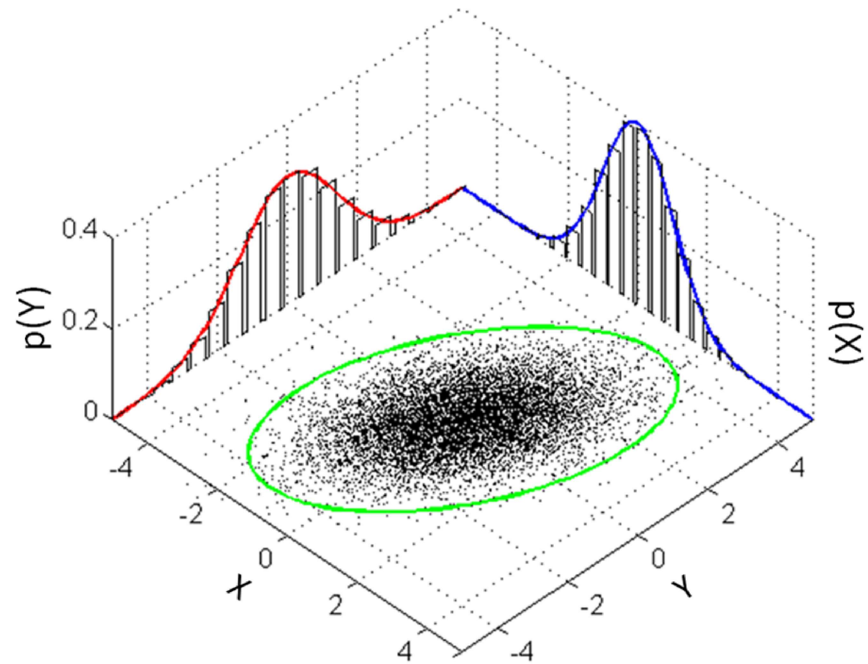


Figure: Bivariate Normal Distribution, with normal marginal distributions.

