Contents

1 Artificial Intelligence, Science and Society

- ECTS Credits: 5

1.1 Course content

Classic approaches in data analysis are use a static procedure for both collecting and processing data. Modern approaches deal with the adaptive procedures which in practice almost always are used.

In this course you will learn how to design systems that adaptively collect and process data in order to make decisions autonomously or in collaboration with humans.

The course applies core principles from machine learning, artificial intelligence and databases to real-world problems in safety, reproducibility, causal reasoning, privacy and fairness.

1.2 Prerequisites

1.2.1 Essential

- Mathematics R1+R2
- Python programming (e.g. IN1900 Introduction to Programming with Scientific Applications).

1.2.2 Recommended

- Elementary knowledge of probability and statistics (STK1000/STK1100)
- Elementary calculus and linear algebra (MAT1100 or MAT1110)

1.3 Learning outcomes

There are two types of learning outcomes. Firstly, those that are the core of the course, and secondly methodologies that are used as part of the course.

Core learning outcomes:

1. Ensuring reproducibility in both science and AI development.
2. Recognising privacy issues and be able to mitigate them using appropriate formalisms.

3. Mitigating issues with potential fairness and discrimination when algorithms are applied at scale.

4. Performing inference when there are causal elements.

5. Developing adaptive experimental design protocols for online and scientific applications.

6. Understanding when it is possible to provide performance guarantees for AI algorithms.

AI learning outcomes:

1. Understanding how to use data for learning, estimation and testing to create reproducible research.

2. Understanding Bayesian inference and decision theory and being able to describe dependencies with graphical models.

3. Understanding neural networks and how to apply stochastic optimisation algorithms.

4. Understanding and using differential privacy as a formalism.

5. Understanding causal inference, interventions and counterfactuals.

6. Understanding the recommendation problem in terms of both modelling and decision making.

1.4 Prerequisites

1.5 Course content

The course is split in 6 modules, which should be taken in sequence.

Module 1. Reproducibility: bootstrapping, Bayesian inference, decision problems, false discovery, confidence bounds. Module 2. Privacy: Databases, k-anonymity, graphical models, differential privacy Module 3. Fairness: Decision diagrams, conditional independence, meritocracy, discrimination. Module 4. The web: Recommendation systems, clustering, latent variable models. Module 5. Causality: Interventions and counterfactuals. Module 6. Adaptive experiment design: Bandit problems, stochastic optimisation, Markov decision processes, dynamic programming.

1.6 Examination

There are 2 projects (formally take-home exams), split into 3 parts each. Each one takes 2-4 hours and is partly done in a tutorial session.

Each question is weighted equally in each home exam, so that by correctly answering the elementary parts of each question, students can be guaranteed a passing grade. Each exam counts for 40% of the score. A final exam is also given by the students. This counts for 20% of the final score.

Criteria for full marks in each part of the exam are the following.

1. Documenting of the work in a way that enables reproduction.
2. Technical correctness of their analysis.
3. Demonstrating that they have understood the assumptions underlying their analysis.
4. Addressing issues of reproducibility in research.
5. Addressing ethical questions where applicable, and if not, clearly explain why they are not.
6. Consulting additional resources beyond the source material with proper citations.

The follow marking guidelines are what one would expect from students attaining each grade.

1.6.1 A

1. Submission of a detailed report from which one can definitely reconstruct their work without referring to their code. There should be no ambiguities in the described methodology. Well-documented code where design decisions are explained.
2. Extensive analysis and discussion. Technical correctness of their analysis. Nearly error-free implementation.
3. The report should detail what models are used and what the assumptions are behind them. The conclusions of the should include appropriate caveats. When the problem includes simple decision making, the optimality metric should be well-defined and justified. Simiarly, when well-defined optimality criteria should given for the experiment

design, when necessary. The design should be (to some degree of approximation, depending on problem complexity) optimal according to this criteria.

4. Appropriate methods to measure reproducibility. Use of cross-validation or hold-out sets to measure performance. Use of an unbiased methodology for algorithm, model or parameter selection. Appropriate reporting of a confidence level (e.g. using bootstrapping) in their analytical results. Relevant assumptions are mentioned when required.

5. When dealing with data relating to humans, privacy and/or fairness should be addressed. A formal definition of privacy and/or should be selected, and the resulting policy should be examined.

6. The report contains some independent thinking, or includes additional resources beyond the source material with proper citations. The students go beyond their way to research material and implement methods not discussed in the course.

