Department of Statistics
STATS 784: Data Mining

Zhi Zhang, 708439475, zzha822
The University of Auckland

August 3, 2017

1 Question 1

1.1 Application1

In "Applied Predictive Modelling" (Kuhn, M. and Johnson, K. (2013).), chapter 3, there is a case that decribes the application of data pre-processing technology.

Data pre-processing techniques generally refer to the addition, deletion, or transformation of training set data. This case is about Cell Segmentation in High-Content Screening. Using High-Content Screening, the medical researchers can measure the cell characteristics from the kinds of samples of number of cells in a living organism or plant. These samples are regarded as the training set that can be used for data pre-processing techniques.

The first process is to do data transformations for individual predictors, which including centering, scaling, and resolving distributional skewness. Centering and scaling are generally used to improve the numerical stability of some calculations. And after skewness transformation, the distribution is not entirely symmetric but these data are better behaved than when they were in the natural units.

Another pre-process is to do data transformations for multiple predictors, using methods to resolve outliers and reduce the dimension of the data. Usually we an identify the outliers on a figure and there are some predictive models resistant to outliers. Data reduction techniques are another class of predictor transformations. These methods reduce the data by generating a smaller set of predictors that seek to capture a majority of the information in the original variables.

Dealing with Missing Values is also important for pre-processing data. It is important to understand why the values are missing. First and foremost, it is important to know if the pattern of missing data is related to the outcome. Missing data can be imputed and imputation has been extensively studied in the statistical literature. One popular technique for imputation is a K-nearest neighbor model. A new sample is imputed by finding the samples in the training set closest to it and averages these nearby points to fill in the value.

Removing predictors is also used to get potential advantages for the data modeling. A rule of thumb for detecting near-zero variance predictors is:

- The fraction of unique values over the sample size is low (say 10%).
- The ratio of the frequency of the most prevalent value to the frequency of the second most prevalent value is large (say around 20).

In addition, collinearity is the technical term for the situation where a pair of predictor variables have a substantial correlation with each other. It is also possible to have relationships between multiple predictors at once (called multicollinearity).

Also adding predictors and binning predictors are also general techniques we used for pre-processing data before modeling. More details are discussing with the case of Cell Segmentation.

1.2 Application2

In "Applied Predictive Modelling" (Kuhn, M. and Johnson, K. (2013).), chapter 12, there is a case that decribes the application of Discriminant Analysis technology.

Discriminant Analysis is one of the techniques of Classification Analysis for Data Mining. Generally it includes the techniques for linear and nonlinear classification models. In this chapter, the case is about Predicting Successful Grant Applications. The data of these applications are from a 2011 Kaggle competition sponsored by the University of Melbourne where there was interest in predicting whether or not a grant application would be accepted. In addition to predicting grant success, the university sought to understand factors that were important in predicting success.

Logistic regression is a very popular model due to its simplicity and ability to make inferential statements about model terms. It is linear in the parameters, and these parameters are obtained by minimizing the sum of the squared residuals. It turns out that the model that minimizes the sum of the squared residuals also produces maximum likelihood estimates of the parameters when it is reasonable to assume that the model residuals follow a normal (i.e., Gaussian) distribution.

Another important technique is called Linear Discriminant Analysis (LDA). This method define a linear discriminant function (may find an optimal discriminant vector) to do analysis and estimation. Examining the coefficients of the linear discriminant function can provide an understanding of the relative importance of predictors. Due to the inherent problem with LDA, as well as its other fundamental requirements, it is recommended that LDA be used on data sets that have at least 510 times more samples than predictors.

The third technique is called Partial Least Squares Discriminant Analysis (PLSDA). For retrospectively or prospectively, measured predictors for any particular problem can be highly correlated or can exceed the number of samples collected. If either of these conditions is true, then the usual LDA approach cannot be directly used to find the optimal discriminant function. So we use PLS for the purpose of discrimination. Applying PLS in the classification setting with a multivariate response has strong mathematical connections to both canonical correlation analysis and LDA.

