Lecture 2
Security and Ethical Aspects of Data

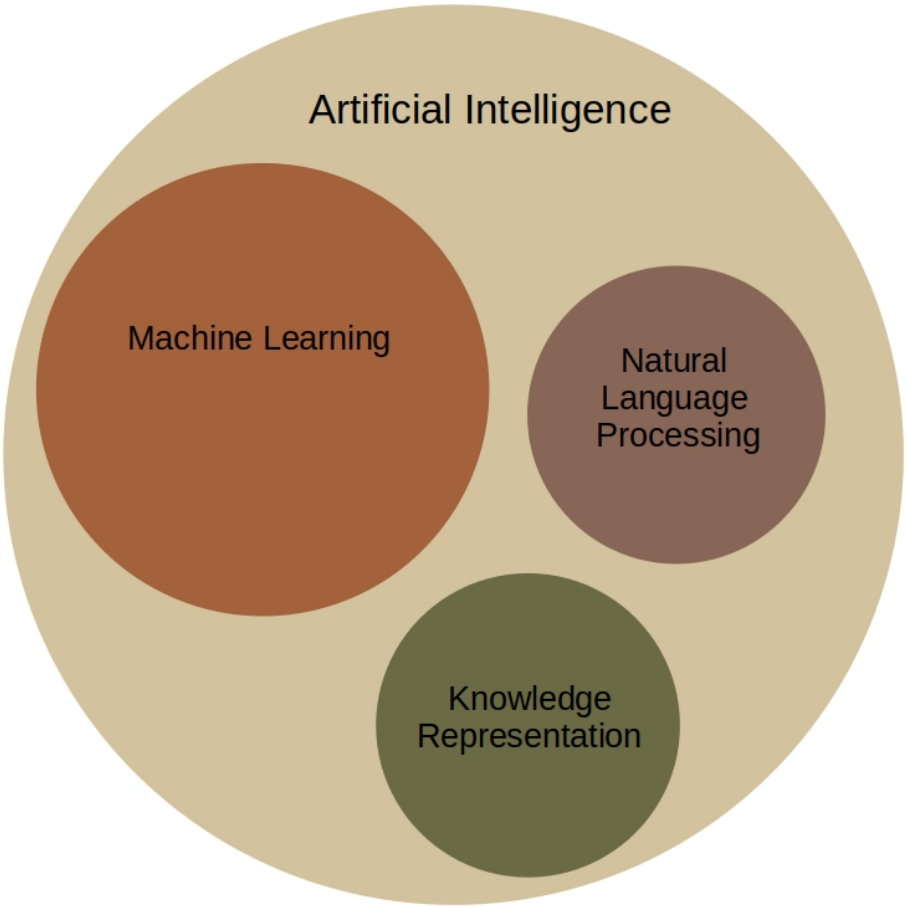
Amaya Nogales Gómez
amaya.nogales-gomez@univ-cotedazur.fr

MSc Data Science & Artificial Intelligence
Université Côte d'Azur

October 18, 2021

Tentative Content

1. Introduction
   - Preliminaries and Machine Learning basics
   - Motivation for ethics in Artificial Intelligence
2. Sources of unfairness
   - Bias in data
   - Algorithmic unfairness: Examples
3. Fairness criteria
   - Types of discrimination
   - Definitions of fairness
4. Algorithms and methods for fair ML
   - Pre-processing methods
   - In-processing methods
   - Post-processing methods
   - Legal and policy perspectives



Supervised Learning

- Classification
- Regression

Unsupervised Learning

- Clustering
- Dimensionality reduction

Reinforcement Learning

Hard-margin SVM: final formulation

The objective function can be replaced by $\Phi(\|\omega\|^\circ)$ for any Φ, increasing in $\mathbb{R}^+$.
Taking $\Phi(t) = \frac{1}{2}t^2$ one obtains an equivalent formulation as a quadratic problem with linear constraints.

$$\min_{\omega, b} \frac{1}{2} \sum_{j=1}^{d} \omega_j^2$$

s.t.

$$y_i(\omega^\top x_i + b) \geq 1 \qquad \forall i = 1, \ldots, n$$
$$\omega \in \mathbb{R}^d$$
$$b \in \mathbb{R}.$$

But...

