EXAMINATION

Adaptive Methods for Data-Based Decision Making (IN-STK 5000/9000), 10 ECTS

Informatics, Mathematics
University of Oslo

Date	**2 December 2019**
Results	**15 December.** Solution suggestions on course web page after the exam.
Allowed aids	Any written or printed material
Exam grade limits	A: 80%, B: 70%, C: 50%, D: 40%, E: 30%, F: ¡30%.

Please observe the following:

- Write in *legible* English (unreadable means no points).
- **Motivate your answers, explain your calculations, and clearly state any assumptions made. Insufficient justifications will lose points. Questions where no calculations are required are clearly marked.**
- *State clearly to which question each answer belongs.*

Good Luck!

No.	1	2	3	4	5	6	7	8	$\sum$	Mark
Max	2	1	2	2	1	4	4	4	**20**	
Points										

Question 1 (Reproducibility): (2 Point(s))

A null hypothesis test at significance level p is constructed by using a test statistic $\pi : \mathcal{X} \to [0, 1)$ mapping from the space of possible data to the interval $[0, 1)$, so that the test rejects the null hypothesis whenever $\pi(x) < p$.

1.1. Select the correct answer (no explanation needed):

a) The probability that the test will falsely reject the null hypothesis is p.

b) The probability that the test will falsely reject the alternative hypothesis is p.

c) Given the data x, the probability that the null hypothesis is true is $\pi(x)$.

d) Given the data x, the probability that the null hypothesis is false is $\pi(x)$.

Solution: (a) The probability that the test will falsely reject the null hypothesis is p. □

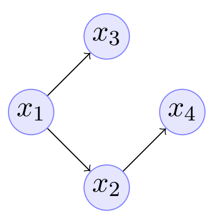
1.2. Being confident about what the test statistic π should mean, you want to use in your own decision problem (e.g. where you want to verify which attributes are significant factors in a classification problem). However, you are still somewhat uncertain about how it would work in practice. Describe a simple synthetic experiment where you can test π's properties.

Solution: The simplest test is to first generate some data from the null hypothesis assumed by the statistic, and then see how often you reject the null hypothesis. In the classification example, you can generate data where the labels are completely random, and see how often the statistic gives a value $< p$ for different features.

Then you need to also generate some data from an alternative hypothesis. For example, you can create a logistic regression model where only some of the features are contributing to the classification, and then see how often the statistic picks those out. □

Question 2 (Graphical models): (1 Point(s))

Factorise the following graphical model of $\mathbf{x} = (x_1, x_2, x_3, x_r)$ with joint distribution $P(\mathbf{x})$. (No explanation needed)



Solution:

$$P(\mathbf{x}) = P(x_1)P(x_2 \mid x_1)P(x_3 \mid x_1)P(x_4 \mid x_2)$$

□

Question 3 (Privacy): (2 Point(s))

Can a k-anonymity algorithm be ϵ-differentially private? Explain why or why not, e.g. through a construction or contradiction.

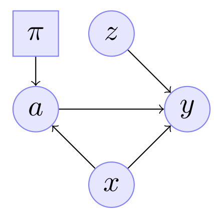
Solution: To have a DP version of k-anonymity we must first we consider all attributes as quasi-identifiers. Secondly, the transformation must be stochastic, and no information about the randomness used must be leaked.

k-anonymity doesn't satisfy the second condition. One can think of a version of k-anonymity where we randomise the quantisation of attributes, so that independent runs of the algorithm will produce different anonymised datasets. However, when we publish the anonymised version we must also explicitly say how we quantised the attributes, so we reveal the randomness and the remaining computation is deterministic.

There are three ways to do a DP version of k-anonymity, but it would not meet the original definition. The first is to use local DP and perform k-anonymity on the DP sample of the data. The second is to perform k-anonymity and then release DP counts of the number of people in each category (remember there are at least k people in each category). The final is to use randomised k-anonymity, but not release the exact quantisation. □

Question 4 (Causality): (2 Point(s))

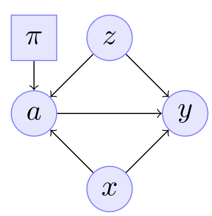
In the diagram below, we have a policy π, an action a, and an outcome variable y.



