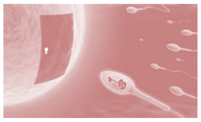


Koskas Florence
Guyon Axel
Buratti Yoann

*CS 229*
*12/12/2014*

# Semen fertility prediction based on lifestyle factors

## 1 Introduction

Although many people still think of infertility as a "woman's problem," in about 40% of infertile couples, the man is the sole cause or a contributing cause of the inability to conceive. Indeed, subfertility affects one in 20 men[1]. Male infertility has many causes : not only hormonal imbalances and/or physical problems but also psychological and/or behavioral problems. The aim of this project is to evaluate the importance of some lifestyle and environmental factors in evaluating infertility.

## 2 Methodology

### 2.1 Study Population

We use the data collected and shared[2] in 2013 by the Department of Biotechnology of University of Alicante : 100 young healthy volunteers among students who were between 18 and 36 years old provided a semen sample for analysis as well as their socio-demographic data, environmental factors, health status, and life habits. Students with previous known reproductive alterations were excluded from the statistical analysis.

### 2.2 Features and labels available from questionnaires

We have 9 features available for our 100 training examples :

1. Season in which the analysis was performed : winter (-1), spring (-0.33), summer (0.33), fall (1)
2. Age at the time of analysis : between 18 years old and 36 years old scaled to [0, 1]
3. Child diseases (i.e. chicken pox, measles, mumps, polio) : yes (0), no (1)
4. Accident or serious trauma : yes (0), no (1)
5. Surgical intervention : yes (0), no (1)
6. High fevers in the last year : less than three months ago (-1), more than three months ago (0), no (1)
7. Frequency of alcohol consumption : several times a day (0), every day (0.2), several times a week (0.4), once a week (0.6), hardly ever (0.8) or never (1)
8. Smoking habit never (-1), occasional (0) or daily (1)
9. Number of hours spent sitting per day between 1 and 16 scaled to [0, 1]

Our output is the semen analysis diagnosis : the label is “N” when semen is normal and “O” if it is altered. We match the label as follow for our algorithms :

. Semen is normal : “N” -> 0
. Semen is altered : “O” -> 1

[1] The National Center for Biotechnology Information, U.S. National Library of Medicine
[2] ttps ://archive.ics.uci.edu/ml/datasets/Fertility

### 2.3 Roadmap

For this project, our main goals are :

- Predict a young man's fertility based on his answers to 9 personal history and lifestyle questions
- Evaluate which of these 9 features have the greatest impact on the prediction's outcome

To achieve these goals, we used different machine learning algorithms and inferred the following insights :

**Logistic regression and Kernelized SVM :** Data is not linearly separable even in the high-dimensional feature space generated by a Gaussian or Sigmoid kernel.

**Decision tree algorithm :** This algorithm gives us better results, with a relatively simple decision tree.

## 3 Logistic regression and Kernelized SVM

### 3.1 Logistic regression

We implement a logistic regression coupled with Newton's method. The results show the data is not linearly separable. Although this approach is unsuccessful, we can drive some insights from the parameters theta found from the regression. Comparing the absolute values of the different coefficients of our parameter vector theta, we can get a sense of which features have the highest impact on the outcome. Our observations confirm our intuitions : for example, we find that age is negatively correlated with fertility, which seems quite reasonable. We obtain :

$$\theta_{Newton} = \begin{pmatrix} 1.8748 \\ -0.2292 \\ -1.8151 \\ -0.1530 \\ 0.05941 \\ -0.0695 \\ 0.2426 \\ 1.2262 \\ -0.0699 \\ -0.8409 \end{pmatrix} \qquad \theta_{features} = \begin{pmatrix} \text{Intercept term} \\ \text{Season} \\ \textbf{Age} \\ \text{Child disease} \\ \text{Accident or trauma} \\ \text{Surgical intervention} \\ \text{High fever} \\ \textbf{Alcohol consumption} \\ \text{Smoking habit} \\ \textbf{Sitting habit} \end{pmatrix}$$

We bolded the most highly weighted features in the logistic regression. As such, the 3 most important features seem to be (in order) : age, frequency of alcohol consumption and sitting habit. The training error is equal to 12% which shows that we can't separate the data (12 positive example out of 100). We also plotted the different features against each other, we don't notice any particular correlations between the features.