In addition, Many classification models utilize penalties (or regularization) to improve the fit to the data, such as the lasso. And penalization strategies can be applied to LDA models. The penalized LDA model was applied to the grant data. The software for this model allows the user to specify the number of retained predictors as a tuning parameter. As the penalty increases and predictors are eliminated, performance improves and remains relatively constant until important factors are removed. At this point, performance falls dramatically. As a result of the tuning process, six predictors were used in the model which is competitive to other models.

The nearest-shrunken centroid model (also known as PAM, for predictive analysis for microarrays) is a linear classification model that is well suited for high-dimensional problems. The nearest shrunken centroid method has one tuning parameter: shrinkage. This model works well for problems with a large number of predictors since it has built-in feature selection that is controlled by the shrinkage tuning parameter. Nearest shrunken centroids were originally developed for RNA profiling data, where the number of predictors is large (in the many thousands) and the number of samples is small.

2 Question 2

Sectioning commands. The first one is the

`\section{The Most Important Features}`

command. Below you shall find examples for further sectioning commands:

2.1 Subsection

Subsection text.

2.1.1 Subsubsection

Subsubsection text.

Paragraph Paragraph text.

Subparagraph Subparagraph text.

Select a part of the text then click on the button Emphasize (H!), or Bold (Fs), or Italic (Kt), or Slanted (Kt) to typeset *Emphasize*, **Bold**, *Italics*, *Slanted* texts.

You can also typeset Roman, Sans Serif, SMALL CAPS, and `Typewriter` texts.

You can also apply the special, mathematics only commands $\mathbb{BLACKBOARD}$ $\mathbb{BOLD}$, $\mathcal{CALLIGRAPHIC}$, and $\mathfrak{fraktur}$. Note that blackboard bold and calligraphic are correct only when applied to uppercase letters A through Z.

You can apply the size tags – Format menu, Font size submenu – tiny, scriptsize, footnotesize, small, normalsize, large, Large, LARGE, huge and Huge.

You can use the `\begin{quote} etc. \end{quote}` environment for typesetting short quotations. Select the text then click on Insert, Quotations, Short Quotations:

> The buck stops here. *Harry Truman*
>
> Ask not what your country can do for you; ask what you can do for your country. *John F Kennedy*
>
> I am not a crook. *Richard Nixon*
>
> I did not have sexual relations with that woman, Miss Lewinsky. *Bill Clinton*

The Quotation environment is used for quotations of more than one paragraph. Following is the beginning of *The Jungle Books* by Rudyard Kipling. (You should select the text first then click on Insert, Quotations, Quotation):

> It was seven o'clock of a very warm evening in the Seeonee Hills when Father Wolf woke up from his day's rest, scratched himself, yawned and

> spread out his paws one after the other to get rid of sleepy feeling in their tips. Mother Wolf lay with her big gray nose dropped across her four tumbling, squealing cubs, and the moon shone into the mouth of the cave where they all lived. “*Augrh*” said Father Wolf, “it is time to hunt again.” And he was going to spring down hill when a little shadow with a bushy tail crossed the threshold and whined: “Good luck go with you, O Chief of the Wolves; and good luck and strong white teeth go with the noble children, that they may never forget the hungry in this world.”
>
> It was the jackal—Tabaqui the Dish-licker—and the wolves of India despise Tabaqui because he runs about making mischief, and telling tales, and eating rags and pieces of leather from the village rubbish-heaps. But they are afraid of him too, because Tabaqui, more than any one else in the jungle, is apt to go mad, and then he forgets that he was afraid of anyone, and runs through the forest biting everything in his way.

Use the Verbatim environment if you want LaTeX to preserve spacing, perhaps when including a fragment from a program such as:

```
#include <iostream>         // < > is used for standard libraries.
void main(void)             // ’’main’’ method always called first.
{
 cout << ’’This is a message.’’;
                            // Send to output stream.
}
```

(After selecting the text click on Insert, Code Environments, Code.)

