Admission Data Prediction Using Machine Learning Methods

Anonymous Author(s)
Affiliation
Address
email

Abstract

The

1 Introduction

NeurIPS requires electronic submissions. The electronic submission site is

`https://cmt3.research.microsoft.com/NeurIPS2020/`

Please read the instructions below carefully and follow them faithfully.

2 Methodology

The dataset we chose is created for prediction of Graduate Admissions from an Indian perspective, which predicting admission from 7 important parameters with 500 students. The output is a number from 0 to 100, which represents the probably a student being admitted. Therefore, we consider it as a regression problem.

2.1 Preprocessing

Firstly, we do the data splitting process, we randomly divide 500 input data into three parts: 320 train data, 80 validation data, and 100 testing data. Secondly, we do subset selection to find the best subset of 7 feature parameters. Based on the RSS loss of linear regression, we find out that the best subset is the total set, we do not need to filter any feature parameter. Then, we normalize the input data before loading them into algorithm models. Additionally, some algorithm may not support regression task, such as logistic regression, LDA, and Naive Bayes, so for these algorithm, we change the regression task into classification task by approximating the output number into 10 neighbor classes: 0, 10, 20, and so on.

2.2 Algorithm

In order to solve this problem, we considering both classification methods and regression methods. The original data is continuous and we will preprocessing it if we want to use classification methods.

Least square Fit a linear model with coefficients $w = (w_1, ..., w_p)$ to minimize the sum of squared residuals between the actual observed data and the predicted data (estimated values) of the data set: $min_w||Xw - y||_2^2$.
Ridge regression Ridge regression solves some problems of ordinary least squares by penalizing the size of the coefficients. What minimizes the ridge coefficiet is the sum of squared residuals with penalties: $min_w||Xw - y||_2^2 + \alpha||w||_2^2$.

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

Lasso regression Lasso regression consists of a linear model with regular terms of l_1-norm. Its minimized objective function is: $min_w \frac{1}{2n_{samples}}||Xw - y||_2^2 + \alpha||w||_1$.

Knn Knn is also a regression method, it is used when the data labels are continuous variables rather than discrete variables. The label assigned to the query point is calculated from the average of its nearest neighbor labels.

Decision tree The nearest neighbor regression is used when the data labels are continuous variables rather than discrete variables. The label assigned to the query point is calculated from the average of its nearest neighbor labels.

SVMIt is very efficient in high-dimensional space, and different kernel functions have a one-to-one correspondence with specific decision functions. Common kernels are already provided, and custom kernels can also be specified.

Boosting The goal of the boosting method is to combine the prediction results of multiple base estimators constructed using a given learning algorithm to obtain better generalization ability/robustness than a single estimator. We mainly focus on Random Forest and AdaBoost.

LDA This is a classification method. It is derived from simple probability models, and these models can be obtained by Bayes' theorem for the relevant distribution $P(X|y = k)$ of each category k.

Naïve Bayes This is a classification method. Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable.

Logistic This is a classification method. Logistic regression is a generalized linear model, so it has many similarities with multiple linear regression analysis. Their model form is basically the same. It gets dependent variable value by logistic function.

3 Experiment

3.1 Preprocessing

In the preprocessing process, after splitting data into training, validation and testing data, and before normalization, we do subset selection to find the best subset of 7 feature parameters. During subset selection, For each $s \in 0, 1, ..., p$, find the subset in size of s that gives lowest RSS, and use cross-validation to esitimate prediction error and select s. Then, we can select the optimal variables. The result is showed below. We can learn from the result that the best subset is the total set, we do not need to filter any feature parameter. In addition, it is quite friedly for us to use the moethod to selection the best subset, since it need a lot of computation and a lot of time when p is too large, but our p is 7, and the dataset is small, so the running time is not too long.

