Machine Learning
Project

Predicting Percentage Prescribed Drugs based on Gender and Years of Experience of Doctor

L. Dickhoff, M. Hoefsloot, F. Raaijmakers

23 December 2016
Amsterdam University College

Problem

The data set we chose provided a lot of information: it held information for about 240,000 practicing doctors[1]. For each doctor, the data included fields for the gender of the doctor, the region in which they work, whether the doctor works in a urban or non-urban environment, the speciality of the doctor and the amount of years the doctor has been practicing their occupation. Additionally, for each doctor, the data set included a list of various drugs prescribed by that doctor and a number to reflect the amount of times which that doctor prescribed the specific drug. Hence, we had to focus on a few features and we we decided that the most interesting ones are the gender and the years of practice of the doctors. Moreover, the numerous different drugs in the original dataset is overwhelming, so in order to be able to go more in depth, we decided to concentrate on comparing two drugs of the same kind (i.e. painkillers), and thus pick a weak one and a stronger one. Finally, the goal here is to perform classification and regression algorithms on this (reduced amount of) data, plot the results, and see if there are any conclusions to be made.

Please find below a bar representation of the total amount of female and male physicians in the whole dataset (with the number of physicians on the y-axis):

[1]The data set: Kaggle prescription based prediction

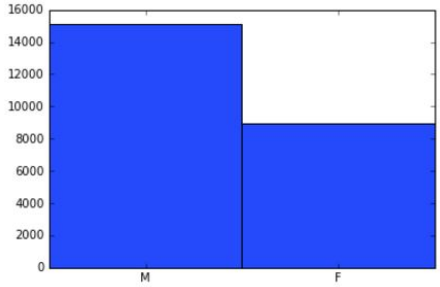

Figure 1: Distribution of male and female physicians

As such, the main question this paper aims to answer is the following:

> Can we observe any trends in the amount of strong vs weak drugs the physicians prescribe according to their gender and their experience?

Approach

As the extensive collection of information provided by our data set was too large for our research, we decided to make a selection of the features given and a selection of the drugs that are prescribed by the physicians. To answer our research question, we only needed information about the gender of the doctors, the years that they have been practicing their occupation and the amount of times they have prescribed a specific drug.

In the data set, the drugs are indicated by their professional name, which made them difficult to distinguish for the common masses. Hence, we decided to select a couple of medicines that would allow for comparison and from which we could possibly draw conclusions. We did this by choosing three types of drugs based on the frequency of their prescription and on the fact that there a weak and a stronger version of the same type occur in the data set. This lead to the choice of two antibiotics, two weak painkillers and two strong painkillers. The two antibiotics we choose to research are Amoxicillin and Penicillin V Potassium. Amoxicillin is a partially synthetic drug while Penicillin is fully synthetic. Some people say that Penicillin is slightly stronger[2]. These two medicines are quite commonly prescribed in various specialities and their effects are quite similar. Therefore, we are interested to find a reason behind choosing one drug over the other.

The two weak painkillers are Ibuprofen and Codeine, of which Codeine is slightly stronger. Both Ibuprofen and Codeine have severe repercussions when taken in strong doses or when combined with other drugs. We think it would be interesting which one of the two reasonably weak painkillers is the "lesser of two evils".

[2]https://www.drugs.com/answers/amoxicillin-caps-500mg-penicillin-500mg-743719.html

Amongst the strongest painkillers available are Morphine, Amitriptyline HCL and Gabapentin [3]. The latter two are very similar and are regarded as having the same strength, therefore we choose to regard Amitriptyline HCL and Gabapentin as one variable. These medicines are used in cases of severe pain as well as dealing with epilepsy and anxiety. Various versions of Morphine exist, all of which are closely related, so we choose to put all different types of Morphine together as one variable. All these drugs are quite common and often also available simply over the counter, however, as the drugs have different effects, it is important that the right drug is used in the right situation. Therefore, it is important that there are no biases in prescribing these drugs based on variables such as gender and the years of experience.

Choices made and justifications

To be able to run our program in a reasonable time, we decided to scale down our data set. After some manual inspection of the data set, we could not find any variable on which the data set was sorted, therefore we decided that we did not need to do a random selection of the data points we were going to use. We simply took the first 24,000 data points to work with, as that was about 10 % of the original data set provided. Additionally, we decided to make a new data set for each type of drug, based on whether a doctor has prescribed at least one of the two drugs of the same type. We used these three data sets for our prediction algorithms as they take away a large portion of the data points that consist out of merely zeros. This gives us a more distinguished trendline.