$$\min_{\omega, b} \frac{1}{2} \sum_{j=1}^{d} \omega_j^2$$

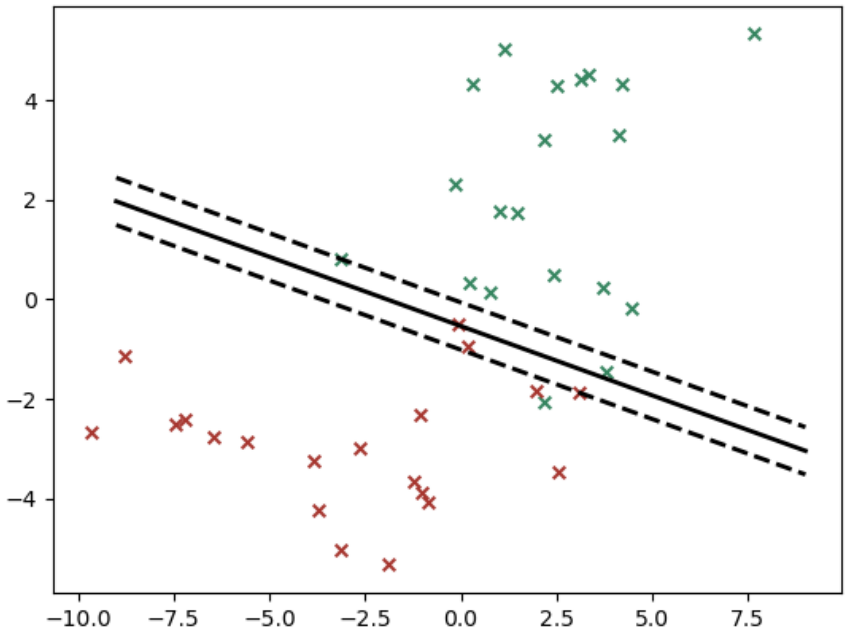
s.t.

$$y_i(\omega^\top x_i + b) \geq 1 \qquad \forall i = 1, \ldots, n$$

$$\omega \in \mathbb{R}^d$$

$$b \in \mathbb{R}.$$

Non-linearly separable data



INFEASIBLE!!

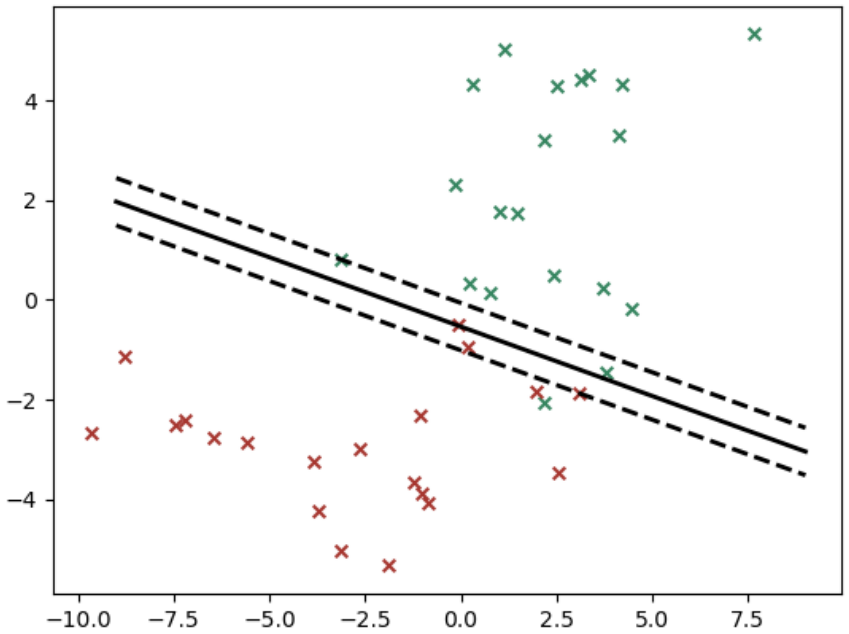
The SVM formulation

$$\min_{\omega,\, b,\, \xi} \; \frac{1}{2}\sum_{j=1}^{d} \omega_j^2 + \frac{C}{n}\sum_{i=1}^{n} \xi_i$$

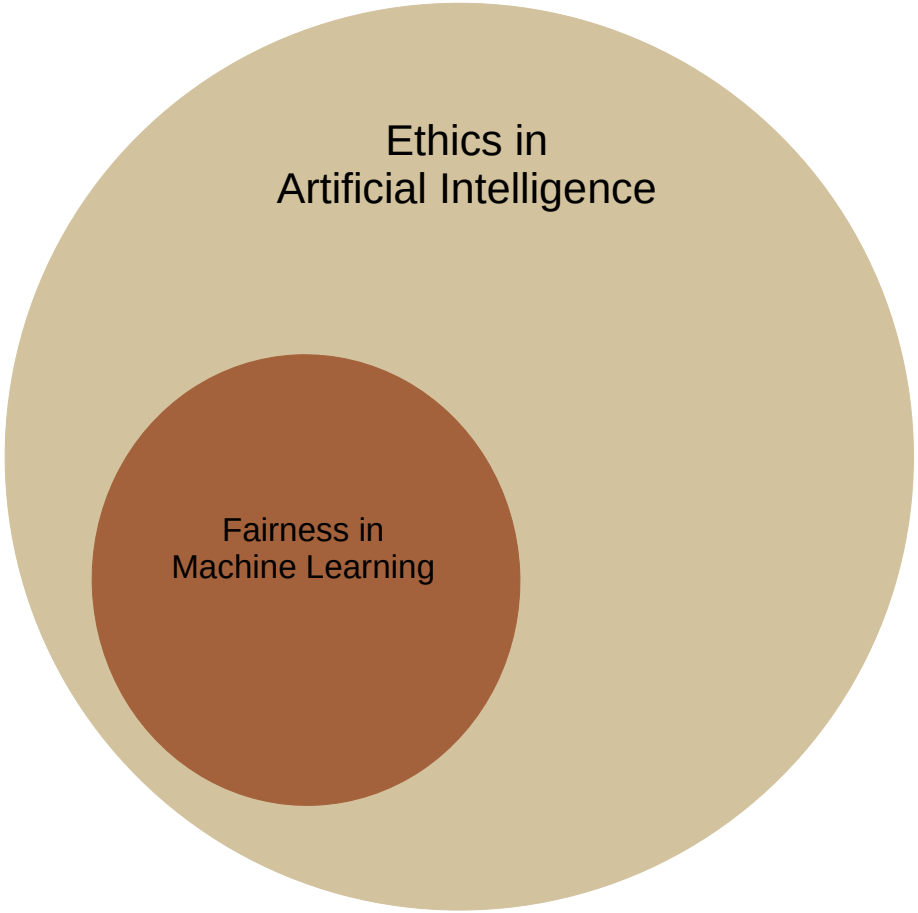
s.t.

