CS 7641: Supervised Learning

Stacey Wieseneck  
swieseneck3@gatech.edu

# Classification problems

For this assignment my two classification problems are:

1. Classifying high or low heating load on a building based on factors of relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area, and glazing area distribution.
2. Classifying high or low cooling load on a building based on factors of relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area, and glazing area distribution.

The reason for choosing this data is largely from having a background in construction and manufacturing. It’s a much more commercialized industry than people realize, and a lot of developments are industry driven with new products promoting enhanced heating and cooling load coming from a corporate supplier. This data is non-trivial in that it will show if there are simple design characteristics, ones that do not make a difference for any company’s bottom line, that can still affect heating and cooling load. Also my electricity bill has been insane lately, would love to know at a high level design principles to look for in buildings where I’m responsible for the power.

# Results

The following table outlines the training and testing error rates I obtained utilizing the variable learning algorithms. For all values below, the test data was 25% of the data available.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Decision Tree | | Neural Networks | | Boosting | | SVM | | KNN | |
|  | Test | Train | Test | Train | Test | Train | Test | Train | Test | Train |
| Question 1- Error Rate | 0.03125 | 0.04513888888888889 |  |  |  |  |  |  |  |  |
| Question 2- Error Rate | 0.020833333333333332 | 0.04861111111111111 |  |  |  |  |  |  |  |  |

## Decision Trees

Starting with decision trees as my learner, both problems we the same thing. Basically this data is very, very well suited for a decision tree. The error rates for almost all splits of test/train data stays static at 0. For both problems we see that as the training data becomes only 10% of the data set overfitting occurs as the training data has less error than the testing data. Overfitting occurs a bit sooner when looking at the second problem (P2).

Looking at the number of iterations run, the data starts to get interesting. Iterations do decrease a bit even as the amount of testing data decreases slightly. While this isn’t a huge increase it is insight into the data set as this would not be expected behavior. This tells me that the data splits are much more clear when we’re working with a larger amount of data, as the data fed into the training model decreases, the bests splits are harder to decipher. Since I’m using old code from a previous class my method for determining which variable and value to split on is a bit archaic, decided through correlation coefficient with the output values and the split value determined by the mean of all the values. A lot of assumptions are made with this split and I would expect a much more reasonable and expected curve for iterations and time if I were to use GINI index or other entropy or entropy adjacent methods.

## Neural Networks

Looking at my neural networks model next. This one behaved much more like one would expect to see. I approached this with a very simple neural network, no hidden layer and optimized the weights iteratively using the gradient approach. The activation function was to “fire” for values greater than or equal to 0.5. For both problems I started by looking at the error rate for a 0.25 rate of test data to confirm the learner was indeed decreasing error. We see the error rate does decrease eventually, after a few dozen iterations on the gradient, however both plateau at a significant level of error, either unable to reduce error further or unable in the set amount of iterations without the system crashing. This shows the simple neural network model is not suitable to account for all of the variable and their interactions with such a simple model.

For test and train error rates in both classification problems, my first and top thought is that there are not many anomalies in the data. The performance of the test and train data mirror each other very, very closely, deviating by at most 0.10. Of course, as expected the training time decreased with the amount of data fed into it as expected.

My favorite aspects of neural networks is looking at the final weights given for each model which is as follows in order for the variables: relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area, and glazing area distribution. For problem 1 (\*). And for problem 2 (\*). This is very direct insight into the factors that are the most directly connected to the variable being classified, heating load for problem 1 and cooling load for problem 2. This aligns with Figure \*\*. Seeing the weight coefficient for X\* (\*\*\*) is tying directly to the overall error through the iterations.

## Boosting

## Fill in \*\*\*

## Support Vector Machines

Covering the Support Vector Machines (SVM’s) next. Overall, SVM’s had a much lower error rate than work with the decision tree or Neural Network. It stayed consistently below 0.5% error rate. However looking at the test/train error rate, even on such a small scale there were some interesting takeaways here. First the training error was significantly larger than the test error for problem 1, specifically for a polynomial kernel. While this is not expected generally, it makes since that this would occur on the low end of the amount of data being fed into the model and I read this to mean it needs more data to work with than 10% of the set available. For Problem 2 with a linear kernel, error values behave much more like I would expect them to with testing error growing as training error decreases, illustrating the smaller the training sample, generally the less representative it is of the test data.

