Unsupervised Learning and Dimensionality Reduction

Brittany N. Tan

Georgia Institute of Technology, brittanytan@gatech.edu

***Abstract* – This project includes analyses on different unsupervised learning and dimensionality reduction algorithms on two different datasets, the spam e-mail dataset which contains 4601 instances and 58 attributes and the diabetes dataset which contains 768 instances and 9 attributes. The algorithms implemented include *k*-Means Clustering, Expectation Maximization, Principal Component Analysis (PCA), Independent Component Analysis (ICA), Randomized Projections, and Feature Selection based on Information Gain.**

*k*-Means Clustering

This algorithm separates the data into *k* number of clusters with randomly picked centers and minimizes the mean squared distance from each data point when clustering [1]. The algorithm was implemented using both datasets with varying amounts of clusters and the within clusters sum of squared errors (SSE) was analyzed. This value represents the sum of the squared differences between the group’s mean and each observed data point [2]. For the spambase dataset, the optimal cluster was found using the elbow method. This method is a visual method where an “elbow” is identified [3]. This is the point where the SSE stops decreasing as rapidly and begins to plateau more. This elbow is indicated in the graph shown in Figure 1.

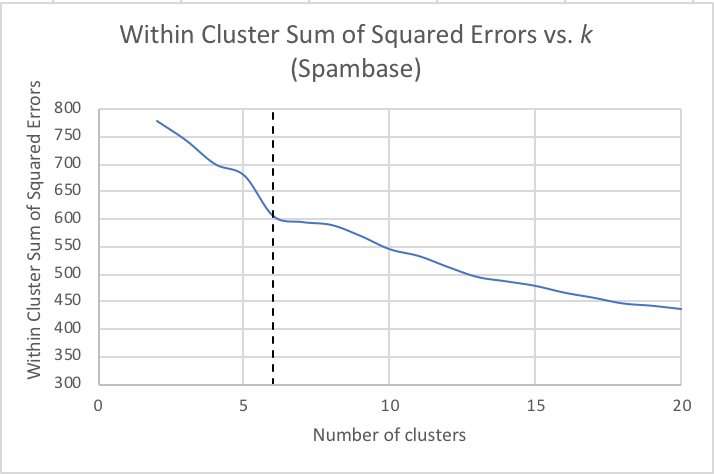


Figure 1

Sum of Squared Errors vs. Number of Clusters (Spambase)

There does not exist an unambiguous elbow, however, in the diabetes dataset. For this dataset, our optimal *k* value is chosen using the calculated silhouette width *s(i)*. The *k*-value with the largest *s(i)* is considered to be the optimal [3]. The silhouette width *s(i)* for entity *i∈ I* is defined in (1).

*s(i) =* . (1)

*a(i)* is the average Euclidean distance between *i* and the other data points in the cluster, and *b(i)* is the minimum of average Euclidean distances between *i* and the other data points. This value is evaluated and indicated in Figure 2.

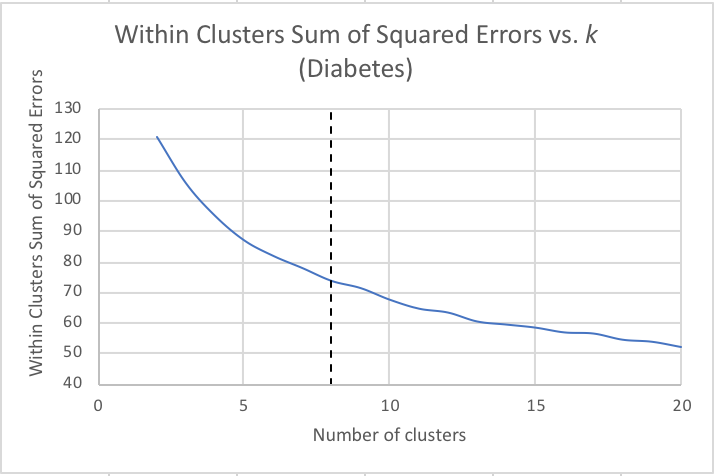


Figure 2

Sum of Squared Errors vs. Number of Clusters (Diabetes)

