Results EDA

Rienk Heins

9/28/2021

```
# Data is loaded from the file
Data <- read.table(file = "breast-cancer-wisconsin.data", sep = ",", na.strings = "?")
# Row names are taken from the names file of the Wisconsin data set and added to the data frame
row_names <- c("ID", "Clump_Thickness", "Uniformity_of_Cell_Size", "Uniformity_of_Cell_Shape", "Marginal
names(Data) <- row_names
# Labels are changed to benign and malignant instead of the 2 and 4 used in the data set for clarity an
Class_numeric <- Data$Class
Data$Class <- factor(Data$Class, labels = c("Benign", "Malignant"))
# Mitosis column gets removed
cleanData <- Data[,-10]
# All NA values are removed
cleanData <- cleanData[complete.cases(cleanData),]
```

Results

In the following section the results of the data interpretation steps and the cleaning of the data will be discussed. First the cleaning of the data will be reviewed.

Cleaning of the data

The only attribute that was directly removed is the mitosis attribute. As it has a fairly low correlation rating of only 0.4, which is especially low compared to the fact that all the other attributes score at least 0.69 with some around or above a score of 0.8. Furthermore the summary showed that most data points in this attribute have a score of 1 on the scale from 1 to 10, as the 1st quartile, median and 3rd quartile as well as the mean all lay at 1.

Furthermore as seen in figure 1, the mitosis attribute also doesn't have a split in it's boxes of benign and malignant cases, rather they overlap at 1. This means that there are both a lot of malignant and benign cases at a score of 1. Given this fact mitosis can never be a good predictor as most cases have a score of 1, yet this score gives almost no information about the malignancy of the tumor cell. There are even some quite high outliers for benign cases in mitosis, meaning that even a high score might not directly be a malignant case. Knowing all this it was decided that the mitosis case should be removed.

The next and last thing that was removed from the data set were the NA cases. There only were 16 NA cases in the data set and they all fell in the bare nuclei attribute. No real reason for these was given by the producers of the data and there are no attributes that might help to give a good score to them through calculation, knowing this it was decided that the patient data with these NA's will be removed. As this still leaves 683 cases which can be used which should be enough to give a good machine learning result.

```
# Data is changed to long data to be used in the multiple box plot
long_data <- pivot_longer(data = Data[,c(2:11)], !Class, names_to = "Attribute", values_to = "Count")
# Multiple boxplot is created using ggplot
ggplot(long_data, aes(y = Count, x = Attribute, fill = Class)) +
  geom_boxplot(width = 0.5, outlier.size = 0.5) +
  ggtitle("Multiple boxplot of all attributes") +
  xlab("Attributes") +
  ylab("Grades(1-10)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

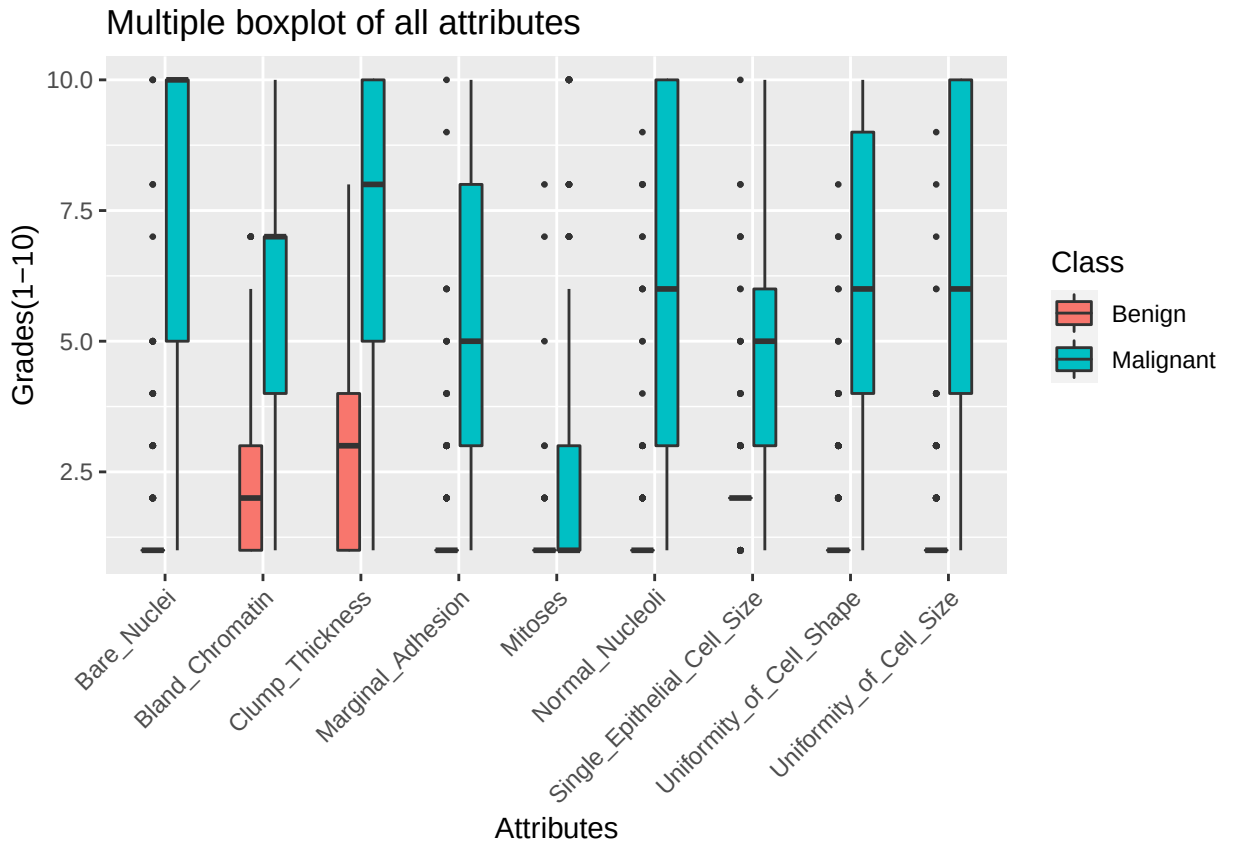

Figure 1: Box plot containg all the attributes from the Wisconsin breast cancer dataset and their spread in benign and malignant cancer cells

