IMT574 Problem Set 4

February 6, 2020

Introduction

This problem set should not be hard per se but it may be very slow in terms of computing. You are mostly asked just to use existing packages to analyze a big survey dataset.

The problem set has two parts: first you try to get as good a prediction as you can using k-NN, logistic regression and SVM methods, and thereafter you analyze how much does knowledge of country of the respondent improve our prediction.

The problem set has 4 goals: a) explore classification methods we have learned in a novel context (a large survey) b) learn more about cross validation c) learn a little bit about the global opinion (but we don't do the latter that rigorously).

World Values Survey

In this database we use World Values Survey data. The data is free to be downloaded from the webpage, just you have to sign up but I expect to download the version on canvas. It is a survey, conducted every few years in a number of countries. Here we use wave 6 data, mostly from 2013-2014. Note that not all countries are participating in each wave.

The questions revolve around different opinion topics, including trust, work, religion, family, gender equality, and nationalism. In this problem set we focus on what the respondents think about abortion: "Please tell if abortion can always be justified, never be justified, or something in between". The responses range between 1 – never justifiable, and 10 – always justifiable. Besides of the numeric range 1..10, a number of cases have negative codes (this applies to many variables). These are various types of missing information (-5: missing, -4: not asked, -3: not applicable, -2: no answer, -1: don't know). We treat all these as just missing below.

The version we use here is a little bit simplified, I have removed a large number of variables that are constructed from the other variables and hence highly collinear with the rest of the data.

I strongly recommend you to browse the documentation before you start, there are two large-ish documentation files provided.

1 Explore and prepare the data (20pt)

As the first step, explore the data.

1. (2pt) Load the data. How many responses and variables do we have?
2. (3pt) Create a summary table over all responses for *V204*: is abortion justifiable. How many non-missing responses (i.e. positive answers) do you find? Describe the the opinion about the abortion among the global pool of respondents.
3. (4pt) Now remove missings. We do it in two ways:

(a) remove everything that are not positive integers for *V204* and *V2* (country).

(b) for all other variables, remove the missings in the sense of missing value on computer. You may leave negative answers in the data, otherwise I am afraid your sample size collapses.

What is the final number of observations?

4. (2pt) In order to simplify the analysis below, create a new binary variable *abortion* as

$$\text{abortion} = \begin{cases} 1 & \text{V204} > 3 \\ 0 & \text{otherwise} \end{cases}$$

5. (5pt) Compute (pearson) correlation table between *abortion* and all other variables in the data. There are many of these!

   Present these variables in descending order according to the absolute value of the correlation. It might look something like:

variable	correlation
abortion	1.000
x1	0.777
x2	-0.666
...	
x33	0.020
x44	-0.011

   Take a look at a few variables that have strong correlation with abortion. What do these represent?

6. (4pt) convert country code *V2* into dummies. First rename *V2* to *country*. Thereafter use `pd.get_dummies` along these lines:

```
data2 = pd.get_dummies(data, columns = ['country'])
```

   Afterwards, remove *country* variable from the data. How many rows/columns do you have now? How many country dummies does the data contain?

   Note that `get_dummies` creates a dummy for every category, so you have to remove one of these dummies in order to avoid perfect multicollinearity.

2 Implement Cross-Validation (40pt)

Now it's time to write your own code that does k-fold CV. I recommend to go the following path:

1. (3pt) Make it as a function that takes `k`, the (unfitted) model, features `X` and the target `y` as arguments.

2. (10pt) Next, one should randomly shuffle the data. However, it is easier to generate a list of indices, and shuffle those randomly.

3. (25pt) Loop the following `k` times

   (a) Select every `k`-th of your indices for validation data

   (b) For training data, select all indices, except those that went into validation data. Hint: check out set operations

(c) Separate the data X and the target **y** into training/validation parts.

(d) Fit the model on training data

(e) Predict outcome on validation data

(f) Compute the resulting statistic (you may compute more than one).

4. (2pt) finally, return mean of the statistics.

Note: This is my suggested path but you may follow another one.

3 Find the best model (40)

In this section your task is to find which model: k-NN, logistic regression, or SVM works best. You will evaluate the model performance using 5-fold cross-validation with accuracy and F-score as the metric. And unlike in all your future work, here *you will use your own CV impementation*!

k-NN and SVM are sensitive to the distance metric, so you may also try to normalized versus non-normalized features. Check out `sklearn.preprocessing.normalize`. Logistic regression is agnostic with respect to the metric, but may benefit from more similar variable values for numerical reasons.

Some of the methods (k-NN, SVM) are slow to compute, so you may start with a subset of data (say, 5000 random lines only). If everything turns out fine, you increase the data size as far as your computer can go.

3.1 k-NN (13pt)

First, use k-NN and experiment with a few different k-s.

1. (2pt) Separate your training data into X (features), and **y** (target). Target will be the *abortion* variable, X are all the other features.

2. (2pt) pick a k and set up the k-NN model. Use your freshly-minted CV routine to cross-validate *accuracy* and F-score of your k-NN model.

3. (5pt) Try a few different k-NN models (pick different k, choose to normalize/not-to-normalize your features).

4. (4pt) Present the results from your best k-NN model. Note: as you are using two metrics here, you may end up with different models performing better according to different measures.

3.2 Logistic Regression (9pt)

1. Now repeat the process above with logistic regression. As we have a myriad of features anyway, we are not going to do any feature engineering. Just a plain logistic regression.

3.3 SVM (15pt)

Now repeat the process with support vector machines while choosing between a few different kernels and kernel options, such as *degree* for polynomial kernels.

Hint: I have mixed experience with *sklearn* version of SVM. I recommend to limit the number of iterations, initially maybe to just 1000, in order to ensure your model actually terminates.

1. (14pt) pick a kernel and repeat the process above.

   Note that some kernels are slower than others, so be careful.

2. (2pt) If your models worked like mine, you may have noticed that while accuracy seems all right, precision and recall are rather low. Explain what does such a phenomenon mean.

3.4 Compare the models (3pt)

1. (2pt) Finally, compare the models. Which ones performed the best in terms of accuracy? Which ones in terms of F-score? Did you encounter other kind of issues with certain models? Which models were fast and which ones slow?

2. (1pt) If you have to repeat the exercise with a single model (and you have, see below), which one will you pick?

4 How large a role does country play? (20pt)

Here we switch from machine learning to social sciences. Public opinion differs from country to country, but also inside the countries. Does the fact that we include country code in data help us to substantially improve the predictions?

You pick the best ML method from above. You estimate two sets of models: one with country information included, and one where it is removed. Is the former noticeably better than the latter?

1. (10pt) Pick your best ML method based you designed above. Cross-validate the accuracy of abortion variable using all the features, including country dummies and report the accuracy. Essentially you repeat here what you did above, so you can also just copy the result from above.

2. (15pt) Now remove all the country dummies, but keep the other variables intact. And repeat.

3. (5pt) Comment what you found. Does country information help to noticeably improve the prediction?