The truly interesting part of this learner is the difference that the kernel makes on its performance. Generally in machine learning, higher order functions are thought to model the data better, create a better representation, despite having a risk of overfitting. However, for this model on both of my questions, a polynomial kernel (of degree 8) had consistently more error compared to it’s linear kernel counter part. This is perhaps my favorite insight into the data across all of the machine learning models. A kernel at its most basic function in SVM’s is a distance measure. Seeing that a linear kernel is performing better than a polynumeric one tells me very clearly that there is a linear relationship between the points and their values. Looking at the variables relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area, and glazing area distribution- this makes a lot of sense. Basic heat transfer principles are generally linear relationships with factors like area. Seeing this reflected in the two different kernel models is exciting.

Additional, training time on models performed as expected generally. Polynomial kernels took less time to train than linear. And general time decreased as the amount of data that it was training on decreased as well.

## *k-*Nearest Neighbors

# Analysis

Improved pruning

Improved split vals for decision trees

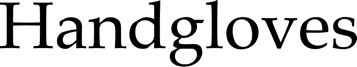
Multilayer neural network

SVM deeper learning on what type of formula to use

k-NN graphically looking at this and varying k

* Values
  + Error rate for test and train for both questions
  + Time and iterations for the algorithms that are timed
  + Decision tree
  + Time is dependent on data set size
  + Grouping of continuous factors is necessary for the most part, layering this by both factor and splitval
* Expand on
  + Second question is not showing as clear of a relationship
    - That’s what the dataset was originally used for

# Analysis

All text in JDF should be set in the Palatino typeface. It is available practically everywhere as a system font: Microsoft ships a version called *Palatino Linotype,* and Apple uses a version simply called *Palatino.* Those without either can look for “Book Antiqua” in their fonts list or download [TEX Gyre Pagella](https://www.ctan.org/tex-archive/fonts/tex-gyre/opentype) from CTAN. They all come in regular and bold weights with matching italics.

Handgloves

1. Palatino. Make sure the live text (bottom) uses the same font as the image (top).

## Body text

Body text is set in the regular weight at 11 points with 17 points of line spacing, and 8.5 points of spacing added after each paragraph. It should be justified with hyphenation enabled where available. Paragraphs should not be indented. These styles can be automatically applied using the *Normal* paragraph style.

**Bold** and *italics* should be used for emphasis. [Hyperlinks](https://foo.bar/) may be inserted in the text, as well as in-line code, superscripts like this, and subscripts like this.

## Title *&* subtitle

The paper title should be set in the regular weight at 17 points with 22 points of line spacing, centered at the top of the first page. The title may span up to three lines. For typical assignments, the document title may be as simple as “Assignment 1.” More specialized assignments may warrant more unique paper names, like “A Proposal to Create a New Document Format.”

The author’s name and email should come next *unless* you want to or were asked to submit anonymously, in which case this can be omitted. They should be set in the same size and weight as body text, centered. These styles can be applied using the *Title* and *Subtitle* paragraph styles.

## Abstract

If your paper requires an abstract, it should be placed at the top of the first page underneath the title block, preceded by the word *Abstract* in bold italics. An extra 0.5″ should be added to both sides. Not all papers require abstracts; only those that would benefit from a high-level summary of the project or its background.

## Headings

Headings should all be set in the same size as body text (11 points) in bold. With the exception of *Heading 1,* they should all have 8.5 points of space added before and after. They should be hierarchically numbered: Word, Pages, and LaTeX will do this automatically when you use the appropriate paragraph styles, but Google Docs users will need to number their headings manually. Headings should not span more than one line.

### Heading 1

*Heading 1* should be set in all caps, with letterspacing expanded by 0.66 points (6%). It should have 11 pt of space added before and 8.5 pt of space added after.

### Headings 2–4

Besides *Heading 1,* which is set in caps, headings should always use sentence case (i.e., first word capitalized) rather than title case; after all, they are not titles. *Heading 2* should be set in bold roman (upright), and *Heading 3* should be set in bold italics. The use of headings beyond *Heading 3* is discouraged.

Heading 4—*Heading 4* is provided as a run-in sidehead. Like *Heading 3,* it is set in bold italics, but it should be followed by an em dash and flow right into the text, as seen at the beginning of the current paragraph. It should be used more as a list style than a heading style, e.g. to set off a list of principles in a heuristic evaluation.

