CHALMERS
EXAMINATION / TENTAMEN

Course code/kurskod	Course name/kursnamn			
DAT246	Emperical Software Engineering			
Anonymous code Anonym kod		Examination date Tentamensdatum	Number of pages Antal blad	Grade Betyg
DAT246-072-HWC		22-10-24	Tes	5

* I confirm that I've no mobile or other similar electronic equipment available during the examination.
Jag intygar att jag inte har mobiltelefon eller annan liknande elektronisk utrustning tillgänglig under eximinationen.

Solved task Behandlade uppgifter No/nr		Points per task Poäng på uppgiften	Observe: Areas with bold contour are to completed by the teacher. Anmärkning: Rutor inom bred kontur ifylles av lärare.
1	X	5	
2	X	9.5	
3	X	3	
4	X	8	
5	X	6	
6	X	6	
7	X	3	
8	X	6	
9	X	3,5	
10	X	13,5	
11			
12			
13			
14			
15			
16			
17			
Bonus poäng			
Total examination points Summa poäng		73,5	

DAT246-072-HWC

Empirical Software Engineering

Write your answers directly on these pages; there's always a risk that loose papers disappear. Use the back also if possible.

On November 8 at 10.00–11.00 you are welcome to room 6217 in the EDIT house (Johanneberg), with questions about the grading. Before the meeting you *must* send Richard an email clearly pointing out where you think the error is, what you wrote, and why you believe the grading was not correct. If I don't receive such an email before 10.00 on November 8, then I will not meet with you.

Sivajeet Chand will be at the written exam twice (first time after approximately one hour).

Ce que nous connaissons est peu de chose, ce que nous ignorons est immense

— Pierre-Simon Laplace

Grade 3: 47 points; ~50%
Grade 4: 65 points; ~70%
Grade 5: 84 points; ~90%
Maximum: 93 points

DAT246-072-Hwc

5p

Question 1 :

(**8p**) In *your* opinion, which **steps** are compulsory when conducting Bayesian data analysis? Please **explain** what steps one take when designing models, so that we ultimately can place *some* confidence in the results.

You can either draw a flowchart and explain each step, or write a numbered list explaining each step. (It's ok to write on the backside, if they haven't printed on the backside again...)

γ - Null model: an initial model to start with.

γ - Prior predictive checks: Do we have sane priors.

ʃ - Diagnostics: How good is this model doing?

γ - Posterior predictive checks: Does our model do a good job of capturing the regular features of the data.

~~ʃ~~ - Improve on our model, and perhaps come up with ~~different models~~ as described in the previous steps.

ʃ - Comparisons: Perform comparisons on our models to see how they fare against each other.

It's worth pointing out that it's an iterative process, ʃ

! ~~when~~ we're never truly done, we just stop when we can no longer find reasonable improvements.

No models are perfect, but some are good enough!

Yes!

likelihood?

data prep?

inferential stats?

9,5

DAT246-07E-HWC

Question 2 :

(**12p**) Underfitting and overfitting are two concepts not always stressed a lot with black-box machine learning approaches. In this course, however, you've probably heard me talk about these concepts a hundred times...

Multilevel models can be one way to handle overfitting, i.e., employing partial pooling. Please design (write down) three models. The **first one** should use complete pooling, the **second one** should employ no pooling, and the **final one** should use partial pooling. (Remember to use math notation!) (9p)

Explain the **different behaviors** of each model. (3p)

No pooling (over fitting): The model only describes the mean of each category separately. No information is shared among the categories, each one is separate.

$$y_i \sim N(\mu_i, \sigma)$$
$$\mu_i = \alpha_{cat[i]}$$
$$\alpha_{cat[i]} \sim N(0,1) \text{ for } i \text{ in } 1,\dots,N$$
$$\sigma \sim Exp(1)$$

-1

Complete pooling (underfitting): One grand mean for the entire population. No categories or clusters are considered.

$$y_i \sim N(\mu_i, \sigma)$$
$$\mu_i = \alpha$$
$$\alpha \sim N(0,1)$$
$$\sigma \sim exp(1)$$

A prime example of underfitting.

