

COS 711 Assignment 2

Supervised Neural Network Training

Jeffrey Russell (u16010648)

EBIT

University of Pretoria

Pretoria, South Africa

u16010648@tuks.co.za

Abstract—In this project we were instructed to implement a neural network with automatic architecture selection and curriculum learning on two wine quality datasets. There were numerous hurdles to overcome such as the nature of the data and the imbalance of the classes given. The final results were promising with an accuracy of approximately 70% finally being achieved on the red wine and ~~insert here~~ on the white wine.

Index Terms—neural networks, curriculum learning

I. INTRODUCTION

In this project we were given a dataset containing a number of wine traits (ph, density, etc.) and the subjective quality assigned by a human. My goal was to create a neural network with certain added extras that would be able to predict the quality of the wine given the set of wine traits of any particular wine. The majority of the work fell into four different pieces of this program: preprocessing, manual parameter tuning, auto architecture selection, and a curriculum learning algorithm.

The results were promising, achieving good accuracy on the two datasets, but they did seem a bit underwhelming considering the difficulty to produce them.

II. BACKGROUND

The preprocessing of the data was fairly simple because Python has so many built-in packages for doing data manipulation and processing. The pre processing involved some statistical analysis and the PCA algorithm to remove useless properties of the wine, and reduce the dimensionality of the problem for the NN.

After the preprocessing, the next step was to create the neural network itself. I used Tensorflow for the neural network itself with categorical cross entropy as the loss function (this isn't the full story on the loss function, more on that later). The preprocessing step took care of the test/generalization/global split. The parameter tuning that I did to come up with an optimal neural network architecture and training strategy was fairly empirical, trying different values and seeing what worked and what didn't.

Insert auto architecture selection piece here

The curriculum learning part of the project ended up being fairly simple, though it was very long running. I created

a GA which decides on the order of the samples shown to the neural network in a number of batches. It progressively shows the NN more and more samples until in the final batch all of the samples are present.

III. EXPERIMENTAL SET-UP

I've split the experimental set-up into the four different parts of the project.

A. Preprocessing

I loaded the data through pandas from the csv files and then used scipy and numpy to split up the data and do the preprocessing. The data was first sent through a z-score estimation function (provided by scipy.stats) which assumes a normal distribution of each input variable and then gives each value a z-score based on the mean and standard deviation within the input variable.

The assumption of a normal distribution seemed justifiable since since most of the qualities are based on biological processes within the fermentation of wine that should present normal distribution curves when examined across a sufficiently large sample size. A couple of the input variables were fairly skewed to one side or the other, but the removal of outliers and increase in variation that the z-score introduced would seem to outweigh the possible problems with this approach.

I then removed all rows in the data which had a column that was more than three standard deviations away from the mean. This dealt with outliers nicely and made the rest of the process much simpler. I also reduced the range of the values to between 0 and 1. In order to deal with class imbalance I decided to use a cost-sensitive loss function based on how many samples there were of that kind, the exact equation for the loss is described in the neural network architecture section.

The next step for the preprocessing was to apply a PCA to the dataset in order to throw away the least useful input variables. I decided to throw away two of the input variables since that seemed to work well. That reduced the dimensionality of the input from 11 down to 9.

The final step was to split the data into training/testing and global generalization. The percentages I decided on was 70% for training and then 15% each for testing and global generalization. This split was decided on because it seemed to give good results without cutting down too much on the testing size. This was all done with the red wine dataset in mind which is a bit smaller so that influenced the decision a bit. Because all the processing steps happened before the split between test and training data, there might have been some latent information based on the preprocessing that might have slightly increased the accuracy, but I don't think that it would be enough to warrant a different approach.

B. Neural network

The neural network was created in Tensorflow with Keras as the backend. I did a lot of experimentation trying to work out which particular parameters ended up creating the best and quickest convergence. Not everything that I did I have graphs for, but the most important stuff I do. The loss function that I used was categorical cross entropy (ish, I'll discuss the loss function below since it has some interesting quirks and rationale behind it). I decided on this because it seems to be the standard for measuring loss on categorical data across the sources that I reviewed.