$$\begin{aligned} y_i(\omega^\top x_i + b) &\geq 1 - \xi_i \qquad \forall i = 1, \ldots, n \\ \omega &\in \mathbb{R}^d \\ b &\in \mathbb{R} \\ \xi_i &\geq 0 \qquad \forall i = 1, \ldots, n. \end{aligned}$$

Non-linearly separable data



- An object i will be correctly classified if $0 \leq \xi_i < 1$
- Misclassified if $\xi_i > 1$.
- In the case $\xi_i = 1$, we get a tie (objects coincide with the hyperplane).
- $\sum_{i=1}^{n} \xi_i$ is an upper bound of the number of misclassified objects.



Artificial Intelligence

↓

Machine Learning

ethics

↓

fairness

What is fairness?

ability to ensure that different social salient groups are treated similarly

What is fairness?

ability to ensure that different social salient groups are treated similarly

Fairness in Machine Learning

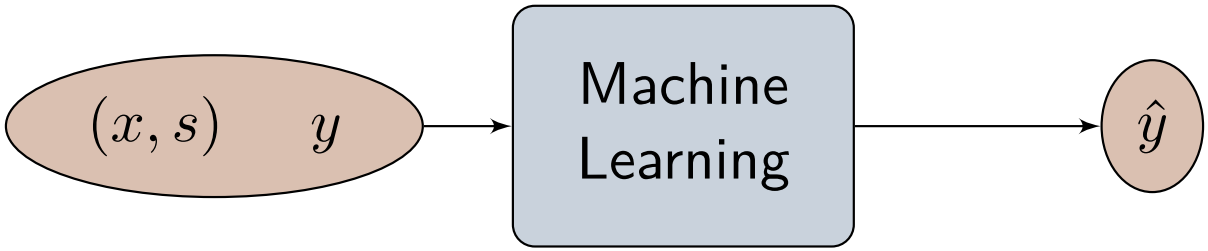
A fair machine learning algorithm implies that its **outcome** should not have a **disproportionately large adverse impact** on a *protected* class of features.

What is fairness?

ability to ensure that different social salient groups are treated similarly

Fairness in Machine Learning

A fair machine learning algorithm implies that its **outcome** should not have a **disproportionately large adverse impact** on a *protected* class of features.



with x =non protected features, s =protected feature, y =true label and $\hat{y}$ =predicted label.

Advertising

Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads

- The authors explore data from a field test of how an algorithm delivered ads promoting job opportunities in the Science, Technology, Engineering and Math (STEM) fields.
- This ad was explicitly intended to be gender-neutral in its delivery.
- **Fewer women saw the ad than men**.

Advertising

Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads

- The authors explore data from a field test of how an algorithm delivered ads promoting job opportunities in the Science, Technology, Engineering and Math (STEM) fields.
- This ad was explicitly intended to be gender-neutral in its delivery.
- **Fewer women saw the ad than men**.
- Why? Because younger women are a prized demographic, i.e., they are more expensive to show ads to.

`https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2852260`

Harm of unfairness in recommender systems

Information Asymmetry

Knowing a piece of information (e.g., a job opportunity) could change one's life.

Matthew effect

Advantaged users, items, or groups get further propagated by recommendations, sometimes not because their good quality but because the recommendation model is dominated by their data.

Echo chambers

Unfair, undiversified exposure to news, tweets, etc. may create echo chamber. Makes it difficult to explore new ideas and opinions different from one's own.

Ethics in the industry

AI Ethics Global Leader: Francesca Rossi

https://www.ibm.com/artificial-intelligence/ethics

IBM AI Ethics Initiative

The purpose of AI is to augment human intelligence

AI should make all of us better at our jobs, and that the benefits of the AI era should touch the many, not just the elite few.