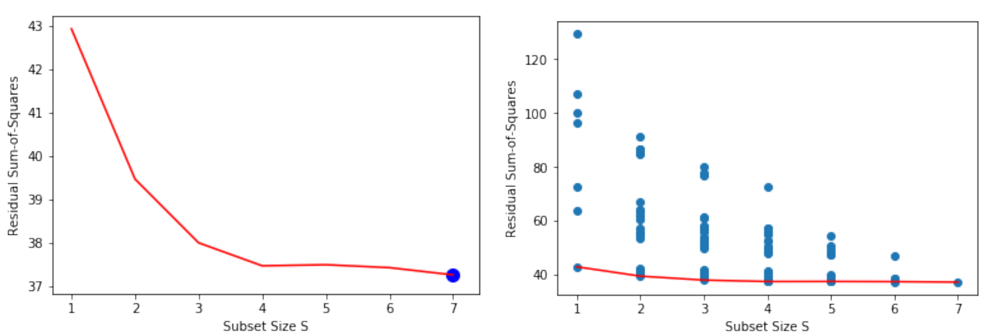

Figure 1: Subset Selection Result.

3.2 Algorithm

3.2.1 Regression

First, we use regression algorithm to fit the admission rate. We use without shrinkage, lasso and ridge models to find the best model, by optimizing the parameter alpha through the analysis of RSS error. Alpha indicates the degree of shrinkage. When alpha approaches 1, it indicates that the degree of shrinkage reaches its maximum; when alpha approaches 0, it indicates that there is no shrinkage. The

RSS of the three methods varies with alpha are shown as follows. It can be seen that the smallest RSS is at lasso regression, and the alpha at this time is 0.3419, the accuracy is 0.8706.

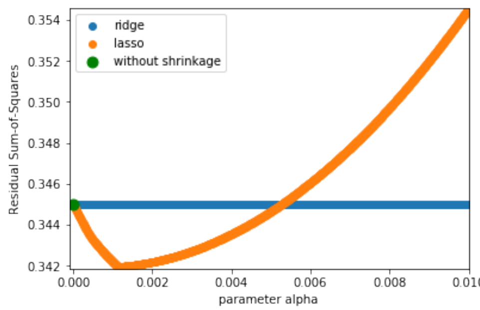

Figure 2: Subset Selection Result.

3.2.2 Decision Tree

In decision tree algorithm, we find the optimal model by finding at which depth we will get the lowest residual sum-of-squares in validation set. From Figure 2 we can know that we can get the lowest residual sum-of-squares at $depth = 4$, and the RSS value is 0.4518. The accuracy of this method is 0.8386.

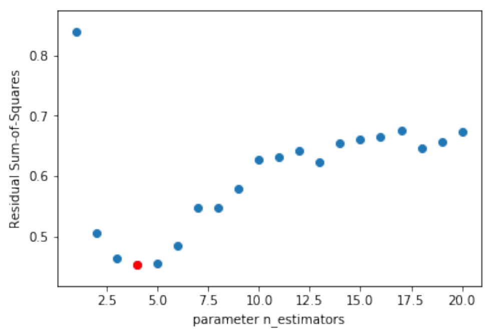

Figure 3: Subset Selection Result.

3.2.3 KNN

Using KNN regression, we need to find the optimal k value. We want to find at which k value we will get the lowest residual sum-of-squares in validation set.

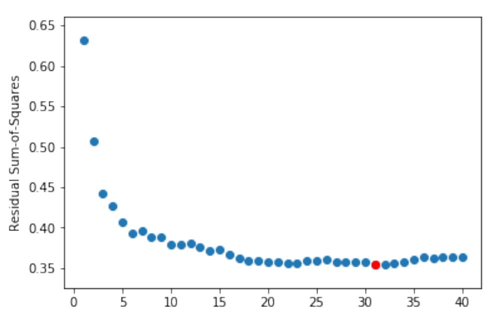

Figure 4: Knn Result.

From Figure 2 we can know that we can get the lowest residual sum-of-squares at $k = 31$, and the RSS value is 0.3546. The accuracy of this method is 0.8706.

3.2.4 SVM

In SVM regression method, we can apply different kernel on it. The min error without any kernel is 0.5596.

- **rbf kernel**: We will find the lowest residual sum-of-squares in validation set at $gamma = 5.0351e - 05$, and the RSS value is 0.4495. The accuracy of this method is 0.6179.

