Guideline and Formatting Instructions for Final Projects (Foundations of Data Science 2020 Course)

Rodrigo Rivera-Castro [1] **Evgeny Burnaev** [1]

Abstract

This document provides guidelines and a basic template for the final project reports at **Foundations of Data Science 2020 course** organized by *Skolkovo University of Science and Technology* (Skoltech). The document is based on ICML 2020 submission guidelines. An abstract must be a single paragraph, ideally between 4–6 sentences long.

1. Online Submission

Submission of final project reports is performed entirely online, via Canvas. The guidelines below are enforced for all submissions. Here is a brief summary:

- Students have to submit two files: compiled PDF and LaTeX source in a `*.zip` file.
- The source of this guideline, based on the icml format, should be used as the LaTeXproject report template for all projects. Do not alter the style template; in particular, do not increase the report size by adding extra vertical spaces.
- Submitted projects must be from **four** (full) to **seven** pages long, **not including references and appendices**. Projects which do not obey this rule will automatically be rejected and receive zero grade.
- Students must **include author information** in reports.
- Your report should be in **10 point Times font**.
- Place figure captions *under* the figure (and omit titles from inside the graphic file itself). Place table captions *over* the table.
- Keep your abstract brief and self-contained, one paragraph and roughly 4–6 sentences describing the motivation and key results of your project.

1.1. Submitting Project Reports

Project Report Submission Deadline: The deadline for project report submission that is advertised on the course canvas page is strict. If your full submission does not reach us on time, it will not be considered for grading.

Students must provide their manuscripts in **PDF** format. Additionally, students have to submit complete LaTeX source packed in a `*.zip` format.

The usage of **Word** is prohibited. Really. We're not joking. Don't send Word or PDFs compiled from Word.

Graphics files should be of a reasonable size, and included from an appropriate format. Use vector formats (.eps/.pdf) for plots, lossless bitmap formats (.png) for raster graphics with sharp lines, and jpeg for photo-like images.

2. Content of the Project Report

The project report should be a solid logical high quality text describing your project and obtained results. The quality of the report should equal to a publication that can go into a minor venue. **The sectioning should be as follows:**

Abstract. Brief and self-contained text. One paragraph (roughly 4–6 sentences) describing the motivation and key results of your project.

Introduction. A gentle introduction to the topic of your report, deeper explanation of motivation (mentioning some recent related work on your topic). The introduction must end with the phrase **the main contributions of this report are as follows** and following concise but still very clear list of 2-4 tasks, problems, improvements, replications, experiments (listed as bullet points) that you performed in your project. This is a usual practice to make the contributions explicit to readers and reviewers. For example, see (Arjovsky et al., 2017; Wehenkel & Louppe, 2019) or almost any other conference paper.

Preliminaries (optional). In some cases, it is reasonable to

[1]Skolkovo Institute of Science and Technology, Moscow, Russia. Correspondence to: Rodrigo Rivera-Castro <rodrigo.riveracastro@skoltech.ru>.

Final Projects of the Machine Learning 2020 Course, Skoltech, Moscow, Russian Federation, 2020.

devote a specific section for notation, key concepts, definitions (especially if the topic is not widespread around data science and machine learning community). Students may include the contents of section to introduction or related work section, depending on the situation.

Related work Review of old, recent and state-of-the art methods for solving the problem students encounter in their project. At least 4-5 references should be mentioned with a brief discussion of their drawbacks and advantages.

Algorithms and Models and **Experiments and Results.**

Two main sections describing key students' results. All the relevant content should be distributed among these sections based on the topic of project, stated goals, project plan and students' decision. In general, these sections should contain clear experimental setup and a link to a **github repo** with a fully reproducible code. **Projects without a github repo with a reproducible code will be graded as zero.**

Students have to explicitly describe the algorithms, models, methods, approaches they used for solving their project's problem. Students should explain the motivation for choosing the models, possible benefits and drawbacks of the choice in application to their problem. The used metrics for accessing the quality of the results should also be described.

The section(s) must contain a **complete description of the datasets** used for experiments with all the required download links. This includes number of features, samples, types of features (categorial, real, pixels, etc.), description of key features, etc. If well-known datasets are used, e.g. MNIST, CIFAR, etc., it is enough to put a link to a dataset (or related paper) without a detailed description.