Data and insights belong to their creator

Clients' data is their data, and their insights are their insights. Government data policies should be fair and equitable and prioritize openness.

Technology must be **transparent** and **explainable**

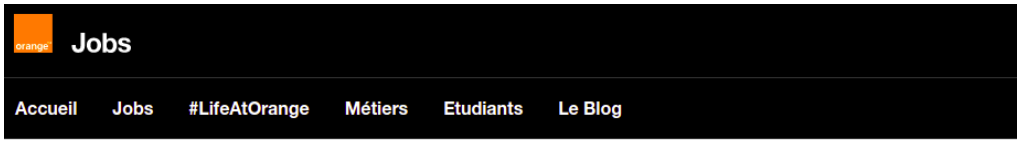
Companies must be clear about who trains their AI systems, what data was used in training and, most importantly, what went into their algorithms' recommendations.

Artificial Intelligence Act: European Parliament

Trustworthy AI by Design:
Responsible, Trustworthy AI requires awareness from all parties involved, from the first line of code.
The way in which we design our technology is shaping the future of our society.

`https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/694212/EPRS_BRI(2021)694212_EN.pdf`

Some "more" real motivation: Orange



POST DOC : Ethics of AI in federated learning F/H

Europe defends responsible AI and published in mid-April 2021 the "AI act" aimed at legislating on high-risk AI. Societal concerns about ***discrimination****, privacy, transparency,* ***explainability*** *and the responsibility of data scientists and companies are growing. At the same time, technology continues to improve the performance and accuracy of models. You will tackle the problem of* ***fairness*** *due to training data and the mechanism of training a model.*

`https://orange.jobs/jobs/offer.do?joid=105070&lang=FR`

Interpretability: Why?

- What are the most influential features towards the decision?
- Is the system "fair" by relying on sensitive attributes such as age and marital status?
- I didn't get the loan; what should I do to get next time I apply?

Interpretability: How?

Transparency	Explainability
inherent/model interpretability	post-hoc/decision interpretability
the level to which a system provides information about its internal and its training	the level to which a system can provide clarifications (explanations) for its decisions/outputs

Explainability

Why was I denied the loan? What should I change in order for my application to be approved?

Counterfactual explanations

You were denied a loan because your annual income was 30000€. If your income had been 45000€ you would have been offered a loan.

Machine Learning Lifecycle

Data Collection

- Before any analysis or learning happens, data must first be collected from the world.
- Data collection involves selecting a population, as well as picking and measuring features and labels to use.
- There exist already many repositories of real-life datasets.

`https://archive.ics.uci.edu/ml/datasets.php`

Machine Learning Lifecycle

Data Preparation

- Depending on the data modality and task, different types of preprocessing may be applied to the dataset before using it.
- Datasets are usually split into a training data used during model development, and testing data used during model evaluation.
- Part of the training data may be further set aside as validation data.

Machine Learning Lifecycle

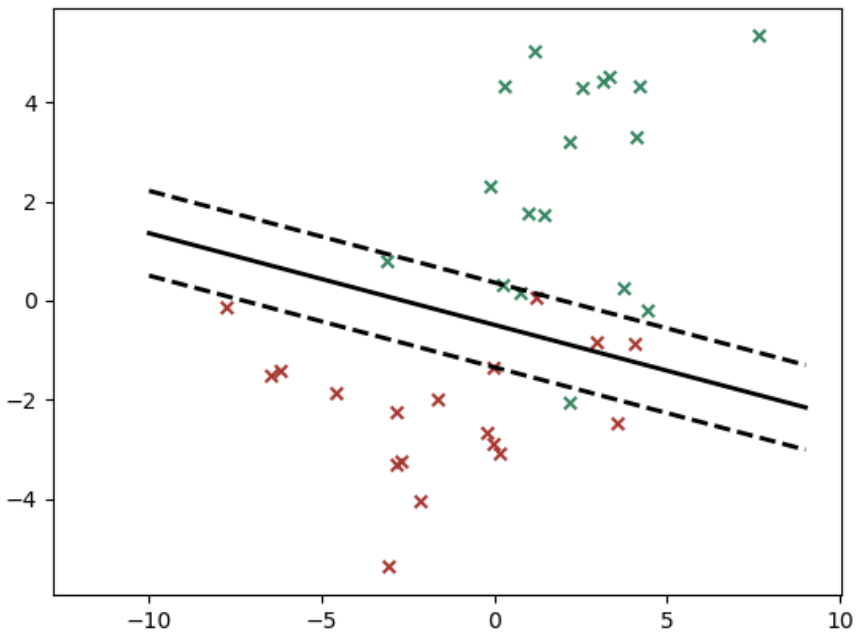
Model Development

- A model is then built using the training data.
- A number of different model types, hyperparameters, and optimization methods may be tested out at this point; usually these different configurations are compared based on their performance on the testing data, and the best one chosen.
- The particular performance metric(s) used in such comparisons are chosen based on the task and data characteristics; common choices are accuracy, false or true positive rates (FPR/TPR).

