Decision trees and ways on removing noisy labels
Identify costumers in unsound service models

Yannick Misteli

Julius Bär

yannick.misteli@juliusbaer.com

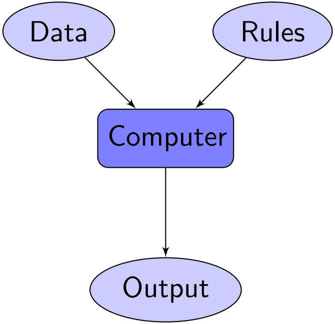
September 4, 2018

Overview

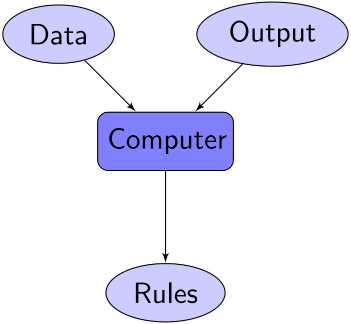
1. Interpretable Models and general aspects of ML
   - Introduction
   - Interpretability
   - Interpretable Models
2. Use Case
   - Service Models
   - Use case
   - Dataset
   - Decision tree
3. After Math
   - Noisy Labels
   - Dealing with Noise
   - Noise Sensitivity of ML algorithms

Interpretable Models and general aspects of ML

Traditional Programming



Supervised Machine Learning



Interpretability

Interpretability

Interpretability is the degree to which a human can understand the cause of a decision[1]

- The importance of interpretability or **what vs why and finding meaning in the world** (Regulator)
- Criteria for interpretability methods or **intrinsic vs post hoc**
- Human-friendly explanations or **what is a good explanation?**

[1]Miller, Tim. 2017. "Explanation in Artificial Intelligence: Insights from the Social Sciences." arXiv Preprint arXiv:1706.07269

Interpretable Models

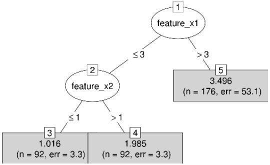
$$y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p} + \epsilon_i$$

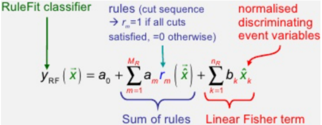
$$P(y_i = 1) = \frac{1}{1+exp(-(\beta_0+\beta_1 x_{i,1}+\ldots+\beta_p x_{i,p}))}$$

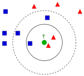
$$P(C_k \mid x) = \frac{1}{Z} P(C_k) \prod_{i=1}^{n} P(x_i \mid C_k)$$

Models

- Linear models
- Logistic regression
- Naive Bayes
- Decision trees
- RuleFit[2]
- k-Nearest Neighbours







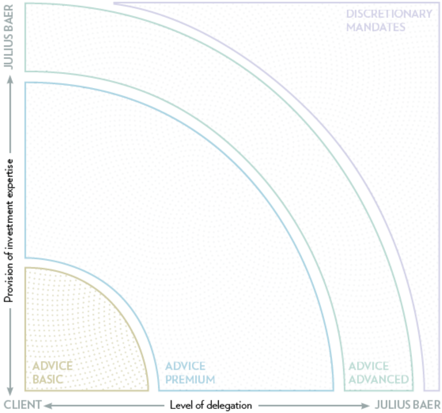
[2] Friedman-Popescu, Tech Rep, Stat. Dpt, Standford U., 2003

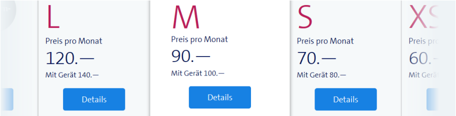
Use Case

Advisory Service Models

1. Basic
2. Premium
3. Advanced

Every advised client signs a service model agreement. Hence, according to preferences and service needs either a basic, premium or advanced service contract is put in place.





