STATG019/3019/M019 SELECTED TOPICS IN STATISTICS — IN-COURSE ASSESSMENT 1 (2017/18 SESSION)

General rules applicable to this ICA:

- The ICA consists of two parts, part A and part B. Part A may be worked on as part of a group, while part B should be worked on individually.
- Your solutions should be submitted digitally, including the full, well-commented code as an appendix, on moodle-TurnitIn - both before the deadline.
- The submitted solution must be *anonymous*, please take care to remove any indication that would disclose your, or your group's identity from your submission. Note that this includes comments or file names referenced in the code.
- Late submission will incur a penalty unless there are extenuating circumstances (e.g. medical) supported by appropriate documentation. Penalties are set out in the latest editions of the Statistical Science Department student handbooks.
- The formal submission time is when your last digital submission process is *complete*, rather than when it was *initiated*. In particular, when submitting digitally via TurnitIn, in your own interest, make sure you do so early on. Since initiating a submission close to a deadline may result in a submission after the deadline, thus incur late submission penalties. You can replace an earlier submission by an updated one, and the time of the latest submission counts (with potentially implied penalties).
- Failure to submit this in-course assessment will mean that your overall examination mark is recorded as "non-complete", i.e. you will not obtain a pass for the course.
- Any plagiarism or collusion will normally result in zero marks for all students involved, and may also mean that your overall examination mark is recorded as non-complete. Guidelines as to what constitutes plagiarism or collusion may be found in the Departmental Student Handbooks and are handed out as a print copy in the course. The Turn-It-In plagiarism detection system may be used to scan your submission for evidence of plagiarism and collusion.
- The graded ICA, including comments feedback, will be made availabe for you on TurnitIn, and you will receive a provisional grade — *grades are provisional until confirmed by the Statistics Examiners' Meeting in June 2018*. You will be able to view the grades and comments for your submission until the TurnitIn archiving cycle in the summer break.

Rules specific to group work, applicable to part A:

- You are allowed to choose your own group, but only subject to the constraints stated in the student handbook: maximally 3 members in a group, and no two members of the group have already been working together on the group work part of the first ICA.
- You will have to register your choice of group at latest by the submission deadline, via the digital form available on the course moodle. You are not allowed to change groups after that (except in the presence of suitable extenuating circumstances).
- You are not allowed to exchange information, especially solutions, about the ICA with anybody outside your registered group except the course lecturer (F. Király). Doing so may constitute collusion or plagiarism.
- In particular, exchanging information with tentative group members before registration of a group is not allowed and may constitute collusion or plagiarism.
- All members of a group get the same mark on part A of this ICA.
- Please be always respectful towards your group partners, whoever they are. In particular, please be aware that your collaboration on the project is of professional nature, so please be careful to respect personal boundaries.
- If there is any problem that hinders or prevents collaboration within your group, please inform the course lecturer (F. Király) as soon as possible.

Rules specific to individual work, applicable to part B:

- You are not allowed to exchange information, especially solutions, about the ICA with anybody except the course lecturer (F. Király). In particular, you are not allowed to work on part B with group members you collaborate with on part A. Doing so may constitute collusion or plagiarism.

STATG019/3019/M019: In-Course Assessment 2 (model validation/assessment/selection)

Your second in-course assessment is worth 50 marks, representing 50% of the credit for the course STATG019/3019/M019. In total, 50 marks can be reached by completing the tasks not marked as "bonus", 40 marks on part A an 10 marks on part B. There are 5 additional bonus points (marked with a star*) to achieve in part B, which will also count as points towards your final grade, except if you would exceed a total of 50 marks on this ICA, in which case your total will be 50 marks. The numbers in square brackets indicate the total marks attached to each question. The non-bonus marks attached to each question are as follows: A1 (5), A2 (2), A3 (5), A4 (9), A5 (9), A6 (10), B.a (2), B.b (2), B.c (6).

You should tackle the tasks below and produce one report for each of the two parts, providing a well-written exposition of your findings. Illustrate your report by graphs, tables, mathematical derivations, or other content as appropriate. Please attach all code used in producing your results in part A as an appendix to your report.

While the exclusive use of R or python for all data analysis is compulsory, you are free to use a word processor of your choice for typesetting (e.g. TeX, LibreOffice, OpenOffice or Microsoft Word). Part Your report on part A should be digitally typeset. Your report on part B should either be digitally typeset, or consist of scans of hand-written documents.

Part A is to be worked on in groups of up to three; part B is for individual submission. For each part, you should submit *one* digital copy of the report, in part A together with your R and/or python code, via the respective TurnitIn module on the STATG019/3019/M019 Moodle page. Note that there will be two submission modules, one for part A (group work) and one for part B (individual work). In particular, do *not* submit your solution to part B to the submission module for part A since this will count as collusion and plagiarism, as it makes your individual solution to part B visible to the other members of your group for part A.