-0.5

Partial pooling: Varying intercept: Here each cluster informs its own mean but also shares information with the other categories. This is an example of partially pooled mean intercept between the categories, using the hyperparameter $\bar{\alpha}$ with its hyperprior.

$$y_i \sim N(\mu_i, \sigma)$$
$$\mu_i = \alpha_{cat[i]}$$
$$\alpha_{cat[i]} \sim N(\bar{\alpha}, 1)?$$
$$\sigma \sim exp(1)$$
$$\bar{\alpha} \sim N(0,1)$$

-1

DAT246-072-HWC 3

Question 3 :

(**5p**) What is **epistemological justification** and how does it differ from **ontological justification**, when we design models and choose likelihoods/priors? Please provide **an example** where you **list** epistemological and ontological reasons for a likelihood.

Epistemological just. is one based on Information theory using the right (maxent) distro. to model the data.

Ex: A count where mean = variance, the Poisson is used.

Ontological just. is one based on our Scientific and natural knowledge of physics, biology ... of the underlying data generation process.

Ex: A natural pheromena with noise and fluctuations is best modelled with a Normal distribution.

Ontological, but epist?

DAT246-072-HWC

8

Question 4 :

(**8p**) **List and explain** the four main benefits of using multilevel models.

* Improved estimations with repeated sampling:
When data is collected across days or weeks or from different settings, situations. We want knowledge shared. Single Level models run the risk of overfitting or underfitting in that case

* Improved estimations when imbalanced sampling:
We want knowledge shared between clusters of different sample sizes, to compensate for difference in size & to prevent one cluster from dominating.

* Estimations of variance:
MLM's allow us to model & measure uncertainty & variance of the different cluster separately while also contributing to the overall variance of the data.

* Avoid averaging, retain variance.
Many scholars using traditional Single-level models tend to average their data resulting in non-isomorphic translations that badly affect the predictive power of the model.

DAT246-072-Hwc 6

Question 5 :

(**6p**) Below you see a Generalized Linear Model

$$y_i \sim \text{Poisson}(\lambda)$$
$$f(\lambda) = \alpha + \beta x_i$$

What is $f()$, **and** why is it needed? (2p)

Provide at least **two examples** of $f()$ **and** when you would use them? (4p)

$f()$ is a link fuction used to:

- translate from linear model to likelihood.
- To avoid absurd values. ✓

Examples of a link function:

✓ log. Used with Poisson & Gamma-Poisson.

✓ logit. Used with for example Binomial.

DAT246-072-HWC 6

Question 6 :

(**6p**) What is the **purpose** and **limitations** of using *Laboratory Experiments* and *Field Experiments* as a research strategy? Provide **examples**, i.e., methods for each of the two categories, and **clarify** if one use mostly qualitative or quantitative approaches (or both).

Laboratory Experiment:

- Maximizes potential for precision of measurment of behaviour. ✓
- Minimizes potential for realism of context & generelizability over actors. ✓
- More obtrusive. ✓
- Contrived Setting. ✓
- Increasingly more universal contexts & systems. ?
- Qualitative but also quantitave. mostly!

Field Experiment:

- Maximizes potential for realism of context. ✓
- Minimizes potential for generalizability over actors & precison measuremont of behaviour. ~
- Less obtrusive than Laboratory Experiment. ✓
- Natural Setting. ✓
- Increasingly more specific contexts & systems. ✓
- Both qualitative & quantitative.

DAT246-072-HWC

3

Question 7 :

(**8p**) Below follows an abstract from a research paper. Answer the questions,

- Which of the eight research strategies presented in the ABC framework does this paper likely fit? **Justify and argue!**
- Can you argue the main validity threats of the paper, based on the research strategy you picked?
  - It would be very good if you can **list threats in the four common categories** we usually work with in software engineering, i.e., internal, external, construct, and conclusion validity threats.

Context: The term technical debt (TD) describes the aggregation of sub-optimal solutions that serve to impede the evolution and maintenance of a system. Some claim that the broken windows theory (BWT), a concept borrowed from criminology, also applies to software development projects. The theory states that the presence of indications of previous crime (such as a broken window) will increase the likelihood of further criminal activity; TD could be considered the broken windows of software systems.

Objective: To empirically investigate the causal relationship between the TD density of a system and the propensity of developers to introduce new TD during the extension of that system.

Method: The study used a mixed-methods research strategy consisting of a controlled experiment with an accompanying survey and follow-up interviews. The experiment had a total of 29 developers of varying experience levels completing a system extension tasks in an already existing systems with high or low TD density. The solutions were scanned for TD, both manually and automatically. Six of the subjects participated in follow-up interviews, where the results were analyzed using thematic analysis.