The input layer uses the tanh activation function. The hidden layers ended up just being a set of five layers with 9 neurons each all with activation function tanh. The output layer has 10 neurons with sigmoid activation. The input activation function is tanh because as soon as I tried any other activation function I immediately got a NaN loss in tensorflow. It seems to be because of the range of the output combined with the loss function which is categorical crossentropy which has issues with predictions of exactly zero since it has to do the log of that which causes undefined math operations. The hidden layers are all tanh since that had the best accuracy at a low number of epochs and a comparable accuracy further on. See fig 1 and 2 which compared relu, tanh, sigmoid and linear over 500 epochs running each model 100 times. Interestingly sigmoid is absolutely terrible up until about generation 300, and I have no idea why this might be.

Justify number of hidden layers;

Because of the data shown in figure 1 and 2 I decided to have a minimum of 200 epochs with an ending function. This function takes a look at the last five epochs. For each epoch it considers the loss function result on the test data and compares that to the result for the current epoch. If 3 out of the most recent five have lower losses than the current epoch then the training is aborted. This exact function seems to strike an excellent balance between stopping too early based on a false positive and stopping too late because it didn't see the trend. As soon as the test loss starts increasing consistently it stops the training. This tends to happen right around the 200 epoch mark hence my use of 200 as my

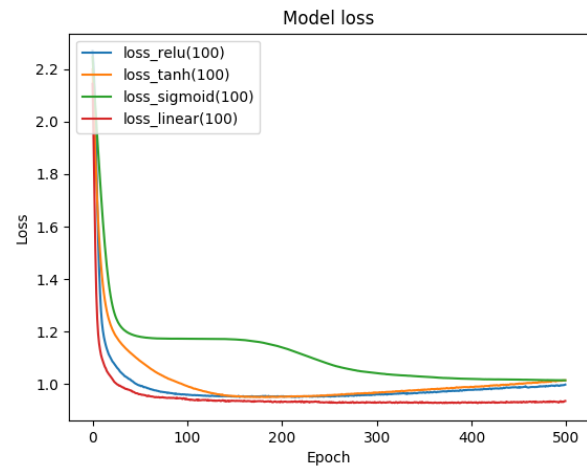


Fig. 1. Epoch number vs loss for four different activation functions

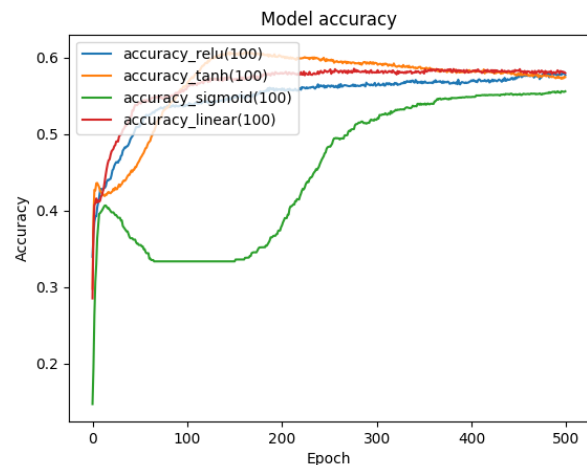


Fig. 2. Epoch number vs model accuracy for four different activation functions

minimum epoch number.

Justify batch size;

C. Automatic architecture selection

say something;

D. Curriculum learning

The curriculum learning approach was to create an ordering of the samples and then showing the NN the first set and then expanding that set of samples over time. The best ordering of the samples was created using a very basic GA with only mutation enabled. The genes of each member of the population was a list of numbers which correspond to samples in the data. The data coming in is always in the same order, so a particular sample of the data can be referred to simply by its row number. This is how the individuals in the population determine the ordering of the samples being

shown to the neural network.

The method of showing samples to the NN works in batches. I decided on five batches because it neatly divided the dataset into sets of about 200 which was the optimal number of epochs discovered in the previous section (These don't really have any real relation that I could think of). In the case of five batches the NN is shown the first fifth of the training set (according to the GA output ordering) for 40 epochs (200/5). In the next 40 epochs it is then shown the first two fifths of the training set, until for the final 40 epochs the NN is shown the entire data set to train on. This worked quite well, and it gives an easily obtainable difficulty for samples depending on where they tend to end up in the order that the GA produces. I stumbled upon this way of creating the curriculum after reading through a number of different methodologies and realizing that this particular method fit well with the type of data and the other pieces of the puzzle that make up the classifier.