1.6.2 B

1. Submission of a report from which one can plausibly reconstruct their work without referring to their code. There should be no major ambiguities in the described methodology.

2. Technical correctness of their analysis, with a good discussion. Possibly minor errors in the implementation.

3. The report should detail what models are used, as well as the optimality criteria, including for the experiment design. The conclusions of the report must contain appropriate caveats.

4. Use of cross-validation or hold-out sets to measure performance. Use of an unbiased methodology for algorithm, model or parameter selection.

5. When dealing with data relating to humans, privacy and/or fairness should be addressed. While an analysis of this issue may not be performed, there is a substantial discussion of the issue that clearly shows understanding by the student.

6. The report contains some independent thinking, or the students mention other methods beyond the source material, with proper citations, but do not further investigate them.

1.6.3 C

1. Submission of a report from which one can partially reconstruct most of their work without referring to their code. There might be some ambiguities in parts of the described methodology.

2. Technical correctness of their analysis, with an adequate discussion. Some errors in a part of the implementation.

3. The report should detail what models are used, as well as the optimality criteria and the choice of experiment design. Analysis caveats are not included.

4. Either use of cross-validation or hold-out sets to measure performance, or use of an unbiased methodology for algorithm, model or parameter selection - but in a possibly inconsistent manner.

5. When dealing with data relating to humans, privacy and/or fairness are addressed superficially.

6. There is little mention of methods beyond the source material or independent thinking.

1.6.4 D

1. Submission of a report from which one can partially reconstruct most of their work without referring to their code. There might be serious ambiguities in parts of the described methodology.

2. Technical correctness of their analysis with limited discussion. Possibly major errors in a part of the implementation.

3. The report should detail what models are used, as well as the optimality criteria. Analysis caveats are not included.

4. Either use of cross-validation or hold-out sets to measure performance, or use of an unbiased methodology for algorithm, model or parameter selection - but in a possibly inconsistent manner.

5. When dealing with data relating to humans, privacy and/or fairness are addressed superficially or not at all.

6. There is little mention of methods beyond the source material or independent thinking.

1.6.5 E

1. Submission of a report from which one can obtain a high-level idea of their work without referring to their code. There might be serious ambiguities in all of the described methodology.
2. Technical correctness of their analysis with very little discussion. Possibly major errors in only a part of the implementation.
3. The report might mention what models are used or the optimality criteria, but not in sufficient detail and caveats are not mentioned.
4. Use of cross-validation or hold-out sets to simultaneously measure performance and optimise hyperparameters, but possibly in a way that introduces some bias.
5. When dealing with data relating to humans, privacy and/or fairness are addressed superficially or not at all.
6. There is no mention of methods beyond the source material or independent thinking.

1.6.6 F

1. The report does not adequately explain their work.
2. There is very little discussion and major parts of the analysis are technically incorrect, or there are errors in the implementation.
3. The models used might be mentioned, but not any other details.
4. There is no effort to ensure reproducibility or robustness.
5. When applicable: Privacy and fairness are not mentioned.
6. There is no mention of methods beyond the source material or independent thinking.

1.7 Motivation

Algorithms from Artificial Intelligence are becoming ever more complicated and are used in manifold ways in today's society: from prosaic applications like web advertising to scientific research. Their indiscriminate use creates many externalities that can be, however, precisely quantified and mitigated against.

The purpose of this course is to familiarise students with societal and scientific effects due to the use of artificial intelligence at scale. It will equip students with all the requisite knowledge to apply state-of-the-art machine learning tools to a problem, while recognising potential pit-falls. The focus of the course is not on explaining a large set of models. It uses three basic types of models for illustration: k nearest-neighbour, neural networks and probabilistic graphical models, with an emphasis on the latter for interpretability and the first for lab work. It is instead on the issues of reproducibility, data colletion and experiment design, privacy, fairness and safety when applying machine learning algorithms. For that reason, we will cover technical topics not typically covered in an AI course: false discovery rates, differential privacy, fairness, causality and risk. Some familiarity with machine learning concepts and artificial intelligence is expected, but not necessary.

2 External resources:

- Programming differential privacy - A book about DP, for programmers. Near and Abuah, 2021.