Common classification criteria

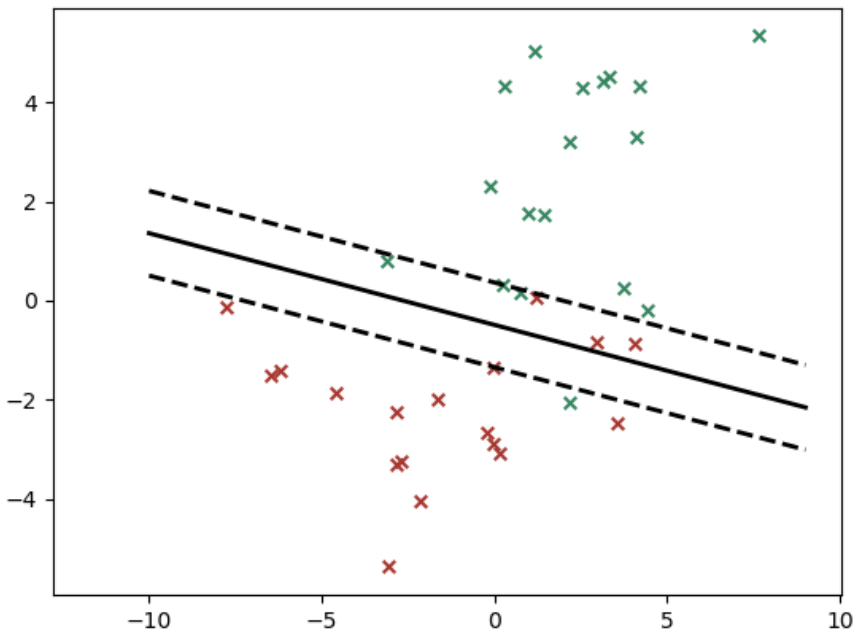
Prediction $\hat{y}$	Label y	Criteria
$+1$	$+1$	True positive rate
-1	$+1$	False negative rate
$+1$	-1	False positive rate
-1	-1	True negative rate

Example: Classification accuracy



- $|\Omega| = 40$
- $\#\{i, y_i = +1\} = 20$
- $\#\{i, y_i = -1\} = 20$
- Accuracy$=\frac{19+17}{40} = 0.9$

Example: Classification accuracy



- $|\Omega| = 40$
- $\#\{i, y_i = +1\} = 20$
- $\#\{i, y_i = -1\} = 20$
- Accuracy$=\frac{19+17}{40} = 0.9$

$\hat{y}$	y	Criteria	
$+1$	$+1$	TPR	$\frac{19}{20}$
-1	$+1$	FNR	$\frac{1}{20}$
$+1$	-1	FPR	$\frac{3}{20}$
-1	-1	TNR	$\frac{17}{20}$

Machine Learning Lifecycle

Model Evaluation

- After the final model and hyperparameters are chosen and the model optimization finished, the final performance of the model on the validation data is reported.
- It is important that the validation data is not used before this step to ensure that the model's performance is a true representation of how it performs on unseen data.
- As in model development, choosing well-suited performance metric(s) is important.

Machine Learning Lifecycle

Model Postprocessing

- Once a model is ready to be used, there are various post-processing steps that may need to be applied.
- For example, if the output of a model performing binary classification is a probability, but the desired output to display to users is a binary answer, there remains a choice of what threshold(s) to use to round the probability to a hard classification.

Machine Learning Lifecycle

Model Deployment

- For a real-world machine learning application, there are many steps that arise when a system is actually deployed.
- For example, a model may need to be changed based on requirements for fairness, or there may be real-time feedback that should be integrated back into the model.

Bias in data

Historical bias

- It arises when there is a misalignment between world as it is and the values or objectives to be encoded and propagated in a model.
- It is a normative concern with the state of the world, and exists even given perfect sampling and feature selection.

Representation bias

- It arises while defining and sampling a development population.
- It occurs when the development population under-represents, and subsequently fails to generalize well, for some part of the use population.

Measurement Bias

- It arises when choosing and measuring features and labels to use; these are often proxies for the desired quantities.
- The chosen set of features and labels may leave out important factors or introduce group or input-dependent noise that leads to differential performance.

Aggregation bias

- It arises during model construction, when distinct populations are inappropriately combined.
- In many applications, the population of interest is heterogeneous and a single model is unlikely to suit all subgroups.

Evaluation bias

- It occurs during model iteration and evaluation.
- It can arise when the testing populations do not equally represent the various parts of the use population.
- Evaluation bias can also arise from the use of performance metrics that are not appropriate for the way in which the model will be used.

Deployment Bias

- It occurs after model deployment, when a system is used or interpreted in inappropriate ways.

Historical Bias

- Historical bias arises even if the data is perfectly measured and sampled.
- It happens if the world as it is or was leads a model to produce outcomes that are not fair.
- Such a system, even if it reflects the world accurately, can still inflict harm on a population.
- Considerations of historical bias often involve evaluating the representational harm (such as reinforcing a stereotype) to a particular identity group.