```
# Correlation function is used
correlation <- cor(x = Class_numeric, y = Data[,c(2:10)], use = "complete.obs")
# Correlation data is set into a data frame
cor_data <- as.data.frame(correlation)
# Data is sorted for better visualization
sort(cor_data, decreasing = T)
```

```
##   Bare_Nuclei Uniformity_of_Cell_Shape Uniformity_of_Cell_Size Bland_Chromatin
## 1   0.8226959                0.8218909               0.8208014       0.7582276
##   Normal_Nucleoli Clump_Thickness Marginal_Adhesion Single_Epithelial_Cell_Size
## 1       0.7186772       0.7147899         0.7062941                   0.6909582
```

```
##     Mitoses
## 1 0.4234479
```

Next up will be the results of the data analysis.

Data anlysis reults

Figure 1 contains a multiple box plot featuring all attributes in the data set. As discussed earlier the mitosis attribute isn't interesting as it has been removed. A first thing to note about figure 1 is that all attributes are ranked from 1 to 10 and in each case a higher score is linked to a higher change of the tumor cells being malignant. With the attributes bare nuclei, marginal adhesion, normal nucleoli and uniformity of cell size and shape all having a box with a benign mean at 1, making them look like decent predictors as most benign cases are located at the far bottom of the score with most of the malignant cases at a higher score giving a clear distinction in these attributes. Yet there are some outliers in both cases with benign cases with high scores and malignant cases with low scores but luckily as seen in figures 2 and 3 most of the time this is combined with another of the attributes having a high score for a malignant case, or a low score for a benign case. Meaning that a combination of the attributes should still be able to categories these outliers.

The other attributes, bland chromatin, clump thickness and single epithelial cell size, all have there boxes somewhat closer to each other, further they also have more overlap in outliers than the other attributes making them less useful as predictors. Yet they do still have some distinction with there boxes not matching at any point and they all have a score of at least 0.69 in the correlation matrix. Meaning that while they might not be used as the main predictors they might give information for cases of outliers which can't be classified with the main rules of the machine learning algorithm. With this in mind they will still be taken to the machine learning steps of the project.

Two of the best predictors following figure 1 and the correlation matrix are the uniformity of cell size and uniformity of cell shape. Thus they were plotted against each other in figure 2.