- Is z a confounding variable, a sufficient covariate, or a nuisance variable?
- Amend the graph so that (x, z) jointly become a sufficient covariate. (No explanation needed)

  Solution: z is a confounder, as it affects the outcome variable, but is not an input to the action. Consequently, the policy maker cannot use knowledge of z to make predictions.

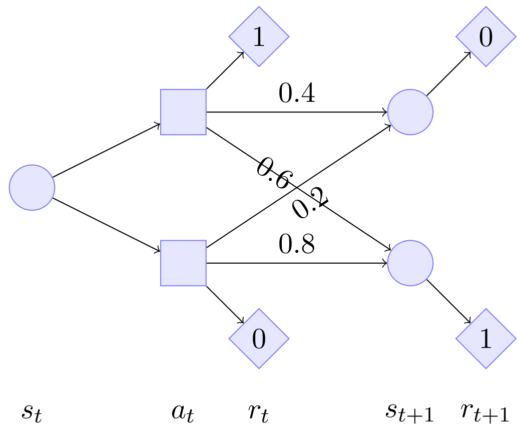
  In the following version of the graph, z links to a. Neither z or x are sufficient by themselves, but jointly they affect y and are both an input to a.



□

Question 5 (Markov decision processes): (1 Point(s))

The following diagram shows a Markov decision process where the circles denote states, the squares denote action notes, and the rhomboids reward nodes. The probability of reaching a state by taking different actions is given on the connecting edges. The process ends after two steps and the utility is defined as the total reward $U_t = r_t + r_{t+1}$. Calculate the expected value of the first state at s_t for the optimal policy, i.e. the one maximising expected utility.



Solution: The top state has value 0 and the bottom has value 1. Hence the top action has value $0.6 + 1$ and the bottom 0.8. Hence the top action is the best, and the optimal policy chooses it. Then the value of the first state is 1.6. □

Question 6 (Probability and expected utility): **(4 Point(s))**

Consider the following basketball matches. Firstly, a match between Helsinki Hippopotami and Gothenburg Geese and secondly a match between the Stockholm Spaniels and the Oslo Occelots.

a) Assume that the outcomes of the matches are independent. You estimate that Helsinki has a 60% chance of beating Gothenburg, so $P(H) = 0.6$ while Stockholm has a 70% chance of winning over Oslo, so $P(S) = 0.7$. There are two bookies, which allow you to bet on the outcome of the matches. The first bookie gives you odds 3/2 that *both Helsinki and Stockholm win*. The second bookie gives you odds 2/1 for a Helsinkini and Stockholm win. (An odds of 3/2 for an event means that you gain 3 NOK for each 2 NOK you bet if the event occurs. So if you bet that *both Helsinki and Stockholm win* with the *first* bookie, and they do, then you get your 1 NOK back and you also gain $3/2 = 1.5$ NOK. Similarly, you can gain 2 NOK if you place the same bet with the second bookie and you win.) Given your assumptions, which of the two bookies gives you the highest expected amount of money for the bet that Helsinki and Stockholm both win? Show your calculations in detail.

Solution: First of all, the chance of the bet comes off is $P(HS) = P(H)P(S) = 0.42$ due to independence. The expected gain in the first case is

$$1.5 \cdot P(HS) - 1 \cdot (1 - P(HS)) = 0.63 - 0.68 = -0.65$$

In the second case, it is

$$2 \cdot P(HS) - 1 \cdot (1 - P(HS)) = 0.84 - 0.68 = 0.16$$

So it is obviously better to take the second bookie's offer if you want to maximise expected money gain. □

b) However, you realise that the Stockholm match is after the Gothenburg match. Due to well-known problems with corruption in sports, there is then a chance that Stockholm would be bribed into losing too, since the match would not be important to them anymore. Specifically, you guess that if Gothenburg loses, then Stockholm has a lower chance of winning: 50%, i.e. that $P(S \mid H) = 0.5$. However, if Gothenburg wins, then you still estimate that Stockholm wins with probability 70%, i.e. that $P(S \mid \neg H) = 0.7$. What is Stockholm's chance of winning without knowing the outcome of the Gothenburg match, i.e. what is $P(S)$ if your assumptions are true?