### 3.2 Kernelized SVM

Applying logistic regression taught us the data is not linearly separable in the current feature space. Hence, we choose to run an SVM algorithm with $l_1$ regularization. The objective is to find a separating hyperplane in a higher-dimensional feature space generated by applying a Kernel to the data. We use a Polynomial, a Gaussian and a Sigmoid Kernel.

We find the data is not linearly separable in the high-dimensional feature space for any of the Kernels used. However the data seems to show structure that is not captured by the Kernelized SVM. To potentially overcome this instance of underfitting and separate the data, we believe we would need additional relevant features characterizing our training set.

Below is the result of the SVM algorithm using a Gaussian Kernel. We use a PCA method to represent the data. The X and Y axes are the 2 vectors that maximize the variance of the projections of the data on a 2-dimensional space. These 2 principal components of the data are linear combinations of the features.

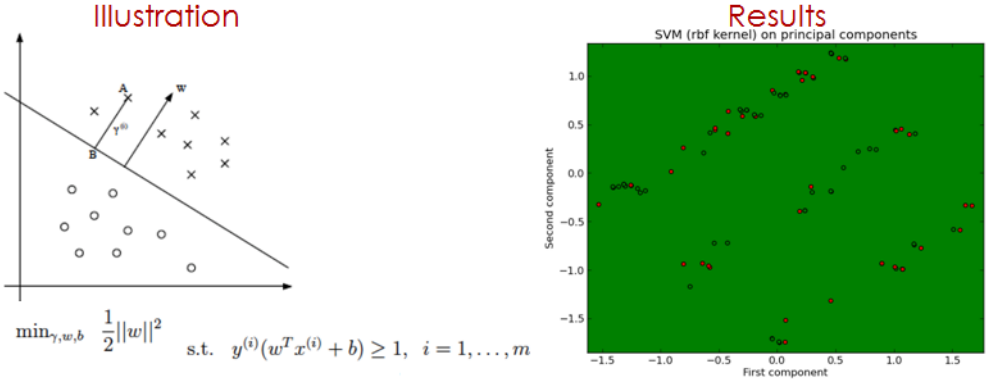


Figure 1 – Red and green data points respectively represent positive and negative outcome examples. The Kernelized SVM predicts all training examples to be negative (green background).

# 4 Decision Tree algorithm

## 4.1 Main idea

A decision tree is a non-linear model that splits training examples into different bags of data, called leaves, so that we can make a more accurate prediction in each leaf. To build the tree and get the data clustered in these leaves, we select the feature with the heaviest weight in a linear regression. We split the data above and below a certain threshold for this feature and we recurse on the newly created subset. When the tree depth exceeds a specific, carefully chosen threshold, we stop and the lastly created subsets are our leaves. The proportion of negative or positive data points in each leaf is an estimate of the probability of being negative or positive respectively, for any new test example that happens to belong to the leaf.

Using the optimal tree depth is key to the performance of the algorithm. Creating a deeper tree will lead to very few data points per leaf and result in overfitting, while stopping too soon will lead to underfitting. How do we find the optimal tree depth named the "Complexity Parameter"?
If we plot the test error as a function of the tree depth, $E_{test} = f(D)$, we can derive the tree depth $D^\star$ that achieves the minimum of the test error. An usual Complexity Parameter is taken to be one standard deviation below this minimum :