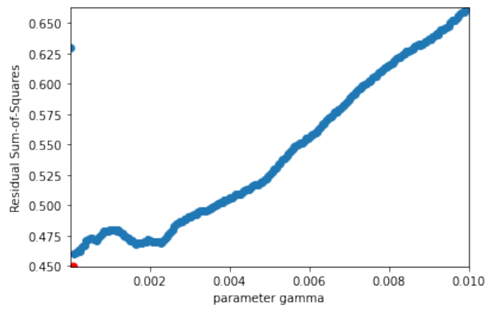

Figure 5: rbf kernel

- **linear kernel**: We will find the lowest residual sum-of-squares in validation set at $C = 0.0918$, and the RSS value is 0.4476. The accuracy of this method is 0.6221.

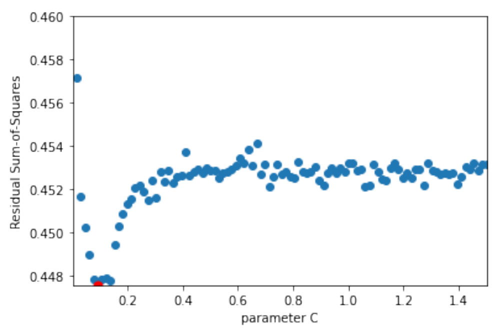

Figure 6: linear kernel

- **poly kernel**: We will find the lowest residual sum-of-squares in validation set at $degree = 1$, and the RSS value is 0.4532. The accuracy of this method is 0.6211.

3.2.5 AdaBoost

When we using AdaBoost method, we are using a lot of weak estimators to regress this problem. So we want to find the optimal number of the weak estimators.
We will find the lowest residual sum-of-squares in validation set at the $n\_estimators = 10$, and the RSS value is 0.3951. The accuracy of this method is 0.6836.
At the same time, we realized that the running time may have some relation with the number of the weak estimators, and then we record the running time of this method with different number of the weak estimators.
The min time is 11035 microseconds at $n\_estimators = 1$, and the running time increases as the number of the weak estimators increases. We can get the conclusion that the optimal RSS may not correspond with the shortest running time.

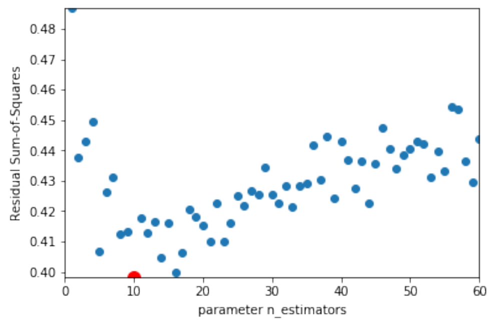

Figure 7: AdaBoost: estimator number and RSS

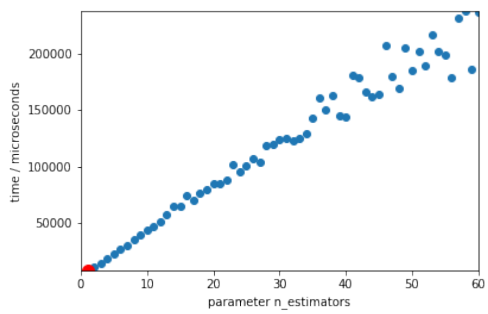

Figure 8: AdaBoost: estimator number and RSS

3.2.6 Random Forest

In random forest algorithm, we need to find the optimal model by finding the optimal n_estimators and optimal depth, by get the lowest residual sum-of-squares in validation set.
Firstly, we choose best n_estimators by running model with diffirent n_estimators and choose the best. But during this process, we find out as n_estimators increase, the time that cost to run the algorithm increase linearly, so its not elegant to choose large n_estimators. After finding the optimal n_estimators, we start finding the optimal depth with optimal n_estimators, the thild figure in Figure 4 shows that at $depth = 4$, the validation error reach maximum. Also, the model start overfitting at around $depth = 3$.
From Figure 3 we can know that we can get the lowest residual sum-of-squares in the validation set at $depth = 4$, $n\_estimators = 13$, and the RSS value is 0.3797. The accuracy of this method is 0.8303.

3.2.7 Other

Besides the regression method above, we also use some classification methods to solve this problem.