Please make sure that all submitted reports are *anonymized*, that is, please make sure there is no information in the report itself that hints at your identity. The deadline for submission is Wednesday, 25$^{\text{th}}$ April 2018, 11:55 pm (five minutes to midnight).

Part A: Artificially Intelligent Wine Tasting

While making use of actual data (the Wine Quality Dataset), the scenario contained in this exercise is entirely fictional and unrelated to the real world data source. No similarity with any actual scenario is intended and any such similarity would be entirely coincidental.

In your first year at a data scientific consultancy, you are assigned to a project team consulting for a new food/AI start-up, which has succeeded in acquiring large amounts of venture capital with their vision of creating artificial intelligences with an appreciation for quality food - more precisely, being capable of tasting and appraising food products, with a long-term outlook towards creating an AI which can cook and eventually create novel cuisine.

As all of the start-up's founders have backgrounds in marketing, finance, or gastronomy, but not in artificial intelligence, they have enlisted your employer's services to scope possible technical avenues by which their vision can be realized. Part of this long-term, large-scale engagement (which both sides wish to keep a secret for PR reasons) is the "AI sommelier" flagship sub-project to which you have been assigned.

Since nowadays even non-experts know that the easiest way for AIs to acquire new capabilities is "learning" from human "teachers", the initial phases of the project focused on acquiring systematic data on quality wines. No expenses were spared in tracking down the most accomplished sommeliers and the highest quality wines, hiring the former to taste and rate the latter. The wines also all underwent extensive chemical and then non-expert gustatory analysis, with the aim to link human ratings with potential outputs from taste sensors developed in the "AI gustometer" project.

Your team of data scientists is now tasked with using the above data to help the AI to acquire its taste. Specifically, the start-up's CTO (Chief Technology Officer) who is also the CGO (Chief Gastronomy Officer) has set out the following questions for your team to investigate, in his most recent vision statement:

1. Can we use the sommelier/wine data to create an AI with super-human performance in wine tasting?
2. Which components of wine make a wine a good wine?
3. Can the AI use the data to create the perfect wine, i.e., wine whose quality exceeds all that we have seen?

Furthermore, due to constant media coverage of the "AI sommelier" project, a number of academic ethicists have raised concerns which your project manager would like you to investigate in addition. While the most prominent such concerns in AI ethicist circles (the AI developing a taste for human blood due to its chemical composition) are already covered by extensive (and expensive) insurances which the

start-up has taken out, one philosopher of ethics, not taken too seriously by her more senior colleagues, has raised the question

4. whether human perception of wine quality is all but subjective, i.e., perhaps there is no empirically verifiable correlate of "good" and "bad" for wine. For any such perception could be entirely a result of human biases, incompetence, unintentional priming, and self-delusion, based on say a high-quality label on the wine bottle, the price tag, or an authoritative expert opinion. And if so, what would it be that AIs could, or would learn from humans?

Your project manager (also a data scientist) believes that this question is not at all ludicrous to ask, especially since it would inform to which extent data scientist manpower should be spent on algorithmic marketing, customer analytics and social media campaigns, rather than on development work for the food applications.

Data Source

The dataset you will work on is the Wine Quality Dataset obtainable from
`https://archive.ics.uci.edu/ml/datasets/Wine+Quality`
or `http://www3.dsi.uminho.pt/pcortez/wine/`
or the STATG019/3019/M019 moodle page. (the data from all sources should agree, but in the unlikely case of discrepancy, e.g., due to a change in the internet versions, the moodle version is the authoritative source)

For the purpose of writing the report, you should pretend that it was collected as part of the scenario described above, rather than as part of the real world scenario it was actually collected in.

The Wine Quality Dataset records, for different samples of white (`winequality-white.csv`) and red wine (`winequality-red.csv`), the following variables:

1. fixed acidity = mass concentration of tartaric acid (g/dm^3)
2. volatile acidity = mass concentration of acetic acid (g/dm^3)
3. mass concentration of citric acid (g/dm^3)
4. residual sugar mass concentration (g/dm^3)
5. mass concentration of sodium chloride (g/dm^3)
6. mass concentration of free sulfur dioxide (mg/dm^3)
7. mass concentration of sulfur dioxide total (mg/dm^3)
8. density (g/cm^3)
9. pH value
10. mass concentration of potassium sulphate (mg/dm^3)
11. alcohol content ($vol\%$)

and additionally, a median sensory preference from up to three sensory assessors ("sommeliers" in the scenario) following blind sensory assessment on a subjective scale of 0 (disgusting) to 10 (excellent).

Please feel free to look at the paper *Modeling wine preferences by data mining from physicochemical properties* by Cortez et al (2009) in which very similar analyses are performed. Note that simply repeating those analyses will not net you too many points on this ICA - due to shortcomings in the published analyses, as well as due to the different background scenario.