```
# Jitter plot if made using ggplot
ggplot(data = Data, aes(x = Uniformity_of_Cell_Size, y = Uniformity_of_Cell_Shape, colour = Class)) +
  geom_jitter(width = 0.2, height = 0.2, alpha = 0.3, size = 0.8) +
  geom_smooth(method = "lm", SE = F) +
  xlab("Uniformity of the cell size(graded)") +
  ylab("Uniformity of the cell shape(graded)") +
  ggtitle("Effects of uniformity of cell size and shape on state of breast cancer cells")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

A quick thing to note about figure 2 is that for both attributes most malignant cases are located at a higher score and most benign cases at a lower score. More specifically, most benign cases fall under a score of 1, then still regular appearing at a score of and lower than 3 for cell size and 4 for cell shape. After these scores there are only a handful of benign cases, giving some pretty clear cutoff points. Further there are almost no benign cases after scores of 8 and 9 and seemingly 0 benign cases and the score of 10. Furthermore there are some malignant points under the scores of 3 for both attributes. Meaning that these attributes alone can't completely predict the outcome of all cases.

The last thing to look at is figure 3, which plots the normal nucleoli and bare nuclei attributes against each other. The thing to be noted from figure 3 is that even though most malignant cases are as expected and following the trend of figure 2 malignant and a high score for one of the attributes, there are quite some cases where this high score in one attribute is combined with a low score in the other one. There are even quite some malignant cases when normal nucleoli is 1. Yet a combination of low scores on both attributes almost always yields a benign case. The same is true on the other end, as two high scores gives a malignant case, so much so that if both attributes score a 6 or higher there is only one benign point.