To research the possible relationship between the different kinds of characteristics of the doctors and the drugs they prescribe, we applied several learning algorithms provided by the SK-learn module. For the question if the prescription of a certain drug depends on the years of experience a doctor has, we need algorithms of the type regression, as we are working with a non-binary variable. However, for the question if the prescription of a certain drug depends on a binary variable such as the gender of a physician, we chose to work with a classification algorithm.

The first algorithm we applied was isotonic regression. This algorithm finds a non-decreasing approximation of a function while minimizing the mean squared error on the training data. The cost function as we know it is based on the calculation of the mean squared error, so this means that the isotonic regression function evaluates its own approximation as it implements the minimization of the cost function we are accustomed to. The benefit of a function like this is that it does not assume a certain target function like linear regression does, it is more flexible and therefore it molds itself to the data set as is visible in the image on the right. We stress this difference between linear and isotonic

[3] http://iytmed.com/list-of-painkillers/

regression by implementing the linear regression, too. We run both algorithms on the same data set as that will give the best comparison of the two learning curves. In both cases we chose to plot the two drugs of the same kind in one graph as this allows for a good comparison of the percentages of times that they are prescribed. Initially, the isotonic regression plot connected all the data points in order of the original data set, so we were not able to read the plot. We chose to sort the data points on increasing years of practice of the doctor, which eliminated the scattering of the isotonic regression trend.

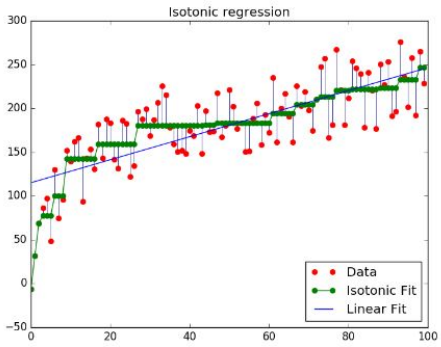

Figure 2: Isotonic Regression

The third learning algorithm we implemented is of a different kind: it's the k-Nearest Neighbour algorithm, which is a classification algorithm instead of a regression algorithm. This allows us to find a decision boundary to be able to group the data points into the male-female distinction we are looking for. In this case, we plot all different medicines separately to be able to recognize which female or male doctors prescribe which type of drug.

Results

Isotonic and Linear regression

All throughout our results, it is clear that of two drugs, the weaker drug of the two is prescribed on a more steady basis than the stronger drug. Both the isotonic and the linear regression trends are more or less horizontal for the weaker version of a certain type of drug. However, the stronger kind increases in percentage as a doctor's years of practicing increases, this shows that the more experienced the physicians are, the more they tend to prescribe stronger drugs.

As we expected, the linear regression produces a linear trend, but the isotonic regression has slight fluctuations. The main characteristic of isotonic regression is the fact that it divides the data points into smaller segments and creates a linear line to go through the data points of that smaller set. This is visible in the plots as the isotonic fits show slight angles in the trendlines. We think that the isotonic regression was a good choice for our case as we have a continuous but

discrete variable, this allows for division into different segments. The variable "years of practice" only has 8 possible values, so a regression