Multivariate Statistics: Multivariate Distributions

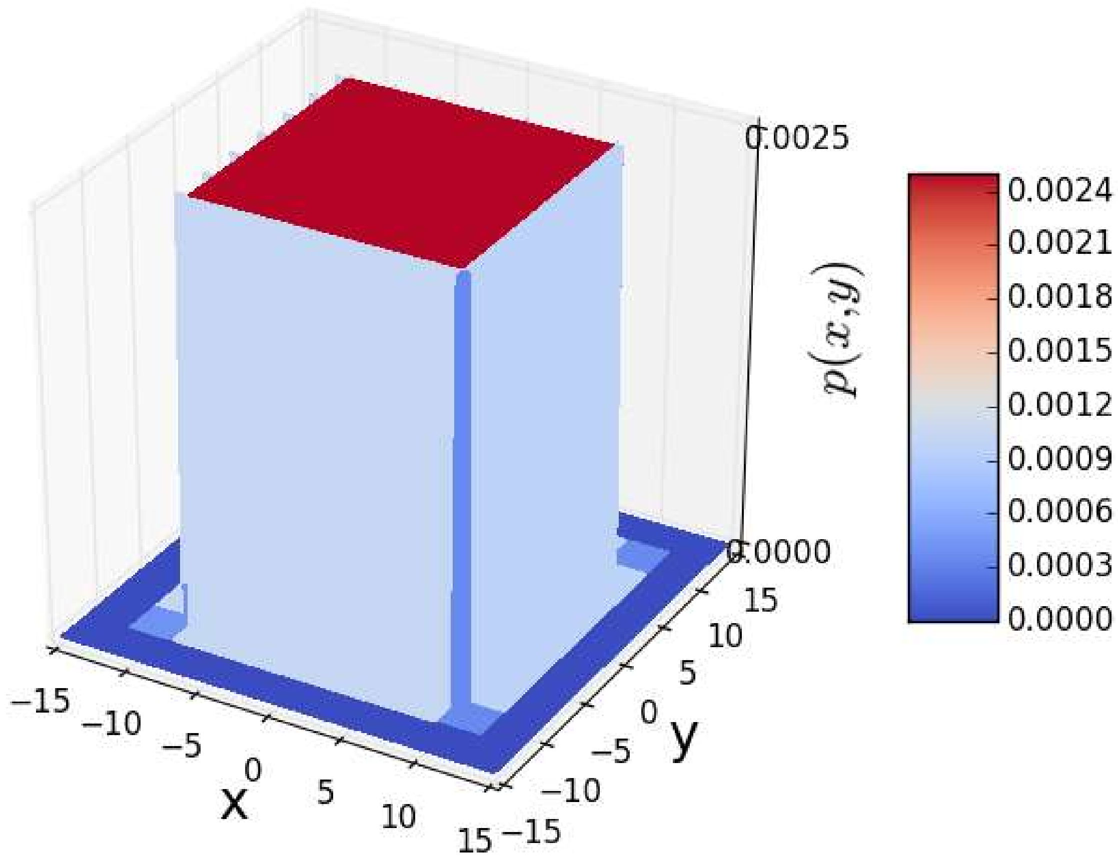


Figure: Multivariate Uniform Distribution

Multivariate Statistics: The Curse of Dimensionality

It seems like using multiple variables gives us a lot of information if we use adequate analyses and methods, and will make our ML & AI algorithms work better and better. Still, problems appear due to the **curse of dimensionality** such as the following:

- **Data Sparsity:** makes it difficult to find significant patterns because the available data cannot populate the space densely.
- **Distances**: distances become less intuitive and, because we consider many dimensions, the difference between farthest and nearest points diminishes, affecting algorithms based on them.
- **Computational Resources**: More variables mean more numbers and more numbers mean we need more computational power to compute different elements.
- **Overfitting**: Algorithms capture more noise than underlying patterns.