3 Schedule

3.1 2022

Date	Lecture	Exercise	Paper
27.9	Algorithms Privacy Fairness Reproducibility	Math Test	The randomised response mechanism
04.10	Privacy and anonymity k-anonymity	k-anonymity	Netflix paper
11.10	Differential Privacy Randomised response Laplace Mechanism	Project1	Staircase mechanism
18.10	Approximate DP Gaussian Mechanism		Renyi DP
25.10	Exponential mechanism Privacy amplification		Shuffle privacy Federated learning
01.11	Group fairness Equalised odds		Kleinberg paper
08.11	Balance Calibration	Project2	
15.11	Meritocracy		top-k
22.11	Smoothness		Fairness through Awareness
29.11	Reproducibility Train/Test	5. Repro	GWA
06.12	GWAS		
13.12	Project presentations	ProjectP	
20.12		Project3	

3.1.1 Papers

1. Randomised Response: A Survey Technique for Eliminating Evasive Answer Bias, Warner, 1965.

2. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization, Ohm, 2009.

3. Robust De-anonymization of Large Sparse Datasets. Narayanan and Shmatikov.

4. Calibrating noise to sensitivity in private data analysis. Dwork et al. 2006. (Approximate DP: See also `https://github.com/frankmcsherry/`

blog/blob/master/posts/2017-02-08.md)

5. Our Data, Ourselves: Privacy Via Distributed Noise Generation, Dwork et al. 2006.

6. The staircase mechanism in differential privacy. Geng et al. 2015.

7. Renyi Differential Privacy, Mironov, 2017.

8. Distributed Differential Privacy via Shuffling. Cheu et al, 2019.

9. Federated Naive Bayes under Differential Privacy. Marchioro et al.

10. Big Data's Disparate Impact. Barocas and Selbst, 2016.

11. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, Chouldechova, 2017.

12. Inherent Trade-Offs in the Fair Determination of Risk Scores, Kleinberg et al. 2016.

13. Meritocratic Fairness for Cross-Population Selection, Kearns et al. 2017.

14. Fairness through awareness, Dwork et al. 2011.

15. Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays, Homer et al. 2008.

16. Needles in the Haystack: Identifying Individuals Present in Pooled Genomic Data, Braun et al. 2009.

17. Privacy Preserving GWAS Data Sharing. Uhlerop et al. 2013.

18. A New Analysis of Differential Privacys Generalization Guarantees, Jung et al. 2019.

3.2 2021

	Theory		Practice
24.8	Decision problems Probability and Utility Decision problems in ML	25.8	Expected utility Conditional probability
27.8	Assignment 1		DEADLINE
31.8	Infinite Decision Spaces Stochastic Gradient Tensor Flow Keras	1.9	Experiment pipeline Basic experiment design
7.9	Conditional Probability Conjugate priors	8.9	n-Meteorologists Beta/Bernoulli
10.9	Assignment 2		DEADLINE
14.9	Bayes-optimal decisions Hypothesis testing	15.9	Beta/Bernoulli for hypothesis testing Hierarchical models [Project introduction]
21.9	Non-Conjugate Priors		TFP Graphical Models
28.9	Privacy and anonymity	29.9	SQL, DB tutorial
30.9	Project 1		PRELIMINARY REPORT
5.09	Lab: Randomised Response	6.09	Laplace and Exponential Mechanisms
12.10	Lab: Exponential vs Laplace Mechanism. OpenDP (optionally) (Dirk)	13.10	Fairness Conditional Independence
19.10	Fairness	20.10	Balance Calibration Meritocracy
22.10	Project 1		DEADLINE
26.10	Fairness	27.10	Latent Variable Models Recommender Systems
2.11	Latent Variable Models Recommender Systems Group work	3.11	Lab: Latent Variables with TFP
5.11	Project 2		Deadline #1
9.11	Causality Interventions Counterfactuals	10.11	Group work
16.11	Markov Decision Processes	17.11	Lab: Project work
18.10	Project 2		Deadline #2
23.11	Group work	24.11	Group work
3.12	Project 2		Final Deadline

3.2.1 Module 1: Decision problems, probability and utility.

Reading: Chapter 1

Here the students get familiar with the concept of expected utility. They perform simple exercises in python. We define utility in terms of the classification accuracy for individual decisions and in terms of the generalisation performance in terms of choosing a specific classifier model.

src/decision-problems/expected-utility.py

3.2.2 Module 2: Experiment design and decision analysis

Reading: Sec. 2.4.1, 2.2, 2.1, 2.6

This includes how data will be collected and processed, focusing on automation of the process. I will encourage students to develop an automated pipeline mainly through simulation, where all the variables can be perfectly controlled.