Example: Image search

In 2018, 5% of Fortune 500 CEOs were women. Should image search results for "CEO" reflect that number? Ultimately, a variety of stakeholders, including affected members of society, should evaluate the particular harms that this result could cause and make a judgment. This decision may be at odds with the available data even if that data is a perfect reflection of the world. Indeed, Google has recently changed their Image Search results for "CEO" to display a higher proportion of women.

Representation bias

- Representation bias occurs when certain parts of the input space are underrepresented.
- Representation bias can arise for several reasons, including:
  1. The sampling methods only reach a portion of the population.
  2. The population of interest has changed or is distinct from the population used during model training.

Example: ImageNet dataset

Geographic diversity in image datasets ImageNet is a widely-used image dataset consisting of 1.2 million labeled images. Approximately $45\%$of the images in ImageNet were taken in the United States, and the majority of the remaining images are from North America or Western Europe. 1% and 2.1% of the images come from China and India, respectively. Shankar et al. (2017) show that the performance of a classifier trained on ImageNet is worse for several categories (such as "bride- groom") on images from under-represented countries such as Pakistan or India versus images from North America and Western Europe.

Shankar, S.; Halpern, Y.; Breck, E.; Atwood, J.; Wilson, J. and Sculley, D. 2017. *No classification without representation: Assessing geodiversity issues in open data sets for the developing world.* arXiv preprint arXiv:1711.08536.

Measurement bias

- Measurement bias occurs when choosing, collection, or computing features and labels to use in a prediction problem.
- After choosing factors to measure, the measurement process itself adds a second layer of noise.
- If the process of choosing and measuring these factors just adds random noise, the model parameters will converge to the those we would expect with the correctly measured quantities.
- Often arises because proxies are generated differently across groups.
- Measurement bias can arise in several ways:
  1. The measurement process varies across groups.
  2. The quality of data varies across groups.
  3. The defined classification task is an oversimplification. Reducing a decision to a single attribute can create a biased proxy label because it only captures a particular aspect of what we really want to measure.

Example: Predictive policing and risk assessments

In predictive policing applications, the proxy variable "arrest" is often used to measure "crime" or some underlying notion of "riskiness". Because minority communities are often more highly policed and have higher arrest rates, there is a different mapping from crime to arrest for people from these communities. Prior arrests and friend/family arrests were two of many differently mismeasured proxy variables used in the recidivism risk prediction tool COMPAS. This was a factor that eventually led to higher false positive rates for black versus white defendants. It is worth noting that such an evaluation is further complicated by the proxy label "rearrest" used to measure "recidivism".

Aggregation bias

- Aggregation bias arises when a one-size-fit-all model is used for groups with different conditional distributions.
- Group membership can be indicative of different backgrounds, cultures or norms, and a given variable can mean something quite different for a person in a different group.
- Aggregation bias can lead to a model that is not optimal for any group, or a model that is fit to the dominant population.

Example: Clinical-aid tools

Diabetes patients have known differences in associated complications across ethnicities. Studies have also suggested that HbA1c levels (widely used to diagnose and monitor diabetes) differ in complex ways across ethnicities and genders. Because these factors have different meanings and importance within different subpopulations, a single model to predict complications is unlikely to be best-suited for any group in the population even if they are equally represented in the training data.

Evaluation bias

- Evaluation bias occurs when the evaluation and/or benchmark data for an algorithm does not represent the target population.
- A model is optimized on its training data, but its quality is often measured on benchmarks (e.g., UCI datasets).
- A misrepresentative benchmark encourages the development of models that only perform well on a subset of the population.
- Evaluation bias can be exacerbated by the particular metrics that are used to report performance.

Example: Commercial facial recognition algorithms

It has been empirically shown a drastically worse performance of commercially-used facial analysis algorithms (performing tasks such as gender- or smiling-detection) on dark-skinned females. Looking at some common facial analysis benchmark datasets, it becomes apparent why such algorithms were considered appropriate for use - just 7.4% and 4.4% of the images in benchmark datasets such as Adience and IJB-A are of dark-skinned female faces. Algorithms that underperform on this slice of the population therefore suffer quite little in their evaluation performance on these benchmarks. The algorithms' underperformance was likely caused by representation bias in the training data, **but the benchmarks failed to discover and penalize this**.
Since this study, other algorithms have been benchmarked on more balanced face datasets, changing the development process to encourage models that perform well across groups.

Deployment bias

- Deployment bias arises when there is a mismatch between the problem a model is intended to solve and the way in which it is actually used.
- This often occurs when a system is built and evaluated as if it were fully autonomous, while in reality, it operates in a complicated sociotechnical system moderated by institutional structures and human decision-makers.