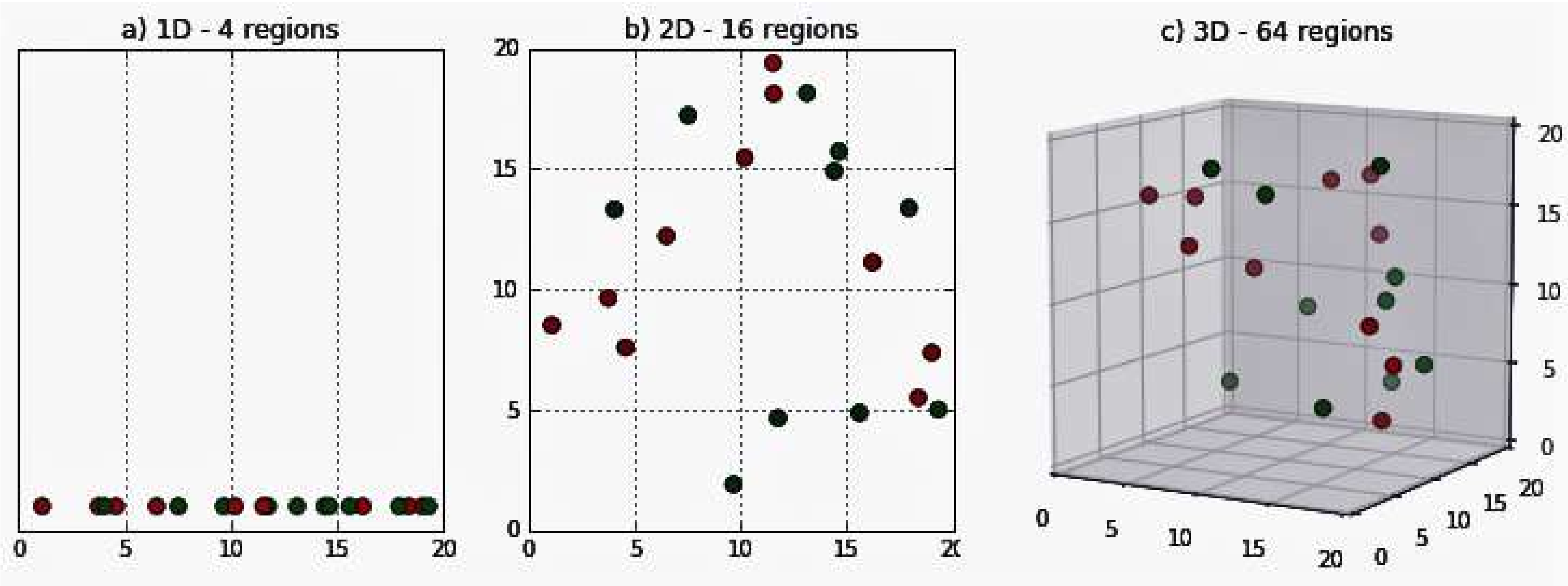


Figure: The regions the algorithm needs to analyze increases with every dimension.