- Optimal decisions in continuous cases: stochastic gradient descent and Bayesian quadrature.

3.2.3 Module 3: Bayesian inference

Reading; Sec 2.3

Introduction to BI through the meteorolgical prediction problem. Dicsussion of simple conjugate priors (Beta, Normal).

Day 1, Part 1

- Graphical model recap (5')
- Conditional probability (5')
- Bayes Theorem (5')
- Marginal distributions (5')
- The n-meteorologists problem (25')

Day 1, Part 2

- Suffficient Statistics / Conjugate priors (15')
- The Beta-Bernoulli conjugate pair (15')
- The Normal-Normal conjugate pair (15')

Day 2, Part 1

- Estimating which classifier is best (45')

– Beta-Bernoulli (15') – Bootstrapping (15')

- Assignment 2 discussion (45')

3.2.4 Module 4: Bayes-optimal Decisions

Reading: Sec. 2.4-2.6, 4.1.3

- Bayesian decisions for models.
- Hypothesis testing: Hierarchical Bayesian models
- Contrast credible intervals with bootstrapping.

3.2.5 Module 5: Non-conjugate priors

Reading: None

Here we will focus on logistic regression as an example, the module will be mainly practical and focus on TF probability.

See `https://arxiv.org/pdf/2001.11819.pdf`

3.2.6 Module 6: Databases and privacy

Reading: Chapter 3.

Introduction to databases, SQL and k-anonymity, consent, and the GDPR. Various mechanisms for DP. Pointers to the opendp.org framework for differential privacy. Comparison of various mechanisms in an ML task.

3.2.7 Module 7: Fairness

Reading: Chapter 4.

Introduction to fairness and condtional independence. Fairness as parity, balance, calibration, meritocracy or smoothness. Measuring conditional independence. Balancing performance with fairness constraints through constrained or penalised optimisation, or Bayesian methods.

3.2.8 Module 8: Latent variable models

Reading: Chapter 5.

Examples: (a) Gaussian mixture model (b) epsilon-contamination model and outliers (c) preferences and attributes in recommendation systems

Practical work with Tensorflow probability, including outlier detection etc.

3.2.9 Module 9: Causality

Reading: Chapter 6.

Confounders, Instrumental variables, Interventions, Counterfactuals. Hands-on: Importance sampling for estimating the impact of decisions. Lab: tf-causalimpact

3.2.10 Module 10: Adaptive experiment design

Reading: Chapter 7.

Here we discuss experiment design in the adaptive setting, where our future experiments depend on data we have not seen yet. Two interesting cases are bandits (e.g. for recommendation systems) and active learning (e.g. for classification).

3.3 2020

19 Aug	L1. Reproducibility, kNN	Christos
20 Aug	A1. Python, scikitlearn, classification, holdouts, overfitting	Dirk
26 Aug	A2. Bootstrapping, XV, project #1 introduction	Dirk
27 Aug	L2. Classification, Decision Problems	Christos
2 Aug	L3. Decisions, inference, optimisation.	Christos
3 Sep	A3. Compare kNN/MLP, discover interesting features	Dirk
9 Sep	L4. Bayesian inference tutorial	Christos
10 Sep	A4. Project Lab	Dirk
16 Sep	L5. Databases, anonymity, privacy	Christos
17 Sep	A5. DB tutorial/distributed computing	Dirk
23 Sep	L6. Differential privacy	Christos
24 Sep	A6. Project DP tutorial: Laplace mechanism	Dirk
25 Sep	Project 1 **Deadline 1**	
30 Sep	L7. Fairness and graphical models	Christos
1 Oct	A7. Production ML: SageMaker/Pipelines	Dirk
7 Oct	L8. Estimating conditional independence	Christos
8 Oct	A8. Project: fairness	Dirk
9 Oct	Project 1 **Deadline 2**	
14 Oct	L9. Recommendation systems [can be skipped?]	Christos
15 Oct	A9. Restful APIs	Dirk
21 Oct	L10. Latent variables and importance sampling	Christos
22 Oct	A10. An example latent variable model?	Dirk
23 Oct	Project 1 **Final Deadline**	
28 Oct	L11. Causality	Christos
29 Oct	A11. Causality lab	Dirk
4 Nov	L12. Interventions and Counterfactuals	Christos
5 Nov	A12. Interventions lab	Dirk
6 Nov	Project 2 **Deadline 1**	
11 Nov	L13. Bandit problems	Christos
12 Nov	A13. Bandit optimisationm lab	Dirk
18 Nov	L14. Experiment design	Christos
19 Nov	A14. Experiment design lab	Dirk
20 Nov	Project 2 **Deadline 2**	
23 Nov	**Exam**	
6 Dec	Project 2 **Final Deadline**	

