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).

Solution

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

— Pierre-Simon Laplace

Grade 3: 39 points; ∼50%
Grade 4: 55 points; ∼70%
Grade 5: 70 points; ∼90%
Maximum: 78 points

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...)

Solution: There is no standard. However, quite recently Gelman *et al.* (arXiv:2011.01808) published a manuscript on arXiv explaining what they believed were the key parts in Bayesian data analysis. In short, a lot of the things they talk about in that paper are things we have covered in this course.

In order to get 8 points there need to be a minimum of 4 areas that are listed and explained, e.g., prior or posterior predictive checks, diagnostics, model comparison, etc.

Start with a null model, do prior checks, check diagnostics, do posterior checks, then conduct inferential statistics if needed. Also, it would be good if you mention **comparisons** of models and that it is an **iterative** approach.

Solution

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)

Solution:

This question can be answered in many different ways. Here's one way.

complete pooling

$$
\begin{aligned}
y_i &\sim \text{Binomial}(n, p_i) \\
\text{logit}(p_i) &= \alpha \\
\alpha &\sim \text{Normal}(0, 2.5)
\end{aligned}
$$

no pooling

$$
\begin{aligned}
y_i &\sim \text{Binomial}(n, p_i) \\
\text{logit}(p_i) &= \alpha_{\text{CLUSTER}[j]} \\
\alpha_j &\sim \text{Normal}(0, 2)
\end{aligned}
$$

partial pooling using hyper-parameters and hyper-priors

$$
\begin{aligned}
y_i &\sim \text{Binomial}(n, p_i) \\
\text{logit}(p_i) &= \alpha_{\text{CLUSTER}[j]} \\
\alpha_j &\sim \text{Normal}(\bar{\alpha}, \sigma) \\
\bar{\alpha} &\sim \text{Normal}(0, 1) \\
\sigma &\sim \text{Exponential}(1)
\end{aligned}
$$

Solution

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.

Solution: Epistemological, rooted in information theory and the concept of maximum entropy distributions. Ontological, the way nature works, i.e., a physical assumption about the world.

One example could be the Normal likelihood for real (continuous) numbers where we have additive noise in the process (ontological), and where we know that in such cases Normal is the maximum entropy distribution.

Solution

Question 4 :

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

Solution:

- Improved estimates for repeat sampling. When more than one observation arises from the same individual, location, or time, then traditional, single-level models either maximally underfit or overfit the data.
- Improved estimates for imbalance in sampling. When some individuals, locations, or times are sampled more than others, multilevel models automatically cope with differing uncertainty across these clusters. This prevents over-sampled clusters from unfairly dominating inference.
- Estimates of variation. If our research questions include variation among individuals or other groups within the data, then multilevel models are a big help, because they model variation explicitly.
- Avoid averaging, retain variation. Frequently, scholars pre-average some data to construct variables. This can be dangerous, because averaging removes variation, and there are also typically several different ways to perform the averaging. Averaging therefore both manufactures false confidence and introduces arbitrary data transformations. Multilevel models allow us to preserve the uncertainty and avoid data transformations.

Solution

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)

Solution: Generalized linear models need a *link function*, because rarely is there a "μ", a parameter describing the average outcome, and rarely are parameters unbounded in both directions, like μ is. For example, the shape of the binomial distribution is determined, like the Gaussian, by two parameters. But unlike the Gaussian, neither of these parameters is the mean. Instead, the mean outcome is np, which is a function of both parameters. For Poisson we have λ which is both the mean and the variance.

The link function f provides a solution to this common problem. A link function's job is to map the linear space of a model like $\alpha + \beta x_i$ onto the non-linear space of a parameter!

Solution

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).

Solution:

Laboratory Experiments: Real actors' behavior

Non-natural settings (contrived)

The setting is more controlled

Obtrusiveness: High level of obtrusiveness

Goal:

Minimize confounding factors and extraneous conditions

Maximize the precision of the measurement of the behavior

Establish causality between variables

Limited number of subjects

Limits the generalizability

Unrealistic context

Examples of research methods: Randomized controlled experiments and quasi-experiments, Benchmark studies

Mostly quantitative data

Field Experiments: Natural settings.