Figure 2: Shows the uniformity of the cell size vs the uniformity of the cell shape and their effects on the malignancy of the cell

```
# Jitter plot is created using ggplot
ggplot(data = Data, aes(x = Bare_Nuclei, y = Normal_Nucleoli, colour = Class)) +
  geom_jitter(width = 0.2, height = 0.2, alpha = 0.5, size = 0.8) +
  geom_smooth(method = "lm", SE = F) +
  ggtitle("Effects of nuclei transfomrations on breast tumor cell malignancy") +
  xlab("Nuclei surrounder by cytoplasm(graded)") +
  ylab("Visibility of nucleoli(graded)")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

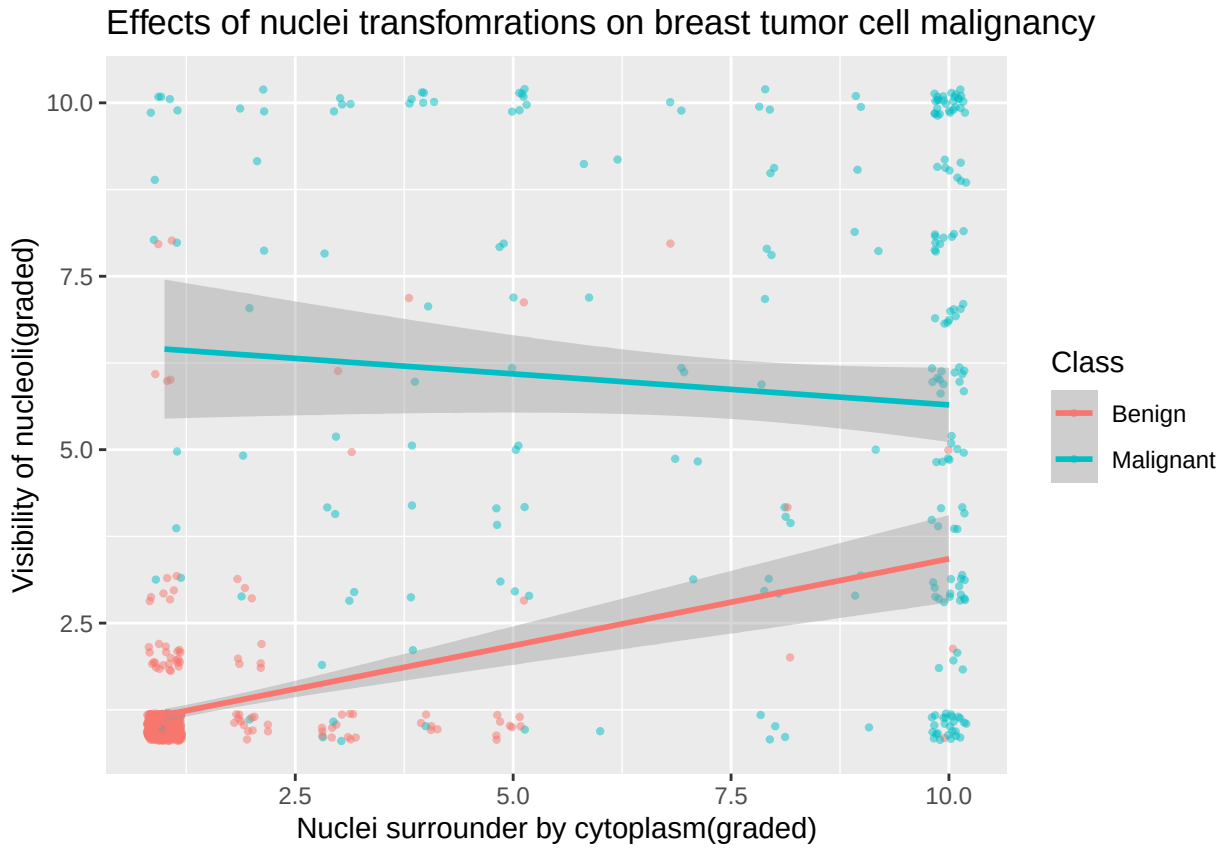

Figure 3: The effects of nuclei changes on the malignancy of a breast cancer cell

Conclusion

The data seems pretty usable for machine learning. The remaining attributes seem to correlate well with the class attribute. The attributes are already normalized by being in a 1-10 score which helps a lot in the machine learning steps as there are no numeric values.

The best attributes to use seem to be bare nuclei and uniformity of cell shape and uniformity of cell size as they have the highest correlation scores and seem decent predictors according to figures 2 and 3. Yet as seen in these figures not a single one of these attributes can be a perfect predictor on it's own, so a multiple of attributes will be needed in the machine learning steps to create an algorithm that predicts the malignancy of the breast cancer cells as best as possible.

While the three attributes named seem to be the best predictors, the other attributes might still hold valuable information as they also correlate decent and thus will be taken to the machine learning steps as they might play a crucial role in classifying the outliers in the three main attributes.

Discussion

Problems with the data

While the data being normalized beforehand might be useful, it also has it's downsides. There is no data set with the original numbers, only the graded one. This means that data that does differ a bit between some patients will be graded on the same score for both of them if they fall in it's boundaries, while the small details may give information about the malignancy of the tumor. A bigger problem with this is that the researchers themselves decided the boundaries for the scores, which gives less freedom for the cleaning of the data in the EDA.

Another problem with the data is that whilst a lower score is mostly linked with a benign tumor cell and a higher score with a malignant one, there are some outliers which do not follow these rules for some attributes. And misclassifying a case is disastrous as this might either mean that a patient gets treatment which might quite possibly be chemotherapy, which is very damaging to the patient, whitest this might be unnecessary. On the other hand when a malignant cancer cell is classified as benign the consequences can be even worse as the tumor cell will continue to grow and when the error is found it might already be to late and the patient could die. So no matter if the machine learning algorithm gets a false positive of a false negative the result will be terrible. This means that the algorithm must become as accurate as possible.

Suggestions for further research

Of course one thing that always is useful is more patients to retrieve data from, as this gives more points which gives more certainty to conclusions made from the data. Another thing that would be useful is a non graded version of the data, as there are some files with data in them, they do not directly match with the attributes from the main data set.