Solution: We simply write the marginal probability

$$P(S) = P(S \mid H)P(H) + P(S \mid \neg H)P(\neg H) = 0.5 \times 0.6 + 0.7 \times 0.4 = 0.58$$

□

Question 7 (Graphical models and conditional probabilitiy): (4 Point(s))

Many patients arriving at an emergency room, suffer from chest pain. This may indicate acute coronary syndrome (ACS). Patients suffering from ACS that go untreated may die with probability 2% in the next few days. Successful diagnosis results lowers the short-term mortality rate to 0.2%. Consequently, a prompt diagnosis is essential.

Approximately 50% of patients presenting with chest pain turn out to suffer from ACS (either acute myocardial infraction or unstable angina pectoris). Approximately 10% suffer from lung cancer.

Of ACS sufferers in general, ⅔ are smokers and ⅓ non-smokers. Only ¼ of non-ACS sufferers are smokers.

In addition, 90% of lung cancer patients are smokers. Only ¼ of non-cancer patients are smokers.

Assumption 1. *A patient may suffer from none, either or both conditions!*

Assumption 2. *When the smoking history of the patient is known, the development of cancer or ACS are independent.*

One can perform an ECG to test for ACS. An ECG test has *sensitivity* of 66.6% (i.e. it correctly detects ⅔ of all patients that suffer from ACS), and a *specificity* of 75% (i.e. ¼ of patients that do not have ACS, still test positive).

An X-ray can diagnose lung cancer with a sensitivity of 90% and a specificity of 90%.

Assumption 3. *Repeated applications of a test produce the same result for the same patient, i.e. that randomness is only due to patient variability.*

Assumption 4. *The existence of lung cancer* does not *affect the probability that the ECG will be positive. Conversely, the existence of ACS* does not *affect the probability that the X-ray will be positive.*

In this exercise, we only worry about making inferences from different tests results.

a) What does the above description imply about the dependencies between the patient condition, smoking and test results? Draw a graphical model for the above problem, clearly differentiating between observed and hidden random variables, assuming both ECG and X-ray tests are performed.

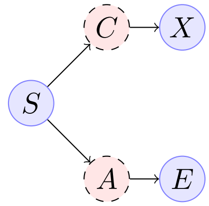
- A: ACS
- C: Lung cancer.
- S: Smoking
- E: Positive ECG result.
- X: Positive X-ray result.

b) What is the probability that the patient suffers from ACS if $S = \texttt{true}$?

c) What is the probability that the patient suffers from ACS if the ECG result is negative?

d) What is the probability that the patient suffers from ACS if the X-ray result is negative and the patient is a smoker?

Solution: **Part 1.** According to our information, only the presence of ACS affects the results of ECG, and only the presence of lung cancer affects the results of the X-ray. Consequently $P(E \mid A, C) = P(E \mid A)$ and $P(X \mid A, C) = P(X \mid C)$. At the same time, smoking is linked with both ACS and cancer.



Part 2. Bayes theorem says, $P(A \mid S) = \frac{P(S|A)P(A)}{P(S)}$, where

$$P(S) = P(S \mid A^{\complement})P(A^{\complement}) + P(S \mid A)P(A) \quad (1)$$
$$= {}^{1}/_{4} \cdot {}^{1}/_{2} + {}^{2}/_{3} \cdot {}^{1}/_{2} = {}^{11}/_{24} \quad (2)$$

is the probability of a smoking patient. Then we plug this in to obtain

$$P(A \mid S) = \frac{{}^{2}/_{6}}{{}^{11}/_{24}} = \frac{8}{11} \approx 72.7\%$$

So smoking is a strong indicator for heart attack.

