Project 2: Data Mining and Prediction

SDS322E

Contents

Modeling

Instructions

General Instructions

Your project will be completed in a single Rmd file called index.Rmd. Clone the following repository, which contains the instructions and template, into your home directory: https://github.com/nathanielwoodward/project2.git

```
cd ~
git clone https://github.com/nathanielwoodward/project2.git
```

When you have finished your project, save all changes and make sure the project knits to HTML correctly.

Submit a link to the Github repository containing your project to Canvas before the due date. Your project should also be live at your_username.github.io/project1 and linked to from the portfolio page on your website. See the last step (6) in this document for how to set this up. Failure to have this done before the deadline will result in a late deduction.

The text of your project document will provide a narrative structure around your code/output. All results presented must have corresponding code. Any answers/results/plots etc. given without the corresponding R code that generated the result will not be considered. Furthermore, all code contained in your final project document must work correctly (knit early, knit often)! Please avoid any extraneous code (code that produces warnings is fine, as long as you understand what the warnings mean!)

Find data:

Find one dataset with at least 5 variables (ideally more!) that you want to use to for mining/classification. At least one should be binary, and at least three should be numeric (taking on more than 10 distinct values). If you can't find a dataset with a binary variable, you can create one by discretizing a numeric variable or collapsing levels of a categorical variable). You will want a minimum of 40 observations (*at least* 5 observations for every explanatory variable you have, ideally 10+ observations/variable). The dataset will need to be tidy, so pick one that is or tidy it before beginning.

It is perfectly fine to use either dataset (or the merged dataset, or a subset of your variables) from Project 1. However, you might wish to take the opportunity to diversify a bit and choose a different dataset to work with (particularly if the variables did not reveal many associations in Project 1 that you want to follow up with). The only requirement/restriction is that you may not use data from any examples we have done in class or lab. It would be a good idea to pick more cohesive data this time around (i.e., variables that you actually thing might have a relationship you would want to examine).

Again, you can use data from anywhere you want (see bottom for resources)! If you want a quick way to see whether a built-in (R) dataset has binary and/or character (i.e., categorical) variables, check out this list: https://vincentarelbundock.github.io/Rdatasets/datasets.html.

Guidelines and Rubric

- **0. Introduction (5 pts)**
  - Introduce your dataset and each of your variables (or just your main variables if you have lots) in a paragraph. Where did you find the data? What are each of the variables measuring? How many observations are there? How many of observations are there per group for your categorical/binary variable(s)?
- **1. Clustering (30 pts)**
  - Perform PAM clustering on at least three of your variables (3 is the bare minimum: using more/all of them will make this much more interesting)! Bonus point for incorporating at least one categorical variable and clustering based on gower dissimilarities.
  - All relevant steps discussed in class (e.g., picking number of clusters based on largest average silhouette width)
  - Visualize the clusters by showing all pairwise combinations of variables colored by cluster assignment (using ggpairs)
  - Include a paragraph or two describing results found, interpreting the clusters in terms of the original variables and observations, discussing goodness of fit of the cluster solution, etc.
- **2. Dimensionality Reduction (20 pts)**
  - Perform PCA on at least three of your numeric variables (3 is the bare minimum: using more/all of them will make this much more interesting)! You can use `eigen()` on the correlation matrix, but `princomp(..., cor=T)` is probably going to be easier.
  - Visualize the observations' PC scores for the PCs you retain (keep at least PC1 and PC2) in ggplot. A biplot with `fviz_pca()` is fine too!
  - Include a paragraph or two describing results with a focus on what it means to score high/low on each PC you retain (interpreting each PC you retained in terms of the original variables/loadings); also discuss how much of the total variance in your dataset is explained by these PCs.
- **3. Linear Classifier and Cross-Validation (20 pts)**
  - Using a linear classifer, (e.g., linear regression, logistic regression, SVM), predict a binary variable (response) from ALL of the rest of the numeric variables in your dataset (if you have 10+, OK to just pick 10).
  - Train the model to the entire dataset and then use it to get predictions for all observations. Run the `class_diag` function or equivalent to get in-sample performance and interpret, including a discussion of how well the model is doing per AUC. Finally, report a confusion matrix.
  - Perform k-fold CV on this same model (fine to use caret). Run the `class_diag` function or equivalent to get out-of-sample performance averaged across your k folds and discuss how well is your model predicting new observations per CV AUC.

  - Discuss the results in a paragraph. How well is your model predicting new observations per CV AUC? Do you see signs of overfitting?

- **3. Non-Parametric Classifier and Cross-Validation (20 pts)**
  - Fit a non-parametric classifier (e.g., k-nearest-neighbors, classification tree) to the exact same dataset/variables you used with the linear classifier (same response variable too).
  - Train the model to the entire dataset and then use it to get predictions for all observations. Run the `class_diag` function or equivalent to get in-sample performance and interpret, including a discussion of how well the model is doing per AUC. Finally, report a confusion matrix.
  - Perform k-fold CV on this same model (fine to use caret). Run the `class_diag` function or equivalent to get out-of-sample performance averaged across your k folds.
  - Discuss the results in a paragraph. How well is your model predicting new observations per CV AUC? Do you see signs of overfitting? How does your nonparametric model compare with the linear model in its cross-validation performance?