$$CP = D^\star - \sigma_{testerror}$$

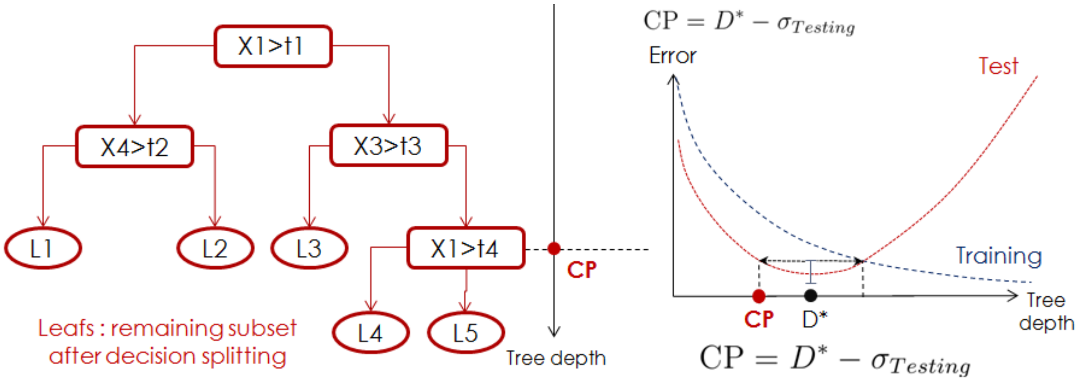

X1>t1
X4>t2
X3>t3
L1
L2
L3
X1>t4
L4
L5
CP
Tree depth
Leafs : remaining subset after decision splitting
$\text{CP} = D^* - \sigma_{Testing}$
Error
Test
Training
CP
D*
Tree depth
$\text{CP} = D^* - \sigma_{Testing}$

Figure 2 – Decision tree algorithm and Complexity Parameter derivation

## 4.2 Fertility decision tree

We randomly divide our data set into two subsets : training set (50 %) and test set (50 %), using the R module : rpart. In figure 3 is the tree we find when training the decision tree algorithm described above on our training set :

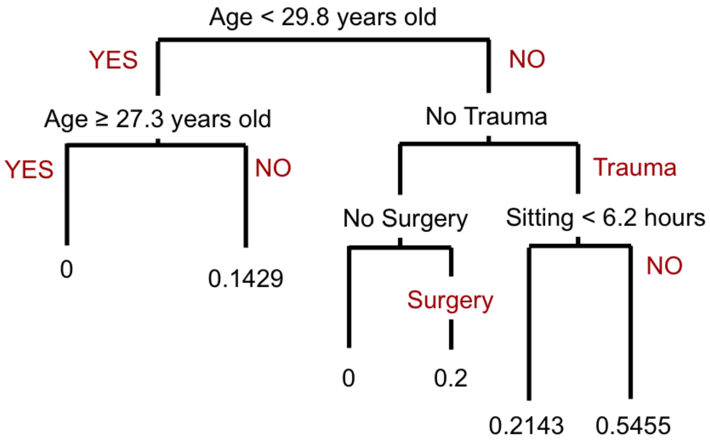


Figure 3 – Decision tree algorithm result

We obtain 6 bags of data. Each leaf displays the probability of the outcome being positive i.e. the man being infertile. For example, according to our data and the model used : a man older than 29.8 years old, who experienced an accident or trauma in the past, and who sits more than 6.2 hours per day has a 54 % chance to be infertile. While the significance of these results can be further debated and analyzed, the analysis suggests at least that fertility is not an intrinsic permanent characteristic but life events like a car accident can potentially alter male fertility.

We repeat the tree generation process a hundred times, and we always get the same decision features. Although the training set is randomly drawn from 50% of our total data set.

Interestingly, compared to logistic regression, the decision tree algorithm gives us a different hierarchy of the features : Age, Traumas, Surgery and Time spent sitting. However, just like with logistic regression, this should be taken with caution given the overall results of the models. We can note for example that our tree tells us that men younger than 27.3 years old are more likely to be infertile than men between 27.3 and 29.8, which is definitely a surprising result that could be questioned. We then test our model on the test set. Below is the confusion matrix describing the test error :