Expectation Maximization

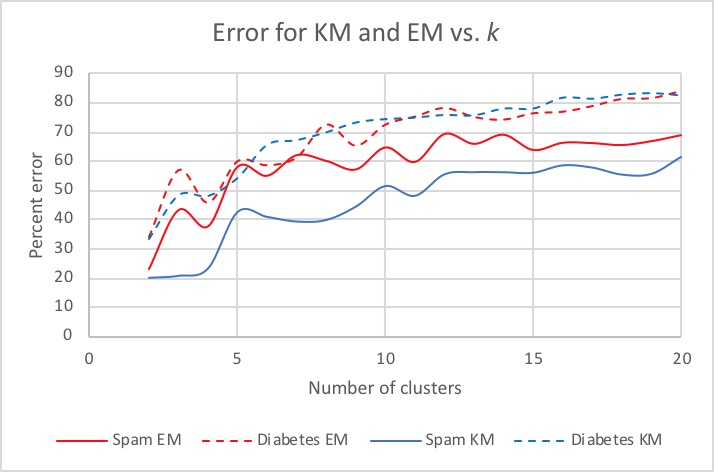


Figure 3

Error vs. Number of Clusters for k-Means and Expectation Maximization

This algorithm is a soft clustering algorithm that alternates between two different probabilistic calculations, expectation and maximization. This was implemented with both datasets, varying the number of clusters between 2 and 20 for each. One observation to note is that this algorithm resulted in a greater percent error than *k*-Means for both datasets. This may be because there may exist outliers or attributes that are not directly pertinent to classifying, and so the domain knowledge was not well-suited for the problems.

Principal Component Analysis (PCA)

Both datasets were filtered using Weka’s Principal Components filter with a 0.95 variance and then run using both clustering algorithms discussed before.

Set all these values using the "FILE" Menu. Select the Page Setup – Margins tab option and click on portrait orientation option. The above margin dimensions can then be inserted into the Page Setup Window. Now select the Paper tab and click on the paper size and select the letter paper size. Next select the Page Setup – Layout tab and set the Header and Footer to 0.5 inches. To set the column width, click anywhere in the document within the two column format, then Select the Columns for the Format menu. Click on the Equal Column Width and set the spacing. Make sure you also apply to "This Section.”

When formatting your document, make consistent use of punctuation marks and spelling. Either American or British formatting is acceptable, but it must be consistent, not a mix. For example,

* Putting all commas and periods either inside (American) or outside (British) of quotation marks
* Use of single quotes, e.g. 'service center' (British) rather than “service center" (American).
* Spellings such as grey and disc (British) vs. gray and disk (American).

Independent Component Analysis (ICA)

The title and author data is in one-column format, while the rest of the paper is in two-column format. To accomplish this, most word processors have a section break that is installed to separate the one and two-column format. There are two ways to setup this format: 1) Use this template as a guide, 2) make your own. If you wish to make your own, it is suggested that you open a new document and begin by inserting the title and author information in the standard one-column format. Please adhere to the following style guidelines:

* **Paper title**: This information should be placed at the top of the first page in 24 point, Times New Roman in Title case, and centered. This style is defined under the style menu of this document as "TITLE.”
* Insert a 11 point blank line between the Title and the Author listing and between the Author listing.
* **Author listing**: 11 point, Times New Roman, centered. This style is defined under the style menu of this document as "Author.” For Special and Panel Session, list all panelists as authors. Include only the author names in the author listing. Use the full first name for the authors. If an author is submitting more than one paper please provide the same name on both papers. Be consistent as the name you provide on the paper will be what is used to create the author index for the proceedings.
* **Author Affiliations**: 10 point, Times New Roman, centered under the author name. This style is defined under the style menu of this document as "Affiliations.” Information for each author such as email, department or college, university, city, state and zip can be listed. If you list the email address make sure the Hyperlink is not attached. If the email address turns a color that means the hyperlink is active. Right click on the address and select Remove Hyperlink. There are 3 basic types of setups that will be encounter: One author, multiple authors from the same institution and multiple authors from multiple institutions. If the paper has one author use the affiliation layout as shown at the top of this document where all the author information is listed under the paper title. For multiple authors, from the same institution, list the common university, then list the email addresses at the end in the same order as the authors appear in the Author Listing. For multiple authors from multiple institutions list the author names under the title, then list the author emails under their names. Additional information for multiple authors can be added to the Author Information section at the end of the paper.