3.4 2019

21 Aug	L1. Reproducibility, kNN	Christos
22 Aug	L2. Classification, Decision Problems, Project Overview	Christos
29 Aug	A1. Python, scikitlearn, classification, holdouts, overfitting	Dirk
29 Aug	A2. Bootstrapping, XV, project #1 introduction	Dirk
30 Aug	Mini-assigment	
4 Sep	L3. Bayesian inference, Networks, SGD	Christos
5 Sep	L4. Bayesian inference tutorial; neural networks	Christos
12 Sep	A3. Compare kNN/MLP, discover interesting features	Dirk
12 Sep	A4. Project Lab	Dirk
18 Sep	Project 1 1st Deadline	
18 Sep	L5. Databases, anonymity, privacy	Christos
19 Sep	L6. Differential privacy	Christos
26 Sep	A5. DB tutorial/distributed computing	Dirk
26 Sep	A6. Project DP tutorial: Laplace mechanism	Dirk
2 Oct	Project 1 2nd Deadline	
2 Oct	L7. Fairness and graphical models	Christos
3 Oct	L8. Estimating conditional independence	Christos
10 Oct	A7. Production ML: SageMaker/Pipelines	Dirk
10 Oct	A8. Project: fairness	Dirk
16 Oct	Project 1 Final Deadline	
16 Oct	L9. Recommendation systems [can be skipped?]	Christos
17 Oct	L10. Latent variables and importance sampling	Christos
24 Oct	A9. Restful APIs	Dirk
24 Oct	A10. An example latent variable model?	Dirk
30 Oct	L11. Causality	Christos
31 Oct	L12. Interventions and Counterfactuals	Christos
7 Nov	A11. Causality lab	Dirk
7 Oct	A12. Causality lab	Dirk
13 Nov	L13. Bandit problems	Christos
14 Nov	L14. Experiment design	Christos
20 Nov	A13. Experiment design lab	Dirk
21 Nov	A14. Experiment design lab	Dirk
2 Dec	Exam: 9AM Lessart Lesesal A Eilert Sundts hus, A-blokka	
11 Dec	Project 2 Deadline	

1. kNN, Reproducibility
2. Bayesian Inference, Decision Problems, Hypothesis Testing
3. Neural Networks, Stochastic Gradient Descent
4. Databases, k-anonymity, differential privacy

5. Fairness, Graphical models
6. Recommendation systems, latent variables, importance sampling
7. Causality, intereventions, counterfactuals
8. Bandit problems and experiment design
9. Markov decision processes
10. Reinforcement learning

4 Exam subjects

Here are some example questions for the exam. Answers can range from simple one-liners to relatively complex designs. Half of the points will come from 10 1-point questions and the remaining from 2 or 3 2-5-point questions.

4.1 Reproducibility

You are given a set of clinical data $x_1, \ldots, x_T$ with associated labels $y_1, \ldots, y_T$, where $y_t \in \{0, 1\}$ indicates whether a patient has a disease. Each point x_t is decomposable into n features $x_{t,1}, \ldots, x_{t,n}$. Discuss how you can use a classification algorithm that estimates $\hat{P}(y|x)$ from the data in order to discover predictive features, and how you can validate your findings in a reproducibile manner.

4.1.1 Possible answer

(Many approaches are possible, the main thing I want to see is that you can validate your findings)

From a statistical point of view, we want to see the strength of the dependence between an individual feature (or set of features) and the data. The strictest possible test is to see whether or not the labels are completely independent of a feature i given the remaining features, i.e. we want to check that

$$y_t \perp x_{t,i} \mid x_{t,-i} \qquad x_{t,-i} \triangleq x_{t,1}, \ldots, x_{t,i-1}, x_{t,i+1}, x_{t,n}$$

However this check is possibly too strict.