All the **preprocessing** and data-handling steps should be presented in these sections. Make sure to answer relevant questions, e.g. the following ones: How data was normalized? How data augmentation was done? How data was cleaned from outliers or anomalies? How the data was splitted for train, test, validation?

All **training parameters** should be listed. Which methods did you try and with which parameters (e.g. neural network architectures, weight initialization, optimizers, optimizer parameters, number of epochs, iterations, cross validation, exact number of restarts, etc.)?

We highly encourage students to additionally present experimental results in a form of **tables and plots** (e.g. generated images for projects related to image generation, segmented images for project related to segmentation, table with scores for projects related to prediction, etc.). All the experimental results must be properly discussed and explained.

Conclusion Concise description of experimental results and outcomes, including possible directions for further work.

References. For references, we highly encourage students to use `references.bib` file provided with the template. To add a reference to the `.bib`, use its Bib-Tex, e.g. provided by Google Scholar. To cite a reference, use `\cite{}`. Students are expected to refer at least to 5 relevant papers.

Appendix A: Contributions of the Team Members.

In this **mandatory** appendix, the contribution of each team member to the project has to be clarified.

Appendix B: Reproducibility Checklist Students **must** answer all the questions listed in the Appendix B (end of this document) regarding reproducibility of the project results.

Filling appendices A and B is mandatory. **Projects with empty or incomplete appendices will be graded as zero.**

3. Format of the Project Report

The source of this guideline, based on the icml format, should be used as the LaTeXproject report template for all projects.

3.1. Dimensions

The text of the report should be formatted in two columns, with an overall width of 6.75 inches, height of 9.0 inches, and 0.25 inches between the columns. The left margin should be 0.75 inches and the top margin 1.0 inch (2.54 cm). The right and bottom margins will depend on whether you print on US letter or A4 paper.

The paper body should be set in 10 point type with a vertical spacing of 11 points.

3.2. Title

The project title should be set in 14 point bold type and centered between two horizontal rules that are 1 point thick, with 1.0 inch between the top rule and the top edge of the page. Capitalize the first letter of content words and put the rest of the title in lower case.

3.3. Abstract

The paper abstract should begin in the left column, 0.4 inches below the final address. The heading 'Abstract' should be centered, bold, and in 11 point type. The abstract body should use 10 point type, with a vertical spacing of 11 points, and should be indented 0.25 inches more than normal on left-hand and right-hand margins. Insert 0.4 inches of blank space after the body. Keep your abstract brief and self-contained, limiting it to one paragraph and roughly 4–6 sentences. Gross violations will require correction at the camera-ready phase.

3.4. Partitioning the Text

You should organize your report into sections and paragraphs to help readers place a structure on the material and understand its contributions.

3.4.1. Sections and Subsections

Section headings should be numbered, flush left, and set in 11 pt bold type with the content words capitalized. Leave 0.25 inches of space before the heading and 0.15 inches after the heading.

Similarly, subsection headings should be numbered, flush left, and set in 10 pt bold type with the content words capitalized. Leave 0.2 inches of space before the heading and 0.13 inches afterward.

Finally, subsubsection headings should be numbered, flush left, and set in 10 pt small caps with the content words capitalized. Leave 0.18 inches of space before the heading and 0.1 inches after the heading.

Please use no more than three levels of headings.

3.4.2. Paragraphs and Footnotes

Within each section or subsection, you should further partition the paper into paragraphs. Do not indent the first line of a given paragraph, but insert a blank line between succeeding ones.

You can use footnotes[1] to provide readers with additional information about a topic without interrupting the flow of the paper. Indicate footnotes with a number in the text where the point is most relevant. Place the footnote in 9 point type at the bottom of the column in which it appears. Precede the first footnote in a column with a horizontal rule of 0.8 inches.[2]

3.5. Figures

You may want to include figures in the paper to illustrate your approach and results. Such artwork should be centered, legible, and separated from the text. Lines should be dark and at least 0.5 points thick for purposes of reproduction, and text should not appear on a gray background.

Label all distinct components of each figure. If the figure takes the form of a graph, then give a name for each axis and include a legend that briefly describes each curve. Do not include a title inside the figure; instead, the caption should serve this function.

[1]Footnotes should be complete sentences.

[2]Multiple footnotes can appear in each column, in the same order as they appear in the text, but spread them across columns and pages if possible.