Result: The analysis revealed significant effects of TD level on the subjects' tendency to re-implement (rather than reuse) functionality, choose non-descriptive variable names, and introduce other code smells identified by the software tool SonarQube, all with at least 95% credible intervals. Additionally, the developers appeared to be, at least partially, aware of when they had introduced TD.

Conclusion: Three separate significant results along with a validating qualitative result combine to form substantial evidence of the BWT's applicability to software engineering contexts. This study finds that existing TD has a mayor impact on developers propensity to introduce new TD of various types during development. While mimicry seems to be part of the explanation it can not alone describe the observed effects.

well...

* Field Experiment. Done in a natural setting. Relatively abtrusive. and somewhat specific context. Realistic context. well...

/4

Runs the risk of recall & practice effect, so not optimal for precision of measurement.

Internal validity: Is TD density the only factor causing introducig TD? (missing control)

External validity: Done on one system. How generelizable is that?

Construct validity: Are surveys, experiments, interview done well? How good is our design & model here?

Conclusion validity: Is there an enough statistical result?

?

DAT246-072-Hwc

6

Question 8 :

(6p)

In the paper *Guidelines for conducting and reporting case study research in software engineering* by Runeson & Höst the authors present an overview of research methodology characteristics. They present their claims in three dimensions: Primary objective, primary data, and design.

Primary objective indicates what the purpose is with the methodology (e.g., explanatory), primary data indicates what type of data we mainly see when using the methodology (e.g., qualitative), while design indicates how flexible the methodology is from a design perspective (e.g., flexible)

Please fill out the table below according to the authors' presentation.

Methodology	Primary objective	Primary data	Design
Survey	descriptive ✓	quant. ✓	fixed ✓
Case study	exploratory ✓	quality ✓	flexible ✓
Experiment	explanatory ✓	quant. ~	fixed ✓
Action research	improving ✓	quality. ✓	flexible ✓

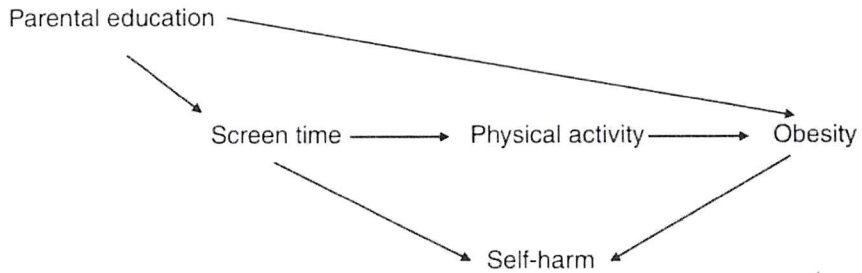
DAT246-072-HWC 3,5

Question 9 :

(**4p**) See the DAG below. We want to estimate the total causal effect of *Screen time* on *Obesity*.

Design a model where *Obesity* is approximately distributed as a Gaussian likelihood, and then write down a linear model for μ to clearly show which variable(s) we should **condition on**.

Also add, what you believe to be, **suitable priors on all parameters**. If needed to you can always state your assumptions.



Screentime → Physial activity → Obesity | closed (not a backdoor)

Screen time ← Parental education → Obesity | open unless we condi. on PE

Screen time → Self-harm ← obesity | Closed no back door

α is the grand mean of weight 40 kg is logical

$$\text{Obesity}_i \sim N(\mu, \sigma)$$

$$\mu_i = \alpha + \beta_1 \text{Screentime}_i + \beta_2 \sum_{j=0} \delta_j$$

$$\alpha \sim N(40, 15)$$

Variance

Don't assume negor pov causality ← $\beta_1 \sim N(0,1)$

No causal assumptions ← $\beta_1 \sim N(0,1)$

$\delta \sim$ ~~exp(1)~~ $\text{Dir}(\alpha)$ –

* We use a Dirichlet to include Parental Education in the model because it's a Likert-scale predictor.

Where β_2 is the total effect of PE and δ_j are summed up to 1 and represent each category in PE.

DAT246-072-Hwc

Question 10 :
([−30, 30]p)

Below follows a number of multiple choice questions. There can be more than one correct answer! Mark the correct answer by crossing the answer. In the case of DAGs, X is the treatment and Y is the outcome.

Correct answer gives 0.5 points, wrong or no answer deducts 0.5 points.