- **Logistic** When we try different penalty function, we find that l_1 function is a little bit better than l_2. The lowest residual sum-of-squares in validation set is 0.7880, and the accuracy is 0.4810.
- **Naïve Bayes** Comparing Gaussian NB and Bernoulli NB, we find that Gaussian NB has better effect. The lowest residual sum-of-squares in validation set is 0.7, and the accuracy is 0.5952.
- **LDA** The LDA method with default solver svd can reach the accuracy 0.5833, and the lowest residual sum-of-squares in validation set is 0.5440.

4 Conclusion

These instructions apply to everyone.

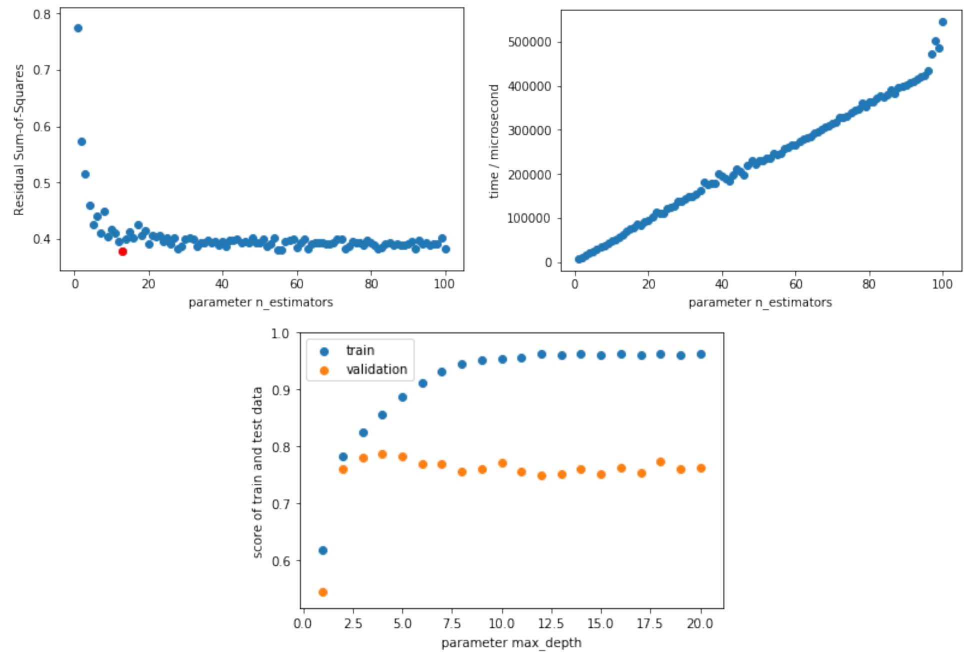

Figure 9: Subset Selection Result.

Table 1: Sample table title

Algorithm	RSS error	Accuracy
regression(Lasso)	**0.3419**	**87.057%**
KNN	**0.3546**	81.570%
Decision Tree	0.4518	**83.863%**
SVM(Linear)	0.4476	62.206%
AdoBoost	0.3982	69.178%
Random Forest	**0.3797**	**83.029%**
LDA	0.5440	58.333%
Naive Bayes(Gaussian)	0.7000	59.5238%
Logistic(l1-penalty)	0.7880	48.0952%

4.1 Citations within the text

The `natbib` package will be loaded for you by default. Citations may be author/year or numeric, as long as you maintain internal consistency. As to the format of the references themselves, any style is acceptable as long as it is used consistently.

The documentation for `natbib` may be found at

`http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf`

Of note is the command `\citet`, which produces citations appropriate for use in inline text. For example,

```
\citet{hasselmo} investigated\dots
```

produces

Hasselmo, et al. (1995) investigated...

If you wish to load the `natbib` package with options, you may add the following before loading the `neurips_2020` package:

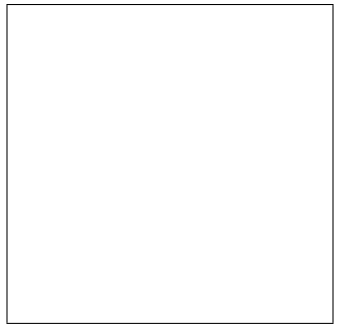

Figure 10: Sample figure caption.

```
\PassOptionsToPackage{options}{natbib}
```

If `natbib` clashes with another package you load, you can add the optional argument `nonatbib` when loading the style file:

```
\usepackage[nonatbib]{neurips_2020}
```

As submission is double blind, refer to your own published work in the third person. That is, use "In the previous work of Jones et al. [4]," not "In our previous work [4]." If you cite your other papers that are not widely available (e.g., a journal paper under review), use anonymous author names in the citation, e.g., an author of the form "A. Anonymous."