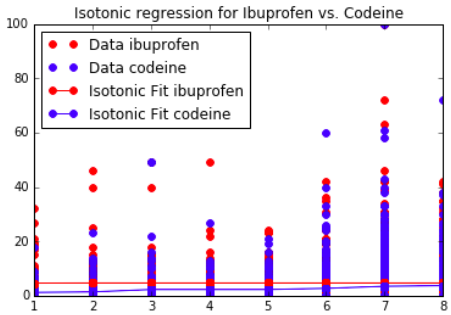

Figure 3: Isotonic Regression Ibuprofen versus Codeine

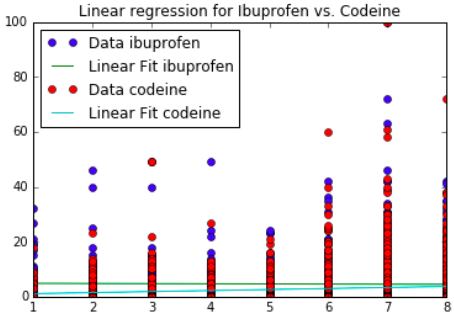

Figure 4: Linear Regression Ibuprofen versus Codeine

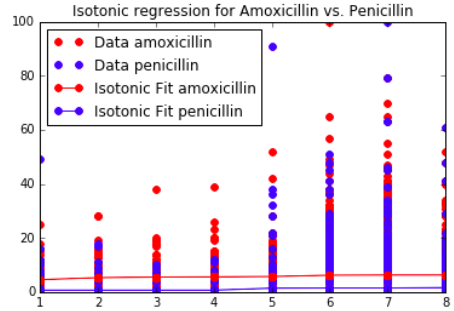

Figure 5: Isotonic Regression Amoxicillin versus Penicillin

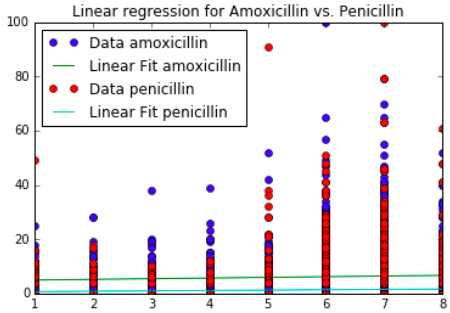

Figure 6: Linear Regression Amoxicillin versus Penicillin

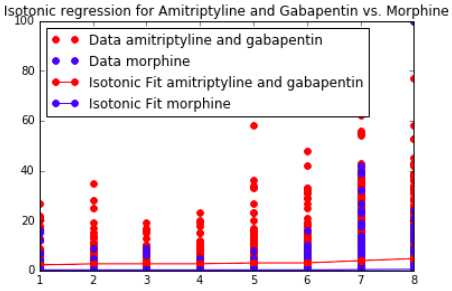

Figure 7: Isotonic Regression Amitriptyline Gabapentin versus Morphine

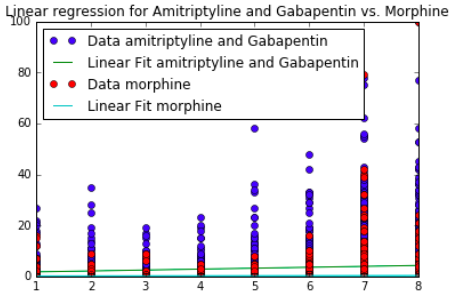

Figure 8: Linear Regression Amitriptyline Gabapentin versus Morphine

K-Neighbors algorithm

In the plots below, the blue coloured clusters represent 'male' and the brown coloured ones represent 'female'. For this analysis we adopted a k value of three. Table 1 displays the precision, recall and f-scores for each individual drug that we tested. As one can easily conclude, the values are high above average, therefore our method is successful in classifying the data based on our parameters. The precision for Amoxicillin for example is as high as 71%. When we compare the plots we observe that the majority of the regions are blue coloured. This is not surprising given the distortion in the male-female balance in our data. Nevertheless, for some drugs there are clear distinctions

present. Take for example Figure 9. We see that there is a noticable divergence between the point that prescribe high percentage of drugs and low percentage; the high values are classified as male no matter the years of experience and the low percentage are (mostly) classified as female. Throughout all the graphs we can make out that there is mostly a horizontal divergence, thus we can conclude that years of experience plays a minimal role in determining the percentage prescribed.