IV. RESEARCH RESULTS

¡Say something¿

V. CONCLUSION

In conclusion, the final accuracy seemed quite good and the project as a whole taught me a great deal about curriculum learning and automatic architecture selection. ¡say something¿

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, ac, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. Units

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive”.
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: “Wb/m²” or “webers per square meter”, not “webers/m²”. Spell out units when they appear in text: “. . . a few henries”, not “. . . a few H”.
- Use a zero before decimal points: “0.25”, not “.25”. Use “cm³”, not “cc”).

C. Equations

Number equations consecutively. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

$$a + b = \gamma \quad (1)$$

Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1)”, not “Eq. (1)” or “equation (1)”, except at the beginning of a sentence: “Equation (1) is . . .”

D. L^AT_EX-Specific Advice

Please use “soft” (e.g., `\eqref{Eq}`) cross references instead of “hard” references (e.g., (1)). That will make it possible to combine sections, add equations, or change the order of figures or citations without having to go through the file line by line.

Please don't use the `{eqnarray}` equation environment. Use `{align}` or `{IEEEeqnarray}` instead. The `{eqnarray}` environment leaves unsightly spaces around relation symbols.

Please note that the `{subequations}` environment in L^AT_EX will increment the main equation counter even when there are no equation numbers displayed. If you forget that, you might write an article in which the equation numbers skip from (17) to (20), causing the copy editors to wonder if you've discovered a new method of counting.

BIB_TE_X does not work by magic. It doesn't get the bibliographic data from thin air but from .bib files. If you use BIB_TE_X to produce a bibliography you must send the .bib files.

L^AT_EX can't read your mind. If you assign the same label to a subsubsection and a table, you might find that Table I has been cross referenced as Table IV-B3.

L^AT_EX does not have precognitive abilities. If you put a `\label` command before the command that updates the counter it's supposed to be using, the label will pick up the last counter to be cross referenced instead. In particular, a `\label` command should not go before the caption of a figure or a table.

Do not use `\nonumber` inside the `{array}` environment. It will not stop equation numbers inside `{array}` (there won't be any anyway) and it might stop a wanted equation number in the surrounding equation.

E. Some Common Mistakes

- The word “data” is plural, not singular.
- The subscript for the permeability of vacuum μ_0 , and other common scientific constants, is zero with subscript formatting, not a lowercase letter “o”.
- In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited,

such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)

- A graph within a graph is an “inset”, not an “insert”. The word alternatively is preferred to the word “alternately” (unless you really mean something that alternates).
- Do not use the word “essentially” to mean “approximately” or “effectively”.
- In your paper title, if the words “that uses” can accurately replace the word “using”, capitalize the “u”; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones “affect” and “effect”, “complement” and “compliment”, “discreet” and “discrete”, “principal” and “principle”.
- Do not confuse “imply” and “infer”.
- The prefix “non” is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the “et” in the Latin abbreviation “et al.”.
- The abbreviation “i.e.” means “that is”, and the abbreviation “e.g.” means “for example”.

F. Authors and Affiliations

The class file is designed for, but not limited to, six authors. A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

G. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and,

conversely, if there are not at least two sub-topics, then no subheads should be introduced.

H. Figures and Tables

a) *Positioning Figures and Tables:* Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 3”, even at the beginning of a sentence.

TABLE I
TABLE TYPE STYLES

Table Head	Table Column Head		
	<i>Table column subhead</i>	<i>Subhead</i>	<i>Subhead</i>
copy	More table copy ^a		

^aSample of a Table footnote.



Fig. 3. Example of a figure caption.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

ACKNOWLEDGMENT

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page. [1]

REFERENCES

- [1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, “Modeling wine preferences by data mining from physicochemical properties,” *Decision Support Systems*, vol. 47, no. 4, pp. 547 – 553, 2009, smart Business Networks: Concepts and Empirical Evidence. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0167923609001377>

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove the template

text from your paper may result in your paper not being published.