4.2 Footnotes

Footnotes should be used sparingly. If you do require a footnote, indicate footnotes with a number[1] in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a horizontal rule of 2 inches (12 picas).

Note that footnotes are properly typeset *after* punctuation marks.[2]

4.3 Figures

All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction. The figure number and caption always appear after the figure. Place one line space before the figure caption and one line space after the figure. The figure caption should be lower case (except for first word and proper nouns); figures are numbered consecutively.

You may use color figures. However, it is best for the figure captions and the paper body to be legible if the paper is printed in either black/white or in color.

4.4 Tables

All tables must be centered, neat, clean and legible. The table number and title always appear before the table. See Table 1.

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

Note that publication-quality tables *do not contain vertical rules.* We strongly suggest the use of the `booktabs` package, which allows for typesetting high-quality, professional tables:

`https://www.ctan.org/pkg/booktabs`

This package was used to typeset Table 1.

[1] Sample of the first footnote.
[2] As in this example.

5 Final instructions

Do not change any aspects of the formatting parameters in the style files. In particular, do not modify the width or length of the rectangle the text should fit into, and do not change font sizes (except perhaps in the **References** section; see below). Please note that pages should be numbered.

6 Preparing PDF files

Please prepare submission files with paper size "US Letter," and not, for example, "A4."

Fonts were the main cause of problems in the past years. Your PDF file must only contain Type 1 or Embedded TrueType fonts. Here are a few instructions to achieve this.

- You should directly generate PDF files using `pdflatex`.
- You can check which fonts a PDF files uses. In Acrobat Reader, select the menu Files>Document Properties>Fonts and select Show All Fonts. You can also use the program `pdffonts` which comes with `xpdf` and is available out-of-the-box on most Linux machines.
- The IEEE has recommendations for generating PDF files whose fonts are also acceptable for NeurIPS. Please see `http://www.emfield.org/icuwb2010/downloads/IEEE-PDF-SpecV32.pdf`
- `xfig` "patterned" shapes are implemented with bitmap fonts. Use "solid" shapes instead.
- The `\bbold` package almost always uses bitmap fonts. You should use the equivalent AMS Fonts:

  ```
  \usepackage{amsfonts}
  ```

  followed by, e.g., `\mathbb{R}`, `\mathbb{N}`, or `\mathbb{C}` for $\mathbb{R}$, $\mathbb{N}$ or $\mathbb{C}$. You can also use the following workaround for reals, natural and complex:

  ```
  \newcommand{\RR}{I\!\!R} %real numbers
  \newcommand{\Nat}{I\!\!N} %natural numbers
  \newcommand{\CC}{I\!\!\!\!C} %complex numbers
  ```

  Note that `amsfonts` is automatically loaded by the `amssymb` package.

If your file contains type 3 fonts or non embedded TrueType fonts, we will ask you to fix it.

6.1 Margins in LaTeX

Most of the margin problems come from figures positioned by hand using `\special` or other commands. We suggest using the command `\includegraphics` from the `graphicx` package. Always specify the figure width as a multiple of the line width as in the example below:

```
\usepackage[pdftex]{graphicx} ...
\includegraphics[width=0.8\linewidth]{myfile.pdf}
```

See Section 4.4 in the graphics bundle documentation (`http://mirrors.ctan.org/macros/latex/required/graphics/grfguide.pdf`)

A number of width problems arise when LaTeX cannot properly hyphenate a line. Please give LaTeX hyphenation hints using the `\-` command when necessary.

References

References follow the acknowledgments. Use unnumbered first-level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to `small` (9 point) when listing the references. **Note that the Reference section does not count towards the eight pages of content that are allowed.**

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer–Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.