Multivariate Statistics: The Curse of Dimensionality

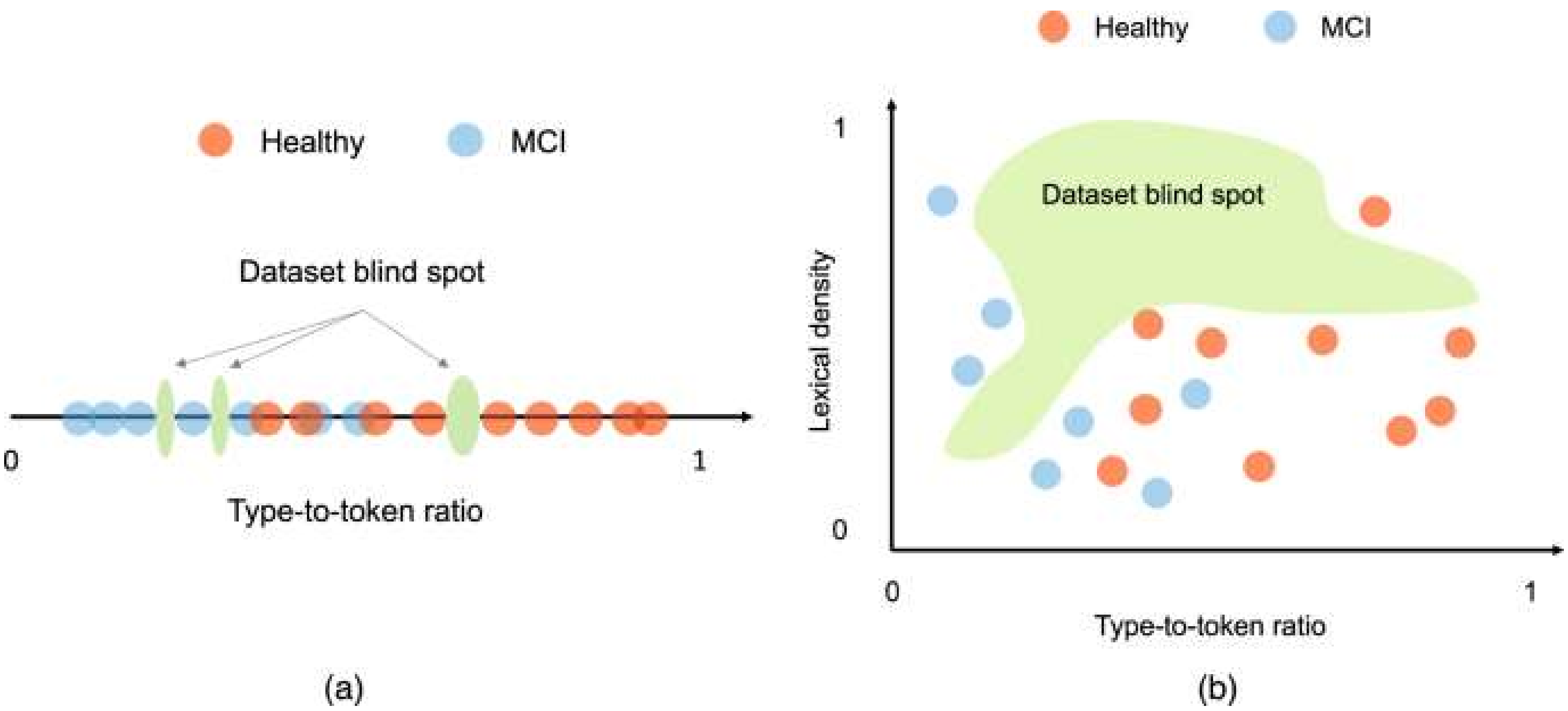


Figure: Datasets shows just a portion of the spaces we want to learn about.

Statistical Learning Theory

Now that we have the basic tools to understand the fundamental statistical and mathematical aspects of modelling with various data, which is foundational for ML & AI, we can go to one of the most interesting questions for newbies: **How do algorithms "learn"?**. The mathematical theory that deals with how algorithms can learn

patterns from data and make predictions is called **statistical learning theory**, and we will take a quick look at some important aspects that characterize ML & AI and differentiates them from fields like statistical inference and econometrics.

- Learning Model and Risk Minimization
- Learning Problems
- Consistency
- Generalization

Statistical Learning Theory: Learning Model & Risk Minimization

The general model for learning is the following:

- A generator (G) of random vectors $\mathbf{x} \in \mathbb{R}^n$, drawn independently from a fixed but unknown probability distribution $F(\mathbf{x})$.
- A supervisor (S) that returns a result y for each vector $\mathbf{x}$, according to the conditional distribution $F(y|\mathbf{x})$, also fixed but unknown. This is the general case, which includes the supervisor using a function such that $y = f(\mathbf{x})$
- A learning machine LM capable of implementing a set of functions $f(\mathbf{x}, \alpha)$ for $\alpha \in \Lambda$, where Λ is the set of parameters. The elements of α might not need to be a vector but any abstract parameter.

Statistical Learning Theory: Learning Model & Risk Minimization

See that I have assumed a supervisor S? This is because this theory was developed for what is called **supervised learning**, but there are other forms:

- **Supervised Learning:** Learning from labeled data to make predictions (e.g., classification and regression).
- **Unsupervised Learning:** Learning from unlabeled data to find structure or patterns (e.g., clustering and dimensionality reduction).
- **Reinforcement Learning:** Learning through rewards and penalties in an environment (e.g., game playing and robotics).

Unsupervised learning is also called **multivariate analysis**. We will focus in supervised methods, given that they are the most famous ones.

Statistical Learning Theory: Learning Model & Risk Minimization

The problem of learning is to select the function from the function set $f(\mathbf{x}, \alpha)$ for $\alpha \in \Lambda$ that best approximates the supervisor's S output y.

The election of the function is based on the training set of l i.i.d observations drawn from $F(y, \mathbf{x}) = F(y|\mathbf{x})F(\mathbf{x})$

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_l, y_l)$$

During the learning process, the LM (learning machine) observes pairs $(\mathbf{x}, y)$ (the training set) and, after training, returns an output $\bar{y}$ for any $\mathbf{x}$. Its goal is for $\bar{y}$ to be as similar as y as possible in some sense.