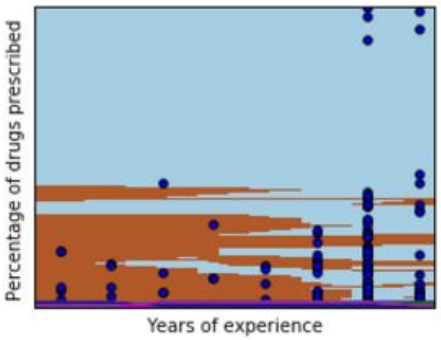

Figure 9: Amitriptyline Gabapentin

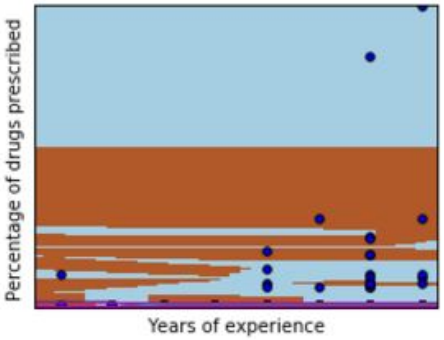

Figure 10: Morphine

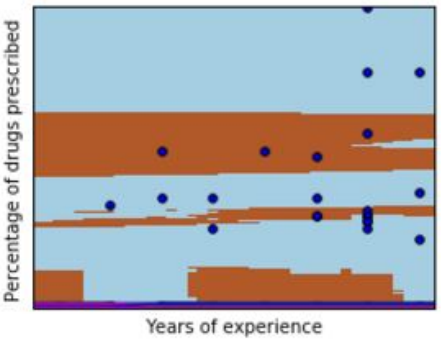

Figure 11: Codeine

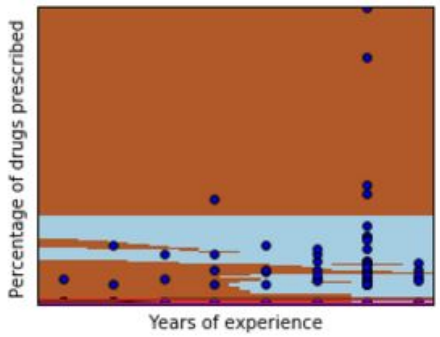

Figure 12: Ibuprofen

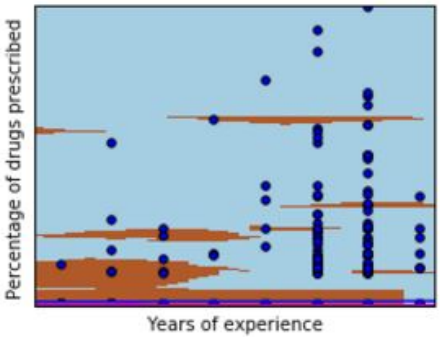

Figure 13: Amoxicillin

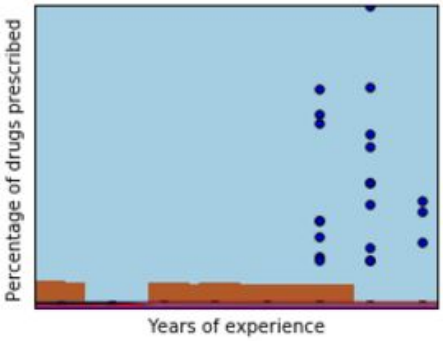

Figure 14: Penicillin

.

Table 1: Results K-Neighbors

Total	Precision	Recall	f-score
Amoxicillin	0.71	0.70	0.67
Penicillin	0.63	0.64	0.63
Ibuprofen	0.69	0.68	0.64
Codeine	0.69	0.68	0.65
Morphine	0.69	0.69	0.66
Amitriptyline Gabapentin	0.71	0.70	0.68

Evaluation

In order to make a well advised decision in determining our machine learning methods, our team did some preliminary study on the data. Incorporating both our hypothesis on the data as well as our preparatory analysis, we determined our methods; isotonic regression and k-neighbors clustering. The results of the isotonic regression highlight the limitations of making educated conclusions based on basic plots and intuitive knowledge. Even though isotonic regression did not discover a trend for the relation between the independent variables 'Years practicing', 'Gender' and the dependent variable 'Percentage drugs prescribed', it does certainly not imply that no trend exist between these variables. This is why we evaluate multiple methods, hence from these results we conclude that based on our research there is no significant regression between the aforementioned variables. On the other hand, the limitations might also lay in the composition of our trained data. We selected a tenth of the total data by taking the first 10% of all doctors in our data set. Even though our data set did not specify that the data has been ordered in any manner, we might have unintentionally selected a sample that did not (in any way) accurately represent our total data pool. Moreover, the composition of the data pool did not represent every category evenly. As an example, the total data pool consists of significantly more males. Furthermore, the within the skewed male-female ratio another distorted trend is hidden; namely that there are a lot more females with little years of practice compared to males, and vice versa a lot more males in our data set have a high number of years of experience. This may have affected the learning for our data as not every combination is accurately represented.

Discarding the possible imperfections in our data, we will continue to analyze our second method of analysis; k-neighbor clustering. Since we used the same data set, the limitations mentioned previously apply here as well. Nevertheless, k-neighboring clustering has demonstrated to be a better method of analysis

than isotonic regression. From the table we can conclude that for every drug the precision was roughly equal. From the plot we can make-out that the clusters are not as clear-cut as we hypothesised, nevertheless we can conclude that there are clear clusters present in our data and thus there is a bias in what percentage drug is prescribed.