

Predicting Wine Quality

Mackenzie Carey and Laura Hiatt

{macarey, lahiatt}@davidson.edu

Davidson College

Davidson, NC 28035

U.S.A.

Abstract

The abstract should be between four and seven sentences long. Introduce the problem you are studying. Describe what you did. Summarize your results — what did you discover, what is the main take-away message?) Basically, you’re trying to sell your paper to the reader, so be brief and to the point. **Do not include any citations in the abstract.**

1 Introduction

Given a set of data points, regression in its simplest form attempts to determine a line of best fit to describe the relationship between both an input and an output variable. However, a straight line may fail to accurately represent the data. A better representation may be through the form of a higher order polynomial. In our experiment, we utilize a random search k-fold cross-validation implementation of stochastic gradient descent to determine the relationship between 11 attributes of wine, Figure 1, and a quality score as determined by wine experts between one and ten.

Feature
<i>fixed acidity</i>
<i>volatile acidity</i>
<i>citric acid</i>
<i>residual sugar</i>
<i>chlorides</i>
<i>free sulfur dioxide</i>
<i>total sulfur dioxide</i>
<i>density</i>
<i>pH</i>
<i>sulphates</i>
<i>alcohol</i>

Figure 1: An example table.

In this section, you should introduce the reader to the problem you are attempting to solve. For example, for the first project: describe the dataset, and the prediction problem that you are investigating. You should also cite and briefly describe other related papers that have tackled this problem (or similar ones) in the past — things that came up during the course of your research. In the AAAI style, citations look like (?) (see the comments in the source

file `intro.tex` to see how this citation was produced). Conclude by summarizing how the remainder of the paper is organized.

Overall, the aim in this section is context-setting: what is the big-picture surrounding the problem you are tackling here?

2 Background

Regression Solvers

Stochastic gradient descent is a widely used algorithm in the field of machine learning and thus there are a wide array of off-the-self regression solvers available for use. Examples include software such as Matlab or, as used in this experiment, the scikit-learn library for Python.

The scikit-learn library provides variations of gradient descent along with

Describe any background information that the reader would need to know to understand your work. You do not have to explain algorithms or ideas that we have seen in class. Rather, use this section to describe techniques that you found elsewhere in the course of your research, that you have decided to bring to bear on the problem at hand. Don’t go overboard here — if what you’re doing is quite detailed, it’s often more helpful to give a sketch of the big ideas of the approaches that you will be using. You can then say something like “the reader is referred to X for a more in-depth description of...”, and include a citation.

This section is also a good place to describe any data pre-processing or feature engineering you may have performed. If you are *only* discussing data wrangling in this section, it’s recommended that you amend the title of the section to “Data Preparation” or something similar; otherwise, use subsections to better organize the flow.

Enumerating

Create bulleted lists by using the `itemize` command (see source code):

- Item 1
- Item 2
- Item 3

Create numbered lists by using the `enumerate` command (see source code):

1. Item 1
2. Item 2
 - (a) Sub-item 2a
 - (b) Sub-item 2b
3. Item 3

Formatting Mathematics

Entire books have been written about typesetting mathematics in \LaTeX , so this guide will barely scratch the surface of what's possible. But it contains enough information to get you started, with pointers to resources where you can learn more. First, the basics: all mathematical content needs to be written in “math-mode” — this is done by enclosing the content within $\$$ symbols. For example, the code to produce $6x + 2 = 8$ is `$\$6x + 2 = 8\$$` . Note that this is only good for inline math; if you would like some stand-alone math on a separate line, use *two* $\$$ symbols. For example, `$\$\$6x + 2 = 8\$\$$` produces:

$$6x + 2 = 8$$

Here are various other useful mathematical symbols and notations — see the source code to see how to produce them.