2.2 Mathematics and Text

It holds [1] the following

Theorem 1 *(The Currant minimax principle.) Let T be completely continuous selfadjoint operator in a Hilbert space H. Let n be an arbitrary integer and let $u_1, \ldots, u_{n-1}$ be an arbitrary system of $n-1$ linearly independent elements of H. Denote*

$$\max_{\substack{v\in H, v\neq 0\\ (v,u_1)=0,\ldots,(v,u_n)=0}} \frac{(Tv,v)}{(v,v)} = m(u_1,\ldots,u_{n-1}) \tag{1}$$

Then the n-th eigenvalue of T is equal to the minimum of these maxima, when minimizing over all linearly independent systems $u_1, \ldots u_{n-1}$ in H,

$$\mu_n = \min_{u_1,\ldots,u_{n-1}\in H} m(u_1,\ldots,u_{n-1}) \tag{2}$$

The above equations are automatically numbered as equation (1) and (2).

2.3 List Environments

You can create numbered, bulleted, and description lists using the tag popup at the bottom left of the screen.

1. List item 1
2. List item 2

   (a) A list item under a list item.

   The typeset style for this level is different than the screen style. The screen shows a lower case alphabetic character followed by a period while the typeset style uses a lower case alphabetic character surrounded by parentheses.

   (b) Just another list item under a list item.

      i. Third level list item under a list item.

         A. Fourth and final level of list items allowed.

- Bullet item 1
- Bullet item 2
  - Second level bullet item.
    - Third level bullet item.
      - Fourth (and final) level bullet item.

Description List Each description list item has a term followed by the description of that term. Double click the term box to enter the term, or to change it.

Bunyip Mythical beast of Australian Aboriginal legends.

2.4 Theorem-like Environments

The following theorem-like environments (in alphabetical order) are available in this style.

Acknowledgement 2 *This is an acknowledgement*

Algorithm 3 *This is an algorithm*

Axiom 4 *This is an axiom*

Case 5 *This is a case*

Claim 6 *This is a claim*

Conclusion 7 *This is a conclusion*

Condition 8 *This is a condition*

Conjecture 9 *This is a conjecture*

Corollary 10 *This is a corollary*

Criterion 11 *This is a criterion*

Definition 12 *This is a definition*

Example 13 *This is an example*

Exercise 14 *This is an exercise*

Lemma 15 *This is a lemma*

Proof. This is the proof of the lemma. ■

Notation 16 *This is notation*

Problem 17 *This is a problem*

Proposition 18 *This is a proposition*

Remark 19 *This is a remark*

Solution 20 *This is a solution*

Summary 21 *This is a summary*

Theorem 22 *This is a theorem*

Proof of the Main Theorem. This is the proof. ■

This text is a sample for a short bibliography. You can cite a book by making use of the command `\cite{KarelRektorys}`: [1]. Papers can be cited similarly: [2]. If you want multiple citations to appear in a single set of square brackets you must type all of the citation keys inside a single citation, separating each with a comma. Here is an example: [2, 3, 4].

References

[1] Rektorys, K., *Variational methods in Mathematics, Science and Engineering*, D. Reidel Publishing Company, Dordrecht-Hollanf/Boston-U.S.A., 2th edition, 1975

[2] Bertóti, E.: *On mixed variational formulation of linear elasticity using nonsymmetric stresses and displacements*, International Journal for Numerical Methods in Engineering., **42**, (1997), 561-578.

[3] Szeidl, G.: *Boundary integral equations for plane problems in terms of stress functions of order one*, Journal of Computational and Applied Mechanics, **2**(2), (2001), 237-261.

[4] Carlson D. E.: *On Günther's stress functions for couple stresses*, Quart. Appl. Math., **25**, (1967), 139-146.

A The First Appendix

The appendix fragment is used only once. Subsequent appendices can be created using the Section Section/Body Tag.