Obtrusiveness: Researcher manipulates some variables/properties to observe an effect. Researcher does not control the setting.

Goal: Investigate/Evaluate/Compare techniques in concrete and realistic settings.

Limitations: Low statistical generalizability (claim analytical generalizability!) Fairly low precision of the measurement of the behavior (confounding factors).

Not the same as an experimental study (changes are made and results observed), but does not have control over the natural setting.

Correlation observed but not causation.

Examples: Evaluative case study, Quasi-experiment, and Action research.

Often qualitative and quantitative.

Solution

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.

Solution: This is a mixed method study, i.e., an experiment (but not an RCT!), but where we also see qualitative parts, like from a field study.

Solution

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			
Case study			
Experiment			
Action research			

Solution:

Table 1 Overview of research methodology characteristics

Methodology	Primary objective	Primary data	Design
Survey	Descriptive	Quantitative	Fixed
Case study	Exploratory	Qualitative	Flexible
Experiment	Explanatory	Quantitative	Fixed
Action research	Improving	Qualitative	Flexible

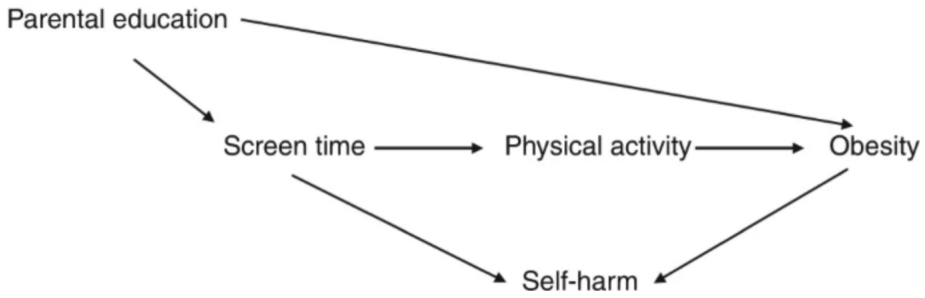
Solution

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.



Solution: 1 p for correct conditioning, 3p for appropriate model as further down below.

```
> dag <- dagitty("dag {
    PE -> O
    S -> PA -> O
    PE -> S -> SH
    O -> SH
    }")
> adjustmentSets(dag, exposure = "S", outcome = "O")
{ PE }
```

and the answer is `{ PE }`, i.e., *Parental education*. In short, check what elementary confounds we have in the DAG (DAGs can only contain four), follow the paths, decide what to condition on. If we condition on Self-harm we're toast...

$$
\begin{aligned}
O_i &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta_{\text{ST}}\text{ST}_i + \beta_{\text{PE}}\text{PE}_i \\
\alpha &\sim \text{Normal}(0, 5) \\
\{\beta_{\text{ST}}, \beta_{\text{PE}}\} &\sim \text{Normal}(0, 1) \\
\sigma &\sim \text{Exponential}(1)
\end{aligned}
$$

Solution

Question 10 :
([−**15**, **15**]**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.**

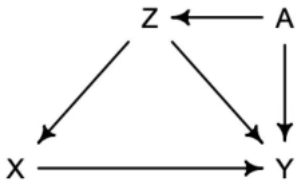
Q1 What is this construct called: $X \leftarrow Z \rightarrow Y$?
{Collider} {Pipe} {**Fork**} {Descendant}

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}

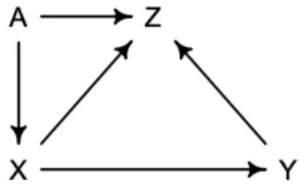
Q4 If we condition on Z we close the path $X \rightarrow Z \leftarrow Y$.
{True} {**False**}

Q5 What should we condition on?



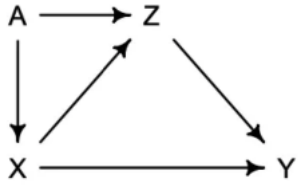
{A} {**Z**} {Z and A} {Nothing}

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?
{$X \leftarrow Z \rightarrow Y$} {$X \rightarrow Z \rightarrow Y$} {$\mathbf{X} \rightarrow \mathbf{Z} \leftarrow \mathbf{Y}$}

solution

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} {T**ranslate 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**}

solution