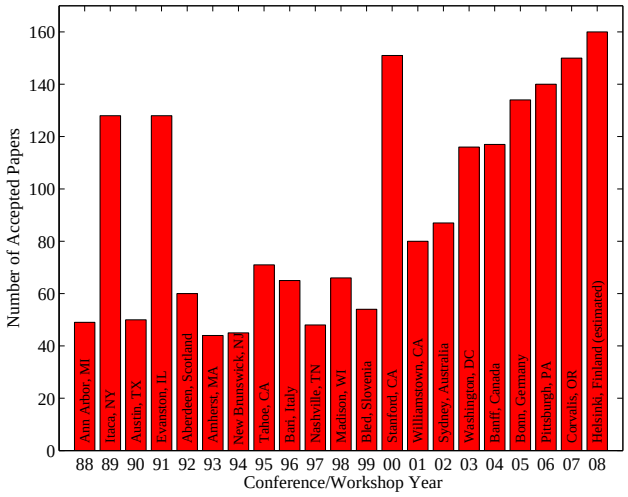

Figure 1. Historical locations and number of accepted papers for International Machine Learning Conferences (ICML 1993 – ICML 2008) and International Workshops on Machine Learning (ML 1988 – ML 1992). At the time this figure was produced, the number of accepted papers for ICML 2008 was unknown and instead estimated.

Algorithm 1 Bubble Sort

Input: data x_i, size m
repeat
  Initialize $noChange = true$.
  for $i = 1$ **to** $m - 1$ **do**
    if $x_i > x_{i+1}$ **then**
      Swap x_i and x_{i+1}
      $noChange = false$
    end if
  end for
until $noChange$ is $true$

Number figures sequentially, placing the figure number and caption *after* the graphics, with at least 0.1 inches of space before the caption and 0.1 inches after it, as in Figure 1. The figure caption should be set in 9 point type and centered unless it runs two or more lines, in which case it should be flush left. You may float figures to the top or bottom of a column, and you may set wide figures across both columns (use the environment `figure*` in LaTeX). Always place two-column figures at the top or bottom of the page.

3.6. Algorithms

Please use the "algorithm" and "algorithmic" environments to format pseudocode. These require the corresponding stylefiles, algorithm.sty and algorithmic.sty, which are supplied with this package. Algorithm 1 shows an example.

Table 1. Classification accuracies for naive Bayes and flexible Bayes on various data sets.

DATA SET	NAIVE	FLEXIBLE	BETTER?
BREAST	95.9± 0.2	96.7± 0.2	√
CLEVELAND	83.3± 0.6	80.0± 0.6	×
GLASS2	61.9± 1.4	83.8± 0.7	√
CREDIT	74.8± 0.5	78.3± 0.6	
HORSE	73.3± 0.9	69.7± 1.0	×
META	67.1± 0.6	76.5± 0.5	√
PIMA	75.1± 0.6	73.9± 0.5	
VEHICLE	44.9± 0.6	61.5± 0.4	√

3.7. Tables

You may also want to include tables that summarize material. Like figures, these should be centered, legible, and numbered consecutively. However, place the title *above* the table with at least 0.1 inches of space before the title and the same after it, as in Table 1. The table title should be set in 9 point type and centered unless it runs two or more lines, in which case it should be flush left.

Tables contain textual material, whereas figures contain graphical material. Specify the contents of each row and column in the table's topmost row. Again, you may float tables to a column's top or bottom, and set wide tables across both columns. Place two-column tables at the top or bottom of the page.

3.8. Citations and References

Please use APA reference format regardless of your formatter or word processor. If you rely on the LaTeX bibliographic facility, use `natbib.sty` and `icml2020.bst` included in the style-file package to obtain this format.

Citations within the text should include the authors' last names and year. If the authors' names are included in the sentence, place only the year in parentheses, for example when referencing Arthur Samuel's pioneering work (1959). Otherwise place the entire reference in parentheses with the authors and year separated by a comma (Samuel, 1959). List multiple references separated by semicolons (Kearns, 1989; Samuel, 1959; Mitchell, 1980). Use the 'et al.' construct only for citations with three or more authors or after listing all authors to a publication in an earlier reference (Michalski et al., 1983).

Use an unnumbered first-level section heading for the references, and use a hanging indent style, with the first line of the reference flush against the left margin and subsequent lines indented by 10 points. The references at the end of this document give examples for journal articles (Samuel, 1959), conference publications (Langley, 2000), book chapters (Newell & Rosenbloom, 1981), books (Duda et al., 2000), edited volumes (Michalski et al., 1983), technical reports (Mitchell, 1980), and dissertations (Kearns, 1989).