Task A1

Discuss, before doing any analyses, which of the questions, numbered 1 to 4 in the scenario outline on p.4-5, you can answer, or to which extent you can come close to answering the questions, using only the data provided. Explicitly state the scientific, empirically quantifiable questions that you are going to address. In the final report, this should be summarized as a separate (optimally, the second) section, which does not contain any analyses. [5]

Task A2

Load the datasets `winequality-white.csv` and `winequality-red.csv` into your preferred data analysis environment (which should be an R or python environment, by the ICA rules). Create a joint dataset which additionally has colour as a variable. Conduct analyses below on this joint dataset taken as a basis. (you will receive the marks for code that creates the joint dataset in a reasonable data structure) [2]

Task A3

Perform an exploratory analysis of your data, by studying univariate summaries (incl mean and five-number-summary), distributional plots such as histograms and density plots, and bi- or multivariate scatter plots. Also investigate the differences between red and white wine. Discuss whether there are univariate variable transforms that make sense before you start with the more advanced analyses; especially in the context that the law of mass action is linear in log-concentration coordinates. Discuss whether you see apparent collinearity, non-linearity, or clustering. In the final report, this should be a section. [5]

Task A4

Conduct predictive benchmarking experiments to determine:

(i) whether, and how well wine quality can be predicted from chemical composition and colour;

(ii) whether wine colour adds predictive power above chemical composition and vice versa, in (i).

Consider the prediction task as *all* of the two following tasks: as deterministic classification, and as univariate regression.
For both tasks, choose and describe a suitable validation set-up for your benchmarking experiment. In the description (which should be part of your final report), make precise and justify all substantial choices. The methods you compare should include at least

- a non-linear support vector machine,
- an ensemble of (at least 10) trees, and
- a neural network with two or more middle layers,

all of them with properly chosen and/or tuned hyper-parameters.
In the final report, this should be a section, in which you report results in appropriate benchmarking tables with appropriate error bars/confidence intervals. [9]

Task A5

Answer the questions, numbered 1 to 4 in the scenario outline on p.4-5, as well as you can, e.g., subject to the restrictions and modifications you formulated in Task A1. For this, use the results of Tasks A3 and A4, and conduct more analyses or experiments where necessary. Also note that some questions may not be appropriately (or easily) phrased as a prediction experiment, but are rather of inferential nature, thus more appropriately addressed by model-specific inference and/or hypothesis testing. [9]

Task A6

Compile a full, joint report on Tasks A1, A3, A4, and A5, whose length is no more than *nine* pages (standard margins, A4, and font size at least 10, excluding appendix and tables/graphs). The report should contain, as its first section, an *executive summary* section at the very top which briefly states goals, findings, conclusions, and limitations, not exceeding *two pages*.

The marks for this task (A6) are awarded for writing of the report - i.e., executive summary, clarity of writing, interpretation and discussion, quality code, appropriate selection of material, clean tables and figure - as opposed to the results themselves, of Tasks A1 and A5. [10]

Part B: Estimating the Generalization Error

This part is concerned with the analysis of the most common variant of "dummy regressor" (aka best uninformed predictor) for univariate, supervised regression tasks, prediction of the training mean, which shares key asymptotics with generalized linear regression except that the asymptotics are easier to compute explicitly than in the general case.

The "dummy regressor" can be described as follows, in non-mathematical pseudocode. (1) For fitting: compute the mean of the training labels. (2) For prediction: always predict the mean computed in (1).

(a) Define appropriate mathematical symbols to describe fitting of and prediction via the "dummy regressor", optimally but not necessarily in concordance with lecture notation, e.g., as a $[\mathcal{X} \to \mathbb{R}]$-valued random variable, for which you specify the value it takes, dependent on an i.i.d. training data batch $(X_1, Y_1), \ldots, (X_N, Y_N) \sim (X, Y)$ of fixed size N. [2]

(b) In the usual i.i.d. supervised learning setting, explicitly compute mean and variance of the expected generalization squared error of the "dummy regressor", in terms of training set size N, test set size M, mean and variance of features and labels (hint: the correct equation depends on only some of these). [2]

(c) Derive MSE, bias and variance for the following estimates of the dummy regressor's generalization squared error: (i) via single-split into a training set of size N and a test set of size M; (ii) via k-fold cross-validation on a dataset of size $k * N$, (iii) via the bootstrap (any reasonable version of your choice) on a dataset of size N. [6]

(d*, bonus) Derive expectation and bias for the following estimates of the dummy regressor's generalization squared error's estimates' variances: (i) variance of the test losses, as an estimate of the variance of (c.i); (ii) the variance estimate for (c.ii) computed "the wrong way", as variance of the fold aggregate sample; (iii) the variance estimate for (c.ii) which is the mean of test loss variances over folds; (iv) a boostrap variance estimate (any sensible variant) of (c.i); a bootstrap variance estimate (any sensible variant) of (c.iii). [5*]

Write up your computations, and submit either scanned or digitally typeset solutions.