Randomized Projections

This is a dimensionality reduction algorithm that

After you enter the title and author information enter a few blank lines and then insert a Continuous section break. Now you must define this section to be in two-column format. To do this in *Word*, under the "Format" menu select "Columns.” This option will open the Columns window. It has an input box for the number of columns. Enter 2 and then set the spacing to 0.2" and select equal column widths. If you have the margin widths set correctly, the width of the column should display as 3.40.” If it does not, go to the "File" menu and select “Page Setup - Margins.” This will open the "Page Setup" window that will allow you to set the top and bottom margins to 1" and the right and left margins to 0.75.” Every word processor will have its own method of accomplishing the above setting, however, most follow the same format.

For "Work in Process" or "Panel Session" submissions, if the paper does not fill the entire page and covers more than 4 inches of the first column, please balance the paper within the two-columns. The easiest way to do this is to insert a section break at the end of the document.

II. Font and Spacing Instructions

Use the full justify option for your columns, and use two-columns in all pages. The two columns must always exhibit equal lengths and you should try to fill your last page as much as possible. To obtain such results, you are free to adjust the figure sizes provided this does not compromise their clarity. Use one line of space between text and section headings. Use one line of space between text and captions, equations, and tables. Use the spelling and grammar checkers. Do not use the “hyphenation” feature in *Word*. Please use the following font and alignment instructions:

* **Body text**: 10 point, Times New Roman, full justified, single space, no blank lines between the paragraphs. Indents - first paragraph of section - none (this style is defined under the style menu of this document as "First Paragraph"), Indent - all other paragraphs - .25" (this style is defined under the style menu of this document as "Body Text"). Follow the examples shown in this document.
* **Section Headings**: 10 point, Times New Roman, bold, centered, use Small Caps, leave one blank line above and below. For example, "Page Layout" on page one of this document is a Section Heading (this style is defined under the style menu of this document as "Section Headings"). To convert the text to small caps, click on the “Home” tab, and select the pull down arrow in the “Font” submenu. This will open the “Font Window” where you can select the Small Cap option.
* **Section Sub-headings**: 10 point, Times New Roman, italic, left justified, leave one blank line above and below. For example, "Font and Spacing Instructions" on this page is a Section Sub Heading (this style is defined under the style menu of this document as "Subheading").
* **Bullets**: 10 point, Times New Roman, left justify and indent the text 0.25.” Insert a blank line after the bullet list but not before, follow the examples in this document (this style is defined under the style menu of this document as "Bullets").

III. Page Numbering

The page number on your manuscript is centered at the bottom of the footer. The final page number will be modified by the conference before the file is saved into the proceedings

Figures, Tables and Equations

All figures and tables must fit either one or two-column width, 3.4" or 7" wide respectively. It is suggested that you use one-column whenever possible. If your table or figure will not fit into one-column, then insert a continuous section break before and after the table or figure, as described above and define it as one-column. To make the paper read easier you may want to position any table or figure that requires one-column either at the bottom of the page or the top of a new page.

Do not abbreviate “Table”; use Roman numerals to number tables. Use the following format guidelines for figures and tables:

* **Figure and table headings**: 8 point, Times New Roman UPPERCASE, centered. Place below the figure and above the Table, (this style is defined under the style menu of this document as "Figure Heading")
* Leave one blank line above and below each Table or Figure.
* **Figure and table captions**: 8 point, Times New Roman, Small Caps, centered. Place below the figure or table headings (this style is defined under the style menu of this document as "Figure Caption")
* **Table text**: 8 point, Times New Roman, (this style is defined under the style menu of this document as "Table text")

Table I and Figure 1 below illustrates proper Table and Figure formatting. Avoid placing figures and tables before their first mention in the text. IEEE has the following rules for inserting graphics as figures:

* The manuscript’s graphics should have resolutions of 600 dpi for monochrome, 300 dpi for grayscale, and 300 dpi for color.
* Graphics should be inserted into the manuscript file using the “Insert – Picture – From File” option in the “Insert” tab. This means you must save every graphics as a separate file. Do not use cut and paste to insert graphics.
* Do not link to a graphic. When inserting figures or tables be sure you insert the figure and not just a link to the figure. The best way to make sure you are doing this correctly is to save your paper, then open the file on a different machine and make sure all your figures are correct. If you insert the link instead of the figure or table, a box with a big red x will appear in the location where the table or figure is supposed to be located.