## Page layout

JDF uses the US Letter paper size (8.5″ × 11″). It has a top margin of 1″, and bottom and side margins of 1.5″. This yields a text block of 5.5″ × 8.5″, which is exactly ½ the size of the page, divided lengthwise.

The page number should be included in the bottom margin, 1″ from the bottom of the page – this creates symmetry with the top margin. No other elements should be placed in the margins.

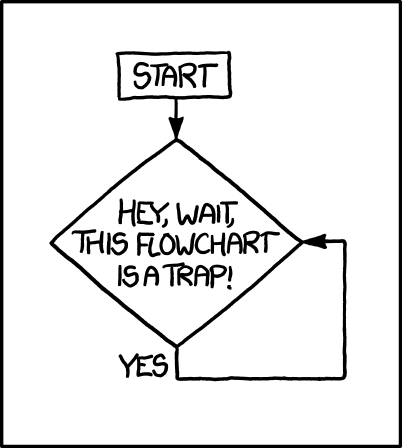
# Presentational elements

You are encouraged to use presentational elements liberally, as long as they add to the clarity of your submissions. They often require less space and fewer accompanying words to explain a given concept, and do a far better job of it.

## Figures

Figures should always be centered on the page, although they may also take up the entire width and height of the text block. Figures should always be referenced in the text, and they should include a descriptive caption. Figures may also be equations, diagrams, or other kinds of content.

If your figure includes a white background (e.g. an interface design or graph), it may aid legibility to add a ¼ point black border.

****

1. Make sure your flowcharts are more useful than this one. Source: [XKCD](https://xkcd.com/1195/).

Figure captions should be placed beneath the corresponding figure, indented 1″ on the left and right sides. The label for the figure, e.g. “Figure 1,” should be set in bold italics followed by an em dash, and the entire caption should be 8.5 points with 14 points of line spacing. The *Figure Caption* paragraph style in Word will number your figures automatically. If need be, you may have one caption corresponding to multiple consecutive figures and use either locational descriptors (e.g. “top left,” “middle”) or labels (e.g. “A”, “B”) to map parts of the caption to parts of the figure. Make sure that caption falls on the same page as the corresponding figure or table; you may need to rearrange text to make this work.

In Microsoft Word, you may need to either change the image’s text wrap settings to “Top and Bottom” or change the line spacing of the image to 1.0.

## Tables

You have freedom to format tables in the way that works best for your data. Generally, text should be left-aligned and numbers should be right-aligned or aligned at the decimal – you can do this in Word using a [decimal tab stop.](https://practicaltypography.com/tabs-and-tab-stops.html) The default table style (below) reduces the text size to be equal to the caption text.

Table captions should be formatted the same way as figure captions, but they should be placed above the table. The popular mnemonic for this is: figures at the foot, tables at the top. The *Table Caption* paragraph style will number your tables automatically. Like figures, tables should not exceed the margins and should be centered on the page.

## Additional elements

There are additional elements you may want to include in your paper, such as in-line or block quotes, lists, and more. For other content types not covered here, you have reasonable flexibility determining how it should be used in this format.

1. Mathematical constants. Notice how the approximations align at the decimal.

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Symbol | Approximation | Description |
| Golden ratio | *φ* | 1.618 | Number such that the ratio of 1 to the number is equal to the ratio of its reciprocal to 1 |
| Euler’s number | *e* | 2.71828 | Exponential growth constant |
| Archimedes’ constant | π | 3.14 | The ratio between circumference and diameter of a circle |
| One hundred | A+ | 100.00 | The grade we hope you’ll all earn in this class |

### Quotes

If you would like to quote an outside source, you may do so in quotation marks followed by a citation. If a quote is fewer than three lines, you may write it in-line. It is acceptable to replace pronouns with their target in brackets for clarity. For example, “Heavy use of peer grading would compromise [the school’s] reputation” (Joyner, 2016). If a quote exceeds three lines, you should set it as its own paragraph with 0.5″ side margins, using the *Blockquote* paragraph style.