Statistical Learning Theory: Learning Model & Risk Minimization

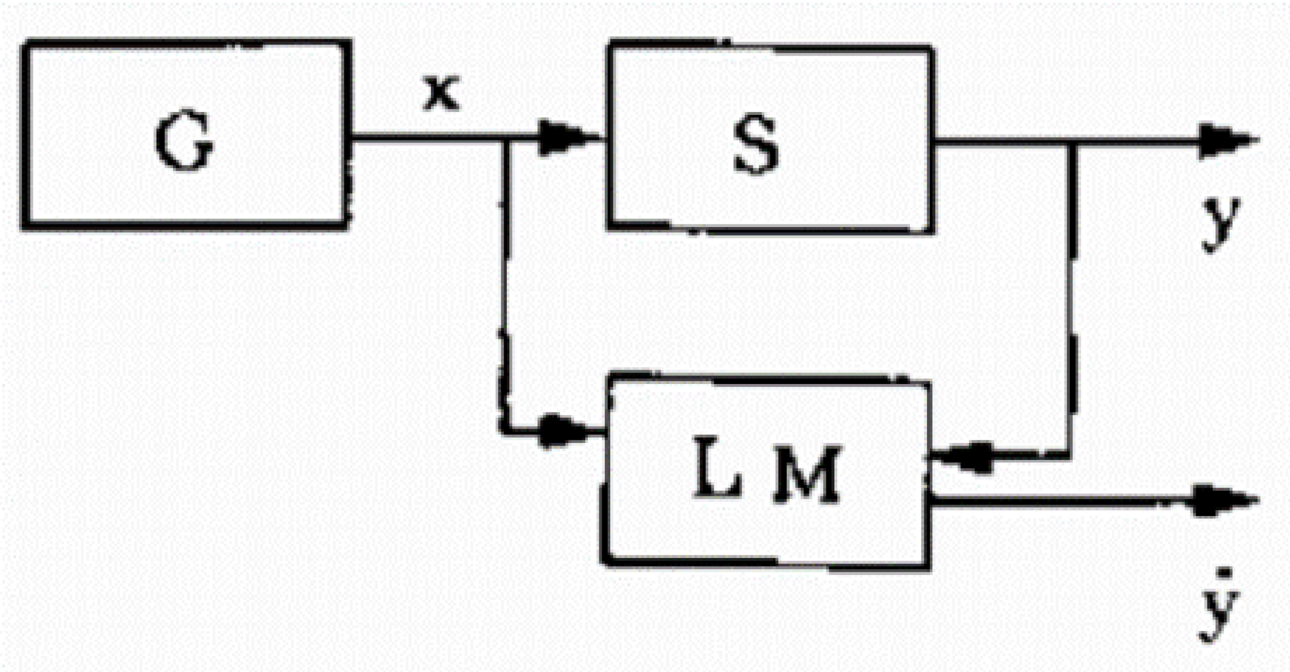


Figure: General Learning Model

Statistical Learning Theory: Learning Model & Risk Minimization

In order to choose the best available approximation to the supervisor's output, one measures the loss or discrepancy $L(y, f(\mathbf{x}, \alpha))$ between S's output y for a given input $\mathbf{x}$ and the response $f(\mathbf{x}, \alpha)$ provided by the LM. Considering the expected risk of this loss, given by the risk functional, the goal is to find a function $f(\mathbf{x}, \alpha_0)$ (from the function class) that minimizes risk $R(\alpha)$, defined as

$$R(\alpha) = \mathbb{E}_{x,y}[L(y, f(\mathbf{x}, \alpha))] = \int L(y, f(\mathbf{x}, \alpha))dF(y, \mathbf{x})$$

whenever $F(y, \mathbf{x})$ is unknown but we just have a set of data $(\mathbf{x}, y)$. Because we deal with a discrete number of data points, we use the empirical risk functional for empirical risk minimization:

$$R_{emp}(\alpha) = (1/l)\sum_{i=1}^{l} Q(z_i, \alpha) \quad \text{where} \quad Q(z_i, \alpha) = L(y_i, f(\mathbf{x}_i, \alpha))$$

Statistical Learning Theory: Learning Problems

This model for learning is very general and encompasses different problems. The most important ones are classification and regression problems (even though we can consider other types).

- **Regression Problems**: Problems in which $y \in \mathbb{R}$ and $f(\mathbf{x}, \alpha)$ is a set of real functions. We predict a continuous output based on input variables (e.g., predicting house prices).
- **Classification Problems**: Problems in which $y \in \{0, ..., m\}$ and $f(\mathbf{x}, \alpha)$ is a set of discrete functions (functions that take a discrete number of values m). We assign data points to predefined categories (e.g., spam vs. non-spam emails).