TABLE I

Point Sizes and Type Styles

|  |  |  |
| --- | --- | --- |
| Points | Place of Text | Type Styles |
| 8  8  8  8  8  8  10  10  10  10  10  10  11  24 | Table number  Table text  Figure and Table Headings  Figure and Table Captions  Footnote  Reference list  Footer  Abstract  Index Terms  Section Titles  Main Text and Equations  Subheadings  Authors’ names  Title | Roman numerals  UPPERCASE  Small Caps  **Bold**  **Bold**  **Small Caps, Bold**  *Italic*, Left justified  Title Case |



Figure 1

Logo of the Institute for Electrical and Electronics Engineers

Number equations consecutively with equation numbers in parenthesis flush with the right margin, as in (1)

*2jk ∂u/∂z = ∂2u/∂x2 + k2 (n2* - β*2) u* . (1)

Refer to “(1),” not “Eq. (1)” or “Equation (1),” except at the beginning of a sentence: “Equation (1) is….”

Make sure you only use the “Symbol Font” for all your symbols, or embed all your different symbol fonts within the file when you save the document.

Headers and Footers

Select Page Layout tab, and then select Margins – More Margins. This set of actions will open the Page Setup Window; there, select the Layout tab. Set the Header and Footer to 0.5 inches. Please use the following format guidelines for the Header and Footer:

* **Header text**: should say Session n as shown on the top of this document. Once your paper is assigned to its final session, the General Chair will modify this to the correct session. Text should be 14 point Times New Roman, right justified, bold (this style is defined under the style menu of this document as "Header")
* **Footer text**: should be 10 point Times New Roman, bold (this style is defined under the style menu of this document as "Footer"). The text of the footer should say the same as shown on the bottom of this document. Please copy and paste this information into your document exactly as shown on this page. Once your paper is assigned to its final session, the General Chair will modify the page number to the correct session.

Acknowledgment

The preferred spelling of the word "acknowledgment" in American English is without an "e" after the "g." Use the singular heading even if you have many acknowledgments. Please put the sponsor acknowledgments in this section; do not use a footnote on the first page.

Copyright Form

An IEEE copyright form should accompany your final submission.

In Text References

All material from any research resource must be accompanied by a bracketed in-text reference. This reference must correspond to its end-text full bibliographic information in the References section. Failure to properly reference all resource material used in a paper leaves the paper’s author open to charges of plagiarism.

Follow these specifications for **in-text references**:

* Bracket all in-text references, for example [1].
* In text references must be **numbered sequentially in the text, beginning with [1]** for the first reference. In other words, the first source from which you quote, paraphrase or use information must be referenced in your paper as [1]. The next source from which your quote, paraphrase or use information must be [2]. If, later in your paper, you use information from the same source and same page as [1], then your in-text reference number will again be [1].
* Do not use “Ref. [3]” or “reference [3].” Simply use the bracketed number.
* For material summarized from several sources, use the appropriate bracketed numbers, for example [3]-[5].
* Bracketed reference numbers should appear after the quotation marks on an in-text quote, but before the final punctuation of the quote. For example, “Here’s the quote” [3]. Bracketed references for paraphrases or summaries should appear after the paraphrase or summary, but before the final punctuation of the sentence or passage. For example, Here’s the paraphrased material [4].

References

1. T. Kanungo, D. Mount, N. Netanyahu, C. Piatko, R. Silverman and A. Wu, "An efficient k-means clustering algorithm: analysis and implementation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 881-892, 2002.
2. G. Milligan and M. Cooper, "An examination of procedures for determining the number of clusters in a data set", *Psychometrika*, vol. 50, no. 2, pp. 159-179, 1985.
3. T. Kodinariya1 and P. Makwana, "Review on determining number of Cluster in K-Means Clustering", *International Journal of Advance Research in Computer Science and Management Studies*, vol. 1, no. 6, pp. 90-95, 2013.

Author Information

**Brittany N. Tan,** Undergraduate Student, College of Computing, Georgia Institute of Technology.