Homework 2: Smoothers, Generalized Additive Models, and Storytelling

Harvard CS 109B, Spring 2017

Feb 2017

Problem 1: Heart Disease Diagnosis

In this problem, the task is to build a model that can diagnose heart disease for a patient presented with chest pain. The data set is provided in the files `dataset_1_train.txt` and `dataset_1_test.txt`, and contains 6 predictors for each patient, along with the diagnosis from a medical professional.

- By visual inspection, do you find that the predictors are good indicators of heart disease in a patient?
- Apply the generalized additive model (GAM) method to fit a binary classification model to the training set and report its classification accuracy on the test set. You may use a smoothing spline basis function wherever relevant, with the smoothing parameter tuned using cross-validation on the training set. Would you be able to apply the smoothing spline basis to categorical predictors? Is there a difference in the way you would handle categorical attributes in `R` compared to `sklearn` in `Python`?
- Plot the smooth of each predictor for the fitted GAM. By visual inspection, do you find any benefit in modeling the numerical predictors using smoothing splines?
- Using a likelihood ratio test, compare the fitted GAM with the following models: (i) a GAM with only the intercept term; (ii) a GAM with only categorical predictors; and (iii) a GAM with all predictors entered linearly.

Hints: You may use the function `gam` in the `gam` library to fit GAM with binary responses. Do not forget to set the attribute `family = binomial(link="logit")`. The `plot` function can be used to visualize the local models fitted by GAM on each predictor. You may use the `anova` function (with attribute `test="Chi"`) to compare two models using a likelihood ratio test.

You may use the following sample code for cross-validation:

```
library(boot)

# Function to compute k-fold cross-validation accuracy for a given classification model
cv_accuracy = function(model, data, k) {
  # Input:
  #   'model' - a fitted classification model
  #   'data' - data frame with training data set used to fit the model
  #   'k' - number of folds for CV
  # Output:
  #   'cv_accuracy' - cross-validation accuracy for the model

  acc <- 1 - cv.glm(data, model, K = k)$delta[1]
  return(acc)
}
```

Problem 2: The Malaria Report

You work for the Gotham Times media organization and have been tasked to write a short report on the World Health Organisation's (WHO) fight against malaria. The WHO Global Malaria Programme (http://www.who.int/malaria/en/) has been working to eliminate the deadly disease over the past several decades, your job is to discuss their work and spotlight the impact they've had. Your writing and graphics should be easily understood by anyone interested in the topic, and not necessarily just physicians and experts.

Key Facts and Quotes on Malaria

Here are some informative key facts and quotes about Malaria that you may want to include in your report:

- RISK: About 3.2 billion people – almost half of the world's population – are at risk of malaria.
- CASES: 214 million malaria cases reported worldwide in 2015.
- INCIDENCE: 37% global decrease in malaria incidence between 2000 and 2015.
- MORTALITY: 60% decrease in global malaria mortality rates between 2000 and 2015.
- "Malaria is a life-threatening disease caused by parasites that are transmitted to people through the bites of infected female mosquitoes."
- "Young children, pregnant women and non-immune travelers from malaria-free areas are particularly vulnerable to the disease when they become infected."
- "Malaria is preventable and curable, and increased efforts are dramatically reducing the malaria burden in many places."

Many of these facts were pulled from the WHO website, where you can find many more.

The Data

The datasets consist of country-level information for 2015, estimated malaria cases over time, and funding values and sources over time.

Dataset 1: data/global-malaria-2015.csv This dataset contains observed and suspected malaria cases as well as other detailed country-level information **for 2015** in 100 countries worldwide. The CSV file consists of the following fields:

- WHO_region, Country, Country Code, UN_population
- At_risk - % of population at risk
- At_high_risk - % of population at high risk
- Suspected_malaria_cases
- Malaria_cases - actual diagnosed cases

Dataset 2: data/global-malaria-2000-2013.csv This dataset contains information about suspected number of malaria cases in the same 100 countries for the years 2000, 2005, 2010, 2013.

Dataset 3: data/global-funding.csv This dataset contains the total funding for malaria control and elimination (in millions USD) provided by donor governments, multilateral organizations, and domestic sources between 2005 and 2013.

Dataset 4: data/africa.topo.json The TopoJSON file (extension of GeoJSON) contains the data of the boundaries for the African countries.

You can also explore the very large database provided by the WHO, though a bit of manual processing may be needed.

Exploratory Data Analysis

Your first task is to use this data for exploratory data analysis (EDA) and to create several visualizations (e.g., bar graphs, line charts, scatterplots, maps) using Tableau or ggplot2. It may also be useful to reshape the data in R before visualizing it – take a look at the R code at the end of this `Rmd` document.

You will notice some regional discrepancies, keep these in mind as you explore the data and gain an understanding of what the data are saying. Try to identify a few key messages that can be supported by the data.

Planning

In planning your report, think about the many facets of information contained in the data, and explore which of the visualziations are most effective to illustrate some key messages. You are free to decide what data you want to use and what kind of story you would like to tell. You may use statistical modeling techniques (e.g., linear regression) if you think they are appropriate, but must explain and justify their use. Consider the visualization and storytelling principles we discussed in class.

Your Report

Your report should have a catchy title and be 500-600 words in length (not more!) with at least two different visualizations, including titles and captions. Structure your report using balanced design and make sure to start with a global overview with context, need, task, and message. The text and visualizations should mutually reinforce your messages through effective redunancy.

Your visualizations must be effective and well designed. Push yourself a little. If you are a Tableau/ggplot2 beginner edit them carefully based on the principles we discussed in class. If you are a Tableau/ggplot2 expert try to do something special that goes beyond the usual graphs and charts.

Submission

Your report can be written in your tool of choice (Word, Google Docs, Markdown, etc.) and should look professionally designed. Upload the report as a PDF into the homework folder. In addition, submit a separate PDF with **all** of the EDA visualizations you created and any code you used.

Grading Criteria

We will grade your report using the following criteria:

- Is the report informative, accurate, and engaging?
- Are the messages clearly expressed?
- Is the writing appropriate for the target audience, the readers of the Gotham Times?
- Are the visualizations effective, accurate, and easy to understand?
- Are the scales, axis, labels, and color maps appropriate?
- Do the titles and captions for the visualizations tell the reader what the message of each visualization is?
- Does the report accurately highlight some interesting facts about the data?
- Does the report follow the storytelling principles discussed in class?