“Whether or not the grades generated by peers are reliably similar to grades generated by experts is only one factor worth considering, however. Student perception is also an important factor. […] Reliance on peer grading is one of the top drivers of high MOOC dropout rates. This problem may be addressed by reintroducing some expert grading where possible.” (Joyner, 2016)

### Lists

Bulleted and numbered lists are indented 0.25″ from the left margin, with the bullet or number hanging by 0.25″ (i.e., flush with the left margin).

* Like this
* And this
* And also this

# Procedural elements

## In-line citations

Articles or sources to which you refer should be cited in-line with the authors’ names and the year of publication.[[1]](#footnote-2) The citation should be placed close in the text to the actual claim, not merely at the end of the paragraph. For example: students in the OMSCS program are older and more likely to be employed than students in the on-campus program (Joyner, 2017). In the event of multiple authors, list them. For example: research finds sentiment analysis of the text of OMSCS reviews corresponds to student-assigned ratings of the course (Newman *&* Joyner, 2018). You may also cite multiple studies together. For example: several studies have found students in the online version of an undergraduate CS1 class performed equally with students in a traditional version (Joyner, 2018a; Joyner, 2018b). If you would like to refer to an author in text, you may also do so by including the year (in parentheses) after the author’s name in the text. If a publication has more than 4 authors, you may list the first author followed by ‘et al.’ For example: Joyner et al. (2016) claim that a round of peer review prior to grading may improve graders’ efficiency and the quality of feedback given. This applies to parenthetical citations as well, e.g. (Joyner et al., 2016).

## Reference lists

References should be placed at the end of the paper in a dedicated section. Reference lists should be numbered and organized alphabetically by first author’s last name. If multiple papers have the same author(s) and year, you may append a letter to the end of the year to allow differentiated in-line text (e.g. Joyner, 2018a and Joyner, 2018b in the section above). If multiple papers have the same author(s), list them in chronological order starting with the older paper. Only works that are cited in-line should be included in the reference list. The reference list does not count against the length requirements.

# References

1. Joyner, D. A., Ashby, W., Irish, L., Lam, Y., Langston, J., Lupiani, I., Lustig, M., Pettoruto, P., Sheahen, D., Smiley, A., Bruckman, A., & Goel, A. (2016). Graders as Meta-Reviewers: Simultaneously Scaling and Improving Expert Evaluation for Large Online Classrooms. In *Proceedings of the Third Annual ACM Conference on Learning at Scale*. Edinburgh, Scotland.
2. Joyner, D. A. (2017). Scaling Expert Feedback: Two Case Studies. In *Proceedings of the Fourth Annual ACM Conference on Learning at Scale*. Cambridge, Massachusetts.
3. Joyner, D. A. (2018a). Intelligent Evaluation and Feedback in Support of a Credit-Bearing MOOC. In *Proceedings of the 19th International Conference on Artificial Intelligence in Education*. London, United Kingdom. Springer.
4. Joyner, D. A. (2018b). Toward CS1 at Scale: Building and Testing a MOOC-for-Credit Candidate. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*. London, United Kingdom. ACM Press.
5. Newman, H. & Joyner, D. A. (2018). Sentiment Analysis of Student Evaluations of Teaching. In *Proceedings of the 19th International Conference on Artificial Intelligence in Education*. London, United Kingdom. Springer.

# Appendices

You may optionally move certain information to appendices at the end of your paper, after the reference list. If you have multiple appendices, you should create a section with a *Heading 1* of “Appendices.” Each appendix should begin with a descriptive *Heading 2;* appendices can thus be referenced in the body text using their heading number and description, e.g. “Appendix 5.1: Survey responses.” If you have only one appendix, you can label it with the word “Appendix” followed by a descriptive title, e.g., “Appendix: Survey responses.”

These appendices do not count against the page limit, but they should not contain any information *required* to answer the question in full. The body text should be sufficient to answer the question, and the appendices should be included only for you to reference or to give additional context. If you decide to move content to an appendix, be sure to summarize the content and note it in relevant place in the body text, e.g., “The raw data can be viewed in *Appendix 5.1: Survey responses.*”

1. In-line citations are preferred over footnotes, and we favor APA citation format for both in-line citations and reference lists. Refer to the [Purdue Online Writing Lab](https://owl.purdue.edu/owl/research_and_citation/apa_style/apa_formatting_and_style_guide/in_text_citations_the_basics.html), or follow the above examples. You should use the *Footnote* paragraph style, with 8.5 point text and 14 point line spacing. [↑](#footnote-ref-2)