Example: Risk assessment tools

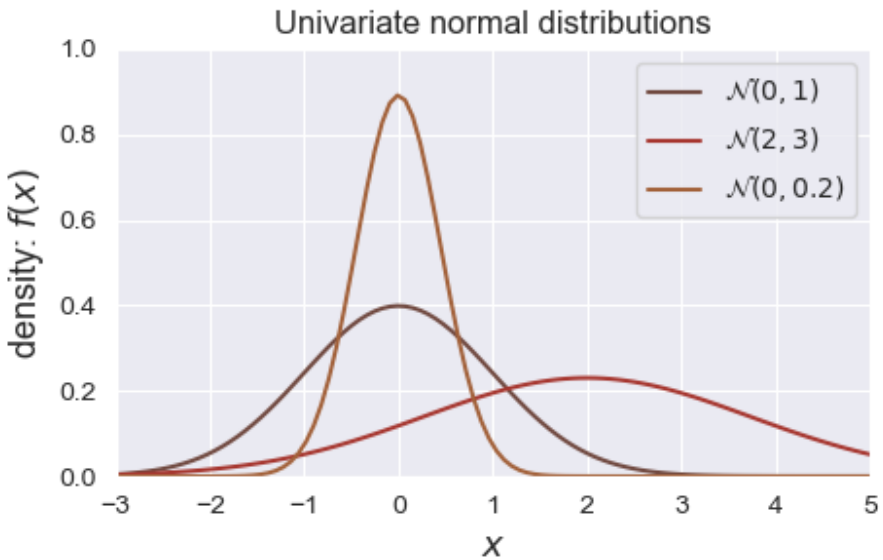
Algorithmic risk assessment tools are models intended to predict a person's likelihood of committing a future crime. In practice, however, these tools may be used in "off-label" ways, such as to help determine the length of a sentence. One of the harmful consequences of risk assessment tools for sentencing, includes the justification of increased incarceration on the basis on personal characteristics.

Gaussian distribution

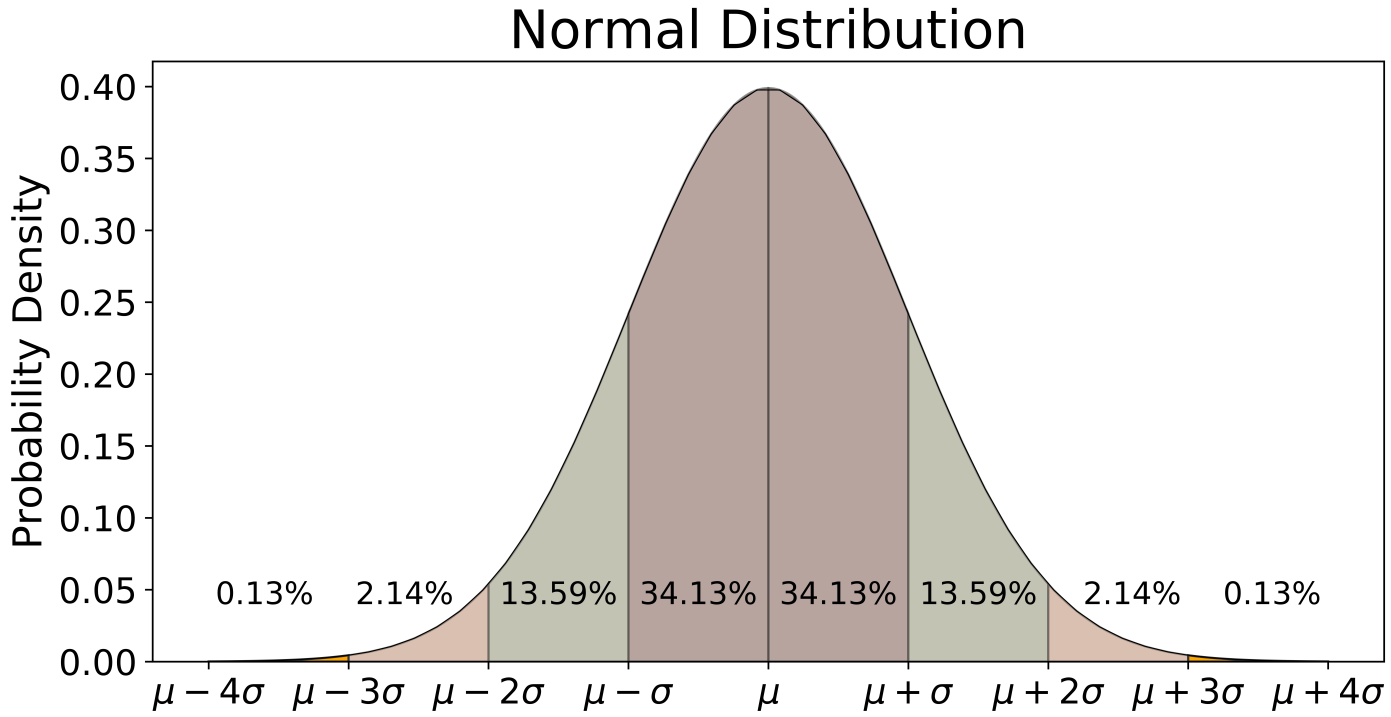
A normal (or Gaussian) distribution is a type of continuous probability distribution for a real-valued random variable. The general form of its probability density function is