Part 3. From Bayes theorem, we have that

$$P(A \mid E^{\complement}) = \frac{P(E^{\complement} \mid A)P(A)}{P(E^{\complement})} \quad (3)$$
$$P(E^{\complement}) = P(E^{\complement} \mid A^{\complement})P(A^{\complement}) + P(E^{\complement} \mid A)P(A) \quad (4)$$

We know that $P(E^{\complement} \mid A) = 1 - P(E \mid A) = 1 - {}^{2}/_{3} = {}^{1}/_{3}$, because the sensitivity is ⅓. We also know that $P(E^{\complement} \mid A^{\complement}) = {}^{3}/_{4}$, directly from the definition of specificity. Consequently, $P(E^{\complement}) = {}^{1}/_{3} \cdot {}^{1}/_{2} + {}^{3}/_{4} \cdot {}^{1}/_{2} = {}^{13}/_{24}$. Plugging in, we obtain

$$P(A \mid E^{\complement}) = \frac{{}^{1}/_{3}}{{}^{13}/_{24}} = \frac{8}{13} \approx 61.5\%.$$

So the existence of a positive ECG on its own, is not sufficient evidence for emergency treatment!

Part 4. The X-ray result does not offer any information on the probability of ACS, if we know the patient is a smoker

$$P(A \mid S, X) = P(A \mid S)$$

as the two events A, X are independent given S □

Question 8 (Income statistics): (4 Point(s))

Consider collecting data of individuals so as to calculate income levels over various cross-sections of society. In particular, we collect the following attributes:

- Income x_i.
- 4-digit Postcode y_i.
- Gender z_i.

(a) We wish to publish differentially-private statistics about the income of people in the country. In particular, we wish to publish the average income for two (2) different genders across 10 different regions, i.e. 20 average incomes in total. Design a strategy so that this data is ϵ-DP.

(b) We analyse the data and obtain the following statistics:

Postcode	0xxx	1xxx	2xxx	3xxx	4xxx	5xxx	6xxx	7xxx	8xxx	9xxx	**Average**
Male	306	472	562	223	544	324	66	688	224	485	155
Female	279	380	525	156	558	321	45	703	175	487	120
Average	292	376	543	189	551	322	55	695	209	486	137

Formulate mathematically one (or more) measurable fairness concept which is likely to be violated given the evidence from these statistics.

Solution: First of all, note that the income is potentially an unbounded number. For that reason, we can first transform the income so it lies in a bounded interval $[0, B]$. We then partition the postcodes into regions

$$R_k = \{1 + 1000(k - 1), \dots, 1000k\}.$$

For each interval k and gender $g \in \{0, 1\}$, there are $n_{k,g}$ persons in the database. To calculate the average wage for each category in an $\epsilon_{k,g}$-DP manner, we can use the Laplace mechanism with sensitivity $B/n_{k,g}$. So the average wage output for each category will be

$$a_{k,g} = \frac{1}{n_{k,g}} \sum_{i: y_i \in R_k, z_i = g} x_i + \omega_{k,g}, \qquad \omega_{k,g} \sim \text{Laplace}(B/n_{k,g}\epsilon_{k,g})$$

Due to composition, the total privacy loss of our mechanism is $\sum_{k,g} \epsilon_{k,g}$ so by setting $\epsilon_{k,g} = \epsilon/20$, we guarantee that our mechanism is going to be ϵ-DP.

The statistics show that the expected income for the two genders is different. So, in this case, even though we do not have a decision variable to take into account, we can formulate the notion of parity:

$$\mathbb{P}(x \mid z) = \mathbb{P}(x).$$

We also see that the average income varies across neighbourhoods. This also points to a non-equitable distribution of income across the population. This can be written again in terms of parity:

$$\mathbb{P}(x \mid y) = \mathbb{P}(x).$$

□