- **4.Regression/Prediction (20 pts)**
  - Fit a linear regression model or regression tree to your entire dataset, predicting one of your numeric variables from at least 2 other variables
  - Report the MSE for the overall dataset
  - Perform k-fold CV on this same model (fine to use caret). Calculate the average MSE across your k testing folds.
  - Does this model show signs of overfitting? Discussion the results in a paragraph

- **5. (10 pts) Python and Reticulate**
  - Include a python code chunk in your project
  - Using `reticulate`, demonstrate how you can share objects between R and python using `r.` and `py$`
  - Include a sentence or two describing what was done

- **6. GitHub and GitHub Pages**
  - Push your cloned project2 folder up to your GitHub. It should contain index.Rmd (your finished project code), the knitted html file (index.html) and optionally, the instructions files. Here's how:

1. Create a remote repository (on your GitHub) called project2.

2. In the terminal, run the following code, changing your username accordingly

```
cd ~/project2 #cd to your project2 directory
git add .
git commit -m "message"
git remote set-url origin https://github.com/YOUR_USERNAME/project1.git
git push origin main
```

- Once it is up there, go to your remote repository and then to Settings > Options and scroll down to GitHub pages. Alternately, go to https://github.com/YOUR_USERNAME/project2/settings/pages.
  - Then, select the main branch and click save. Finally, go back to your homepage repository, go to the Portfolio.Rmd page, and update the link for project2 to YOUR_USERNAME.github.io/project2. Add, commit, and push your updated homepage to GitHub for deployment.
  - Your project MUST be live at YOUR_USERNAME.github.io/project2 and be accessible from your portfolio page to receive a grade. Failure to do this before the due date will result in a late deduction!

Where do I find data again?

You can choose ANY datasets you want that meet the above criteria for variables and observations. You can make it as serious as you want, or not, but keep in mind that you will be incorporating this project into a portfolio webpage for your final in this course, so choose something that really reflects who you are, or something that you feel will advance you in the direction you hope to move career-wise, or something that you think is really neat, or whatever. On the flip side, regardless of what you pick, you will be performing all the same tasks, so it doesn't end up being that big of a deal.

If you are totally clueless and have no direction at all, log into the server and type

```
data(package = .packages(all.available = TRUE))
```

This will print out a list of **ALL datasets in ALL packages** installed on the server (a ton)! Scroll until your eyes bleed! Actually, do not scroll that much... To start with something more manageable, just run the command on your own computer, or just run `data()` to bring up the datasets in your current environment. To read more about a dataset, do `?packagename::datasetname`.

If it is easier for you, and in case you don't have many packages installed, a list of R datasets from a few common packages (also downloadable in CSV format) is given at the following website: https://vincentarelbundock.github.io/Rdatasets/datasets.html.

- A good package to download for fun/relevant data is `fivethiryeight`. Run `install.packages("fivethirtyeight")`, load the packages with `library(fivethirtyeight)`, run `data()`, and then scroll down to view the datasets. Here is an online list of all 127 datasets (with links to the 538 articles). Lots of sports, politics, current events, etc.
- If you have already started to specialize (e.g., ecology, epidemiology) you might look at discipline-specific R packages (vegan, epi, respectively). We will be using some tools from these packages later in the course, but they come with lots of data too, which you can explore according to the directions above
- However, you *emphatically DO NOT* have to use datasets available via R packages! In fact, I would much prefer it if you found the data from completely separate sources and brought them together (a much more realistic experience in the real world)! You can even reuse data from your SDS328M project, provided it shares a variable in common with other data which allows you to merge the two together (e.g., if you still had the timestamp, you could look up the weather that day: https://www.wunderground.com/history/). If you work in a research lab or have access to old data, you could potentially merge it with new data from your lab!
- Here is a curated list of interesting datasets (read-only spreadsheet format): https://docs.google.com/spreadsheets/d/1wZhPLMCHKJvwOkP4juclhjFgqIY8fQFMemwKL2c64vk/edit
- Here is another great compilation of datasets: https://github.com/rfordatascience/tidytuesday
- Here is the UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/index.php
  - See also https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research#Biological_data
- Here is another good general place to look: https://www.kaggle.com/datasets
- To help narrow your search down or to see interesting variable ideas, check out https://www.tylervigen.com/spurious-correlations. This is the spurious correlations website, and it is fun, but if you look at the bottom of each plot you will see sources for the data. This is a good place to find very general data (or at least get a sense of where you can scrape data together from)!
- If you are interested in medical data, check out www.countyhealthrankings.org
- If you are interested in scraping UT data, they make *loads* of data public (e.g., beyond just professor CVs and syllabi). Check out all the data that is available in the statistical handbooks: https://reports.utexas.edu/statistical-handbook

Broader data sources: Data.gov 186,000+ datasets!

Social Explorer is a nice interface to Census and American Community Survey data (more user-friendly than the government sites). May need to sign up for a free trial.

U.S. Bureau of Labor Statistics

U.S. Census Bureau

Gapminder, data about the world.

. . .