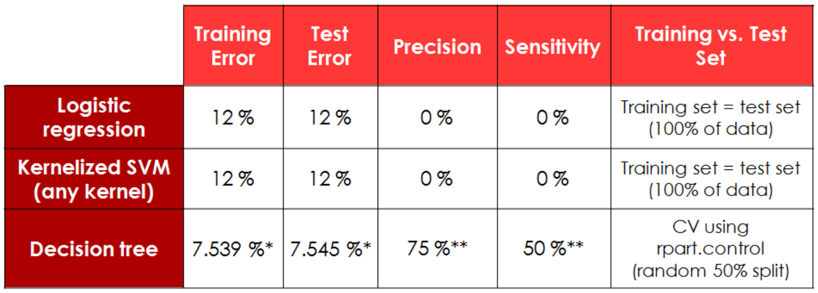
| | | Ground truth | |
|---|---|---|---|
| | | 0 | 1 |
| Prediction | 0 | 86 % | 6 % |
| | 1 | 2 % | 6 % |

A confusion matrix is particularly relevant to analyze results since the number of positive examples is fairly small compared to the total dataset size (12 out of 100). We can see that our tree algorithm only predicts half of the positive examples correctly, while more than 97% of the negative example are predicted correctly. We believe our algorithm does poorly at predicting infertility mainly because the initial database is extremely small and biased.

The next step we are considering to improve this algorithm is to implement bootstrapping on the data set.

# 5 Further discussion & future work

Looking at the results of both SVM and decision tree, we can build the following table :

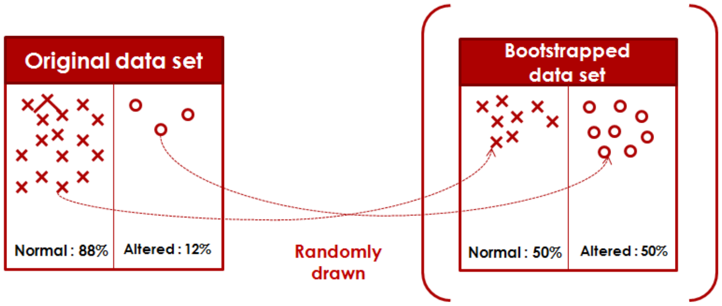


| | Training Error | Test Error | Precision | Sensitivity | Training vs. Test Set |
|---|---|---|---|---|---|
| Logistic regression | 12 % | 12 % | 0 % | 0 % | Training set = test set (100% of data) |
| Kernelized SVM (any kernel) | 12 % | 12 % | 0 % | 0 % | Training set = test set (100% of data) |
| Decision tree | 7.539 %* | 7.545 %* | 75 %** | 50 %** | CV using rpart.control (random 50% split) |

The main challenge with this project was the very small number of data points in the set (100), as well as a limited number of features (9), and we wanted to see how far we could go with such constraints. The bias of the data set towards normal fertility (as in real life actually) was a hindrance, which could be resolved with bootstrapping (we could greatly improve our results by collecting more data, but this has its own challenges).

Bootstrapping consists in artificially obtaining a larger data set from the original one. Given our data set of 100 examples, 40 bags of 34 data points are created (i.e. one third of the size of the original data set). In each bag of 34 data points, are included :

. data points (50%) randomly drawn from the set of normal samples in the original data set
. data points (50%) randomly drawn from the set of altered samples in the original data set



On these 40 bags of data, a decision tree algorithm is then trained. On the test set, each tree gives us a prediction. These predictions are averaged to obtain the final prediction.

Alternatively, as we mentioned in the SVM section, we expect that additional features would enable us to separate the data and would thus greatly improve the value of the data set. In fact, the university of Alicante that published the data also had access to more features to perform their analysis [3].

From discussions with the class teaching team (Dave Deriso, project TA) and visitors at the poster session, it appears that random forest or boosting our tree might be other interesting tree algorithm variant we could try. Finally, the same paper uses CNN (Convolution Neural Networks) as a prediction tool, and using deep learning can help the accuracy of the prediction.

[3] Predicting seminal quality with artificial intelligence methods, *David Gil, Jose Luis Girela, Joaquin De Juan, M. Jose Gomez-Torres, Magnus Johnsson* (2012)