Please put some effort into making references complete, presentable, and consistent. If using bibtex, please protect capital letters of names and abbreviations in titles, for example, use {B}ayesian or {L}ipschitz in your .bib file.

References

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

Duda, R. O., Hart, P. E., and Stork, D. G. *Pattern Classification*. John Wiley and Sons, 2nd edition, 2000.

Kearns, M. J. *Computational Complexity of Machine Learning*. PhD thesis, Department of Computer Science, Harvard University, 1989.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Michalski, R. S., Carbonell, J. G., and Mitchell, T. M. (eds.). *Machine Learning: An Artificial Intelligence Approach, Vol. I.* Tioga, Palo Alto, CA, 1983.

Mitchell, T. M. The need for biases in learning generalizations. Technical report, Computer Science Department, Rutgers University, New Brunswick, MA, 1980.

Newell, A. and Rosenbloom, P. S. Mechanisms of skill acquisition and the law of practice. In Anderson, J. R. (ed.), *Cognitive Skills and Their Acquisition*, chapter 1, pp. 1–51. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, 1981.

Samuel, A. L. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):211–229, 1959.

Wehenkel, A. and Louppe, G. Unconstrained monotonic neural networks. In Wallach, H., Larochelle, H., Beygelzimer, A., d Alche-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 1545–1555. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8433-unconstrained-monotonic-neural-networks.pdf.

A. Team member's contributions

Explicitly stated contributions of each team member to the final project.

Name 1 (20% of work)

- Reviewing literate on the topic (3 papers)
- Coding the main algorithm
- Experimenting with model parameters on MNIST dataset
- Preparing the GitHub Repo
- Preparing the Section N of this report
- ...

Name 2 (25% of work)

- ...

Name 3 (55% of work)

- ...

B. Reproducibility checklist

Answer the questions of following reproducibility checklist. If necessary, you may leave a comment.

1. A ready code was used in this project, e.g. for replication project the code from the corresponding paper was used.

   ☑ Yes.
   ☐ No.
   ☐ Not applicable.

   General comment: If the answer is **yes**, students must explicitly clarify to which extent (e.g. which percentage of your code did you write on your own?) and which code was used.

   Students' comment: None

2. A clear description of the mathematical setting, algorithm, and/or model is included in the report.

   ☐ Yes.
   ☐ No.
   ☐ Not applicable.

   Students' comment: None

3. A link to a downloadable source code, with specification of all dependencies, including external libraries is included in the report.

   ☐ Yes.
   ☐ No.
   ☐ Not applicable.

   Students' comment: None

4. A complete description of the data collection process, including sample size, is included in the report.

   ☐ Yes.
   ☐ No.
   ☐ Not applicable.

   Students' comment: None

5. A link to a downloadable version of the dataset or simulation environment is included in the report.

   ☐ Yes.
   ☐ No.
   ☐ Not applicable.

   Students' comment: None

6. An explanation of any data that were excluded, description of any pre-processing step are included in the report.

   ☐ Yes.
   ☐ No.
   ☐ Not applicable.

   Students' comment: None

7. An explanation of how samples were allocated for training, validation and testing is included in the report.

   ☐ Yes.
   ☐ No.
   ☐ Not applicable.

   Students' comment: None

8. The range of hyper-parameters considered, method to select the best hyper-parameter configuration, and specification of all hyper-parameters used to generate results are included in the report.

   ☐ Yes.
   ☐ No.
   ☐ Not applicable.

   Students' comment: None

9. The exact number of evaluation runs is included.

   ☐ Yes.
   ☐ No.
   ☐ Not applicable.

   Students' comment: None

10. A description of how experiments have been conducted is included.

    ☐ Yes.
    ☐ No.
    ☐ Not applicable.

    Students' comment: None

11. A clear definition of the specific measure or statistics used to report results is included in the report.

    ☐ Yes.
    ☐ No.
    ☐ Not applicable.

    Students' comment: None

12. Clearly defined error bars are included in the report.

    ☐ Yes.
    ☐ No.
    ☐ Not applicable.

    Students' comment: None

13. A description of the computing infrastructure used is included in the report.

    - ☐ Yes.
    - ☐ No.
    - ☐ Not applicable.

    Students' comment: None