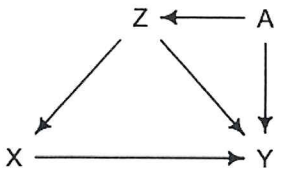
Q1 What is this construct called: $X \leftarrow Z \rightarrow Y$?
{Collider} {Pipe} {Fork} {Descendant}

Q2 What is this construct called: $X \rightarrow Z \leftarrow Y$?
{Collider} {Pipe} {Fork} {Descendant}

Q3 What is this construct called: $X \rightarrow Z \rightarrow Y$?
{Collider} {Pipe} {Fork} {Descendant}

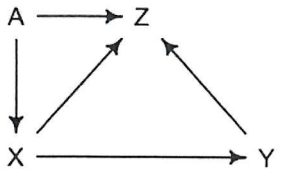
Q4 If we condition on Z we close the path $X \rightarrow Z \leftarrow Y$.
{True} {False}

Q5 What should we condition on?



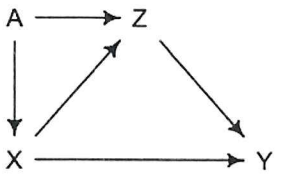
{A} {Z} {Z and A} {Nothing}

Q6 What should we condition on?



{A} {Z} {Z and A} {Nothing}

Q7 What should we condition on?



{A} {Z} {Z and A} {Nothing}

Q8 How can overfitting be avoided?
{cross-validation} {unregularizing priors} {Use a desinformation criteria}

Q9 Berkson's paradox is when two events are...
{correlated but should not be} {correlated but has no causal statement} {not correlated but should be}

Q10 An interaction is an influence of predictor...
{conditional on parameter} {on a parameter} {conditional on other predictor}

Q11 How does an interaction look like in a DAG?
$\{X \leftarrow Z \rightarrow Y\}$ $\{X \rightarrow Z \rightarrow Y\}$ $\{X \rightarrow Z \leftarrow Y\}$

DAT246-072-HWC 13,5

Q12 To measure the same size of an effect in an interaction as in a main/population/β parameter requires, as a rule of thumb, at least a sample size that is...
{4x larger} {16x larger} {16x smaller} {8x smaller}

Q13 We interpret interaction effects mainly through...
{tables} {posterior means and standard deviations} {plots}

Q14 In Hamiltonian Monte Carlo, what does a divergent transition usually indicate?
{A steep region in parameter space} {A flat region in parameter space} {Both}

Q15 Your high $\widehat{R}$ values indicate that you have a non-stationary posterior. What do you do now?
{Visually check you chains} {Run chains for longer} {Check effective sample size} {Check E-BMFI values}

Q16 What distribution maximizes this? $H(p) = -\sum_{i=1}^{n} p_i \log(pi)$
{Flattest} {Most complex} {Most structured} {Distribution that can happen the most ways}

Q17 What distribution to pick if it's a real value in an interval?
{Uniform} {Normal} {Multinomial}

Q18 What distribution to pick if it's a real value with finite variance?
{Gaussian} {Normal} {Binomial}

Q19 Dichotomous variables, varying probability?
{Binomial} {Beta-Binomial} {Negative-Binomial/Gamma-Poisson}

Q20 Non-negative real value with a mean?
{Exponential} {Beta} {Half-Cauchy}

Q21 Natural value, positive, mean and variance are equal?
{Gamma-Poisson} {Multinomial} {Poisson}

Q22 We want to model probabilities?
{Gamma} {Beta} {Delta}

Q23 Unordered (labeled) values?
{Categorical} {Cumulative} {Nominal}

Q24 Why do we use link functions in a GLM?
{Translate from likelihood to linear model} {Translate from linear model to likelihood} {To avoid absurd values}

Q25 On which effect scale are parameters?
{Absolute} {None} {Relative}

Q26 On which effect scale are predictions?
{Absolute} {None} {Relative}

Q27 We can use ... to handle over-dispersion.
{Beta-Binomial} {Negative-Binomial} {Exponential-Poisson}

Q28 In *zi* models we assume the data generation process consist of two disparate parts?
{Yes} {No}

Q29 Ordered categories have...
{a defined maximum and minimum} {continuous value} {an undefined order} {unknown 'distances' between categories}

Q30 When modeling ordered categorical predictors we can express them in math like this: $\beta \sum_{j=0} \delta_j$. Cross correct statement.
{δ is total effect of predictor and β are proportions of total effect} {β is total effect of predictor and δ are proportions of total effect}