- Sub- and super-scripts: e^x , a_n , e^{2x+1} , a_{n+2} , f_{n+1}^i
- Common functions: $\log x$, $\sin x$
- Greek symbols: ϵ , ϕ , π , Π , Φ
- Summations: $\sum_{i=0}^{100} i^2$
- Products: $\prod_{i=0}^{\infty} 2^{-i}$
- Fractions: $3/2$

$$\frac{x+5}{2 \cdot \pi}$$

Other useful resources:

- Find the \LaTeX command you're looking for by drawing what you want to produce¹: <http://detexify.kirelabs.org/classify.html>
- Ask others: <http://tex.stackexchange.com/>
- Every \LaTeX symbol ever: <http://tinyurl.com/6s85po>

3 Experiments

In this section, you should describe your experimental setup. What were the questions you were trying to answer? What was the experimental setup (number of trials, parameter settings, etc.)? What were you measuring? You should justify these choices when necessary. The accepted wisdom is that there should be enough detail in this section that I could reproduce your work *exactly* if I were so motivated.

¹Thanks to Dr. Kate Thompson for pointing me to this resource. Also, this is how you create a footnote. But don't overuse them — prefer citations and use the acknowledgements section when possible. I usually only use footnotes when I want to include a pointer to a web site.

4 Results

Present the results of your experiments. Simply presenting the data is insufficient! You need to analyze your results. What did you discover? What is interesting about your results? Were the results what you expected? Use appropriate visualizations. Prefer graphs and charts to tables as they are easier to read (though tables are often more compact, and can be a better choice if you're squeezed for space). Data you may wish to present include things like learning curves, the results of hyperparameter grid search, etc. — this way, you can justify any choices more rigorously.

From efron paper: cross validation overfits There are a lot more white wine data points than red data points. It is easier to overfit with less data.

Embedding Pictures

See the source code (`results.tex`) for instructions on how to insert figures (like figure 2) or plots into your document.

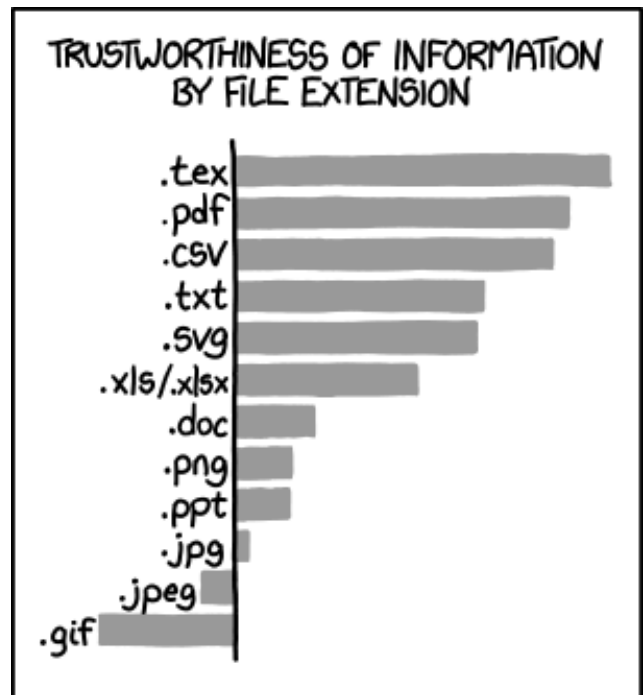


Figure 2: On the trustworthiness of \LaTeX . Image courtesy of xkcd.

Creating Tables

Again, refer to `results.tex` to learn how to create simple tables (like table 3).

Column 1	Column 2	Column 3
1	3.1	2.7
42	-1	1729

Figure 3: An example table.

5 Conclusions

In this section, briefly summarize your paper — what problem did you start out to study, and what did you find? What is the key result / take-away message? It's also traditional to suggest one or two avenues for further work, but this is optional.

6 Contributions

Briefly summarize the contributions of you and your partner in this section. For example: "A.T. wrote the gradient descent code and ran experiments; J.vN. ran the experiments using scikit-learn's built-in regression solver. A.T. wrote the introduction and background sections and prepared figure 2 and table 3. J.vN. wrote the experiments and results sections. Both authors proof-read the entire document." I will be looking for roughly equal contributions from both partners in both aspects of the assignment (i.e., the programming/data preprocessing/experimentation and writing).

7 Acknowledgements

This section is optional. But if there are people you'd like to thank for their help with the project — a person who contributed some insight, friends who volunteered to help out with data collection, etc. — then this is the place to thank them. Keep it short!