If this is the case, then $P(y_t \mid x_t) = P(y_t \mid x_{t,-i})$. One possible method is to fit the classification model of choice $\mu = \hat{P}(y_t \mid x_t)$ and a sequence of

models $\mu_i = \hat{P}(y_t \mid x_{t,-i})$ on a subset D_1 of the dataset. Consequently, we can measure the likelihood of models on the remaining data D_2, so that we obtain

$$\ell(\mu) = \prod_{t \in D_2} \hat{P}(y_t \mid x_t), \qquad \ell(\mu_i) = \prod_{t \in D_2} \hat{P}(y_t \mid x_{t,-i}).$$

We may then consider all features i with $\ell(\mu_i) < \ell(\mu)$ to be redundant. However, this may not be the case for two reasons:

1. If individually redundant features are correlated, then removing all of them may be difficult. For that reason, we may want to also test the performance of models which remove combinations of featutes.

2. Since probably no feature is completely useless, one reason for the apparent lack of predictive ability of some features maybe the amount of data we have. In the limit, if $y_t \perp x_{t,i} \mid x_{t,-i}$ then our estimators will satisfy $\hat{P}(y_t \mid x_t) = \hat{P}(y_t \mid x_{t,-i})$. However, it is hard to verify this condition when the amount of data is little. Conversely, with a lot of data, even weakly dependent features will not satisfy independence.

4.2 Conditional probability and Bayesian inference

A prosecutor claims that the defendant is guilty because they have found DNA matching them on the scene of the crime. He claims that DNA testing has a false positive rate of one in a million (10^{-6}). While this is indeed evidence for the prosecution, it does not mean that the probability that the defendant is innocent is 10^{-6}. What other information would you need to calculate the probability of the defendant being guilty given the evidence, and how would you incorporate it?

4.2.1 Possible answer

Let us define the fact that the defendant committed a crime as C and the converse as $\neg C$. Let us also denote the event that a test is positive as T. Let us also define the case where the DNA being tested is the one being compared to as M. Then the information we have is

$$\mathbb{P}(T \mid \neg M) = 10^{-6} \tag{1}$$

$$T \text{ is true} \tag{2}$$

In order to predict whether somebody has actually committed the crime given the information, we must calculate $\mathbb{P}(C \mid T)$. This means we must

calculate the following

$$\mathbb{P}(C \mid T) = \mathbb{P}(C \mid M)\,\mathbb{P}(M \mid T) + \mathbb{P}(C \mid \neg M)\,\mathbb{P}(\neg M \mid T) \tag{3}$$
$$= \mathbb{P}(C \mid M)[1 - \mathbb{P}(\neg M \mid T) + \mathbb{P}(C \mid \neg M)\,\mathbb{P}(\neg M \mid T)] \tag{4}$$
$$= \mathbb{P}(C \mid M)[1 - \mathbb{P}(T \mid \neg M)\,\mathbb{P}(\neg M)/\,\mathbb{P}(T) + \mathbb{P}(C \mid \neg M)\,\mathbb{P}(T \mid \neg M)\,\mathbb{P}(\neg M)/\,\mathbb{P}(T)], \quad \mathbb{P}(T) = \mathbb{P}(\tag{5}$$

As you can see, we are missing four important quantities.

- $\mathbb{P}(M)$, the *a priori* probability that this is the defendant's DNA
- $\mathbb{P}(T \mid M)$ the probability of a test being positive if the DNA fragments come from the same person.
- $\mathbb{P}(C \mid M)$, the probability that the defendant committed the crime if the DNA was really theirs.
- $\mathbb{P}(C \mid \neg M)$, the probability that the defendant committed the crime if the DNA was not theirs.

So the false positive rate is far from sufficient evidence for a conviction and must be combined with other evidence.

4.3 Utility

If X is our set of rewards, our utility function is $U : X \to \mathbb{R}$ and we prefer reward a to b (and write $a >^* b$) iff $U(a) > U(b)$, then our preferences are transitive. Give an example of a preference relation $>^*$ among objects so that transitivity can be violated, e.g when $X = \mathbb{R}^2$. In that case, we cannot create a utility function that will satisfy the same relation. Back your example with a thought experiment.

4.3.1 Possible answer

A simple example is when $U : \mathbb{R}^2 \to \mathbb{R}$, with rewards having two attributes. Then we might prefer a to b if $a_1 > b_1 + \epsilon$, but if $|a_1 - b_1|\epsilon$ then we prefer a to b if $a_2 > b_2$. An example is if the first attribute is the IQ score of a job candidate and the second attribute their years of experience. We might prefer a brighter candidate as long as they are clearly much better (as IQ scores are fiddly), otherwise we will prefer the ones that have more experience. As an example, consider three candidates