In economics, most of the times we are concerned with these types, but other fields like non-parametric statistics also consider the "density estimation problem".

Statistical Learning Theory: Consistency & Generalization

Consistency in statistical inference is a property of estimators in which the estimator for some parameter $\bar{\alpha}_n$ converges in probability to the real value α, which can be expressed as

$$\lim_{n\to\infty} P(|\bar{\alpha}_n - \alpha| > \epsilon) = 0, \quad \forall\epsilon > 0 \quad \Leftrightarrow \quad \text{plim}_{n\to\infty}\, \bar{\alpha}_n = \alpha$$

This means that our estimator tends (in probability) to the "real" value when we increase observations. We are interested in when a learning machine that minimizes empirical risk can actually achieve a small value of risk (can generalize) and when it cannot.

Statistical Learning Theory: Consistency & Generalization

We say that a method of empirical risk minimization is consistent for the set of functions $Q(z, \alpha)$ for $\alpha \in \Lambda$ and for the probability distribution $F(z)$ whenever the following convergences hold:

$$R(\alpha_\ell) \xrightarrow{P} \inf_{\alpha \in \Lambda} R(\alpha) \quad \text{and} \quad R_{\text{emp}}(\alpha_\ell) \xrightarrow{P} \inf_{\alpha \in \Lambda} R(\alpha).$$

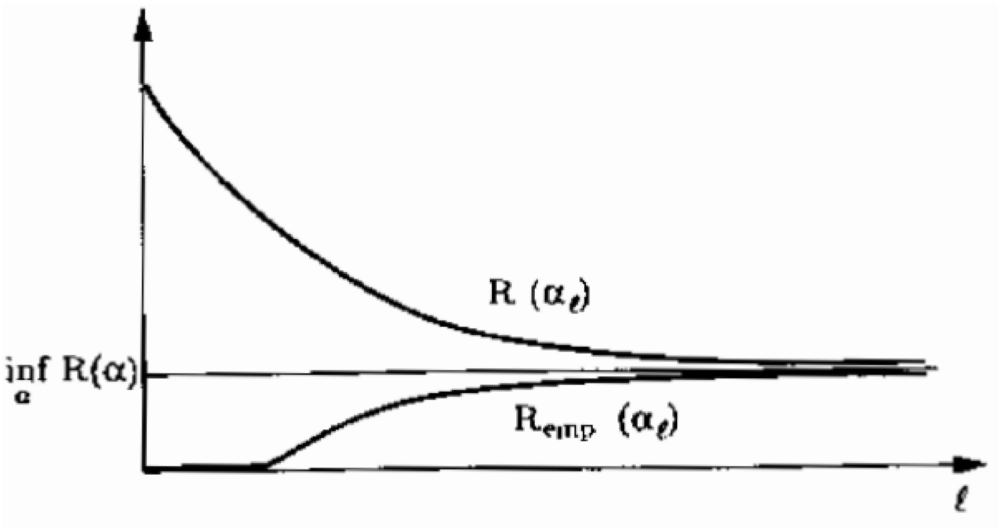


Figure: Convergence of risk and empirical risk to $\inf_\alpha R(\alpha)$

Statistical Learning Theory: Consistency & Generalization

Hence, as our training data set increases, the empirical risk of the learned parameters converge in probability to the true optimal risk (the infimum) and we would be approximating the unknown pattern in an optimal way.

But, why is this important at all? Well, the answer is rooted in the main goal of ML&AI. **We want to generalize to data we have not seen**. Therefore, we just do not care about the **error** we make during training (the **training error**), but also about the error we make on data we have not previously seen, called the **test set**. We would look at this aspect again in the future.

This theory allows us to establish some bounds to the test error and to the expected risk, but this is advanced theory we should not delve into.

Outline