$$X \sim N(\mu, \sigma^2)$$

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



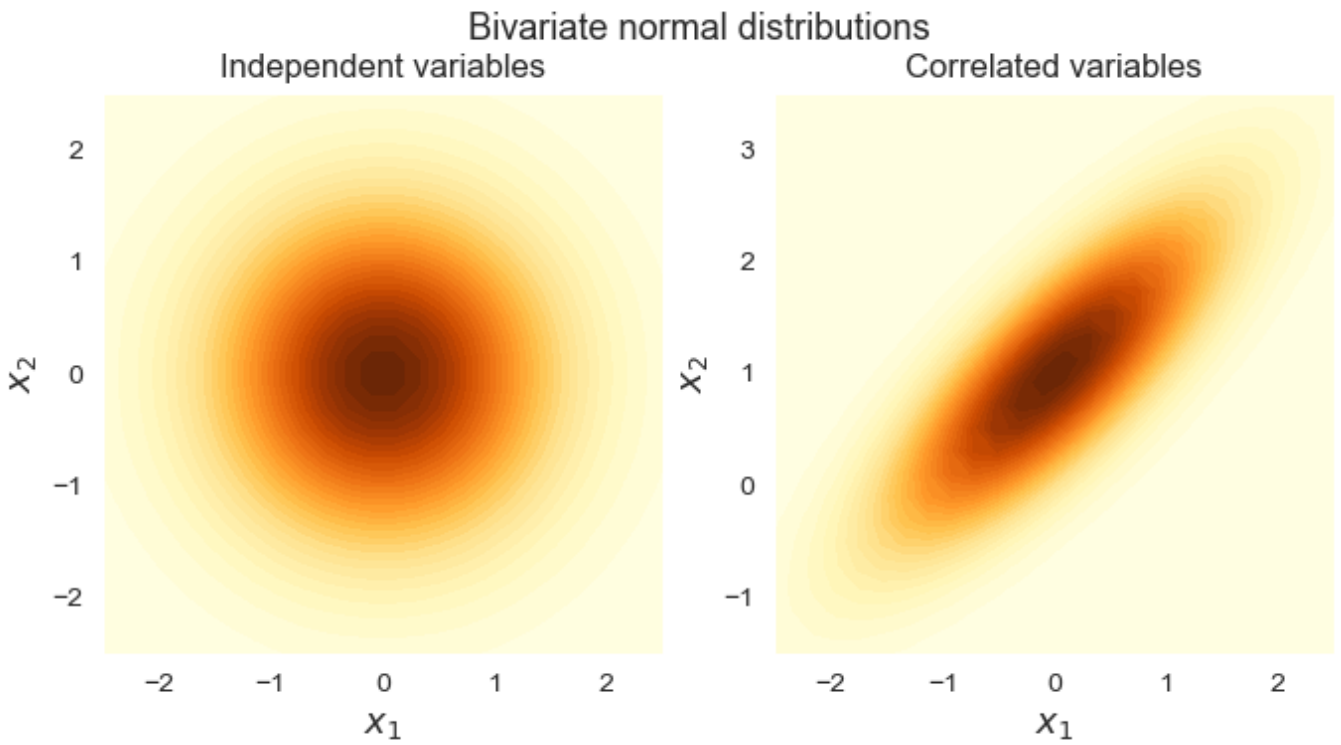
$$P(\mu - n\sigma \leq X \leq \mu + n\sigma) = F(\mu + n\sigma) - F(\mu - n\sigma)$$
$$F(x) = \int_{-\infty}^{x} f(x)dx$$



Multivariate normal distribution

The multivariate normal distribution is a multidimensional generalisation of the one-dimensional normal distribution.
It represents the distribution of a multivariate random variable that is made up of multiple random variables that can be correlated with each other.

$$\begin{aligned} X &\sim N(\mu, \Sigma) \\ f(x) &= \frac{1}{2\pi\sqrt{|\Sigma|}} e^{-\frac{1}{2}\left((x-\mu)^\top \Sigma^{-1}(x-\mu)\right)^2} \end{aligned}$$



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \mu = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

What are some real-world examples of normal distribution?

- Human heights (people of the same gender and age group typically cluster around average with normal distribution)
- IQ scores (the mean is typically 100, SD = 15)
- Marks of students in a class (mean = 60, SD = 20)
- Measure of weight (mean = 80 kg, SD = 10)
- Measure of blood pressure (mean = 120/80, SD = 20)
- Height of trees (measurement in meters; mean = 40 m, SD = 20)

Recommended bibliography

- Y. Li, Y. Ge, and Y. Zhang. 2021. Tutorial on Fairness of Machine Learning in Recommender Systems. SIGIR. https://doi.org/10.1145/3404835.3462814
- http://cs229.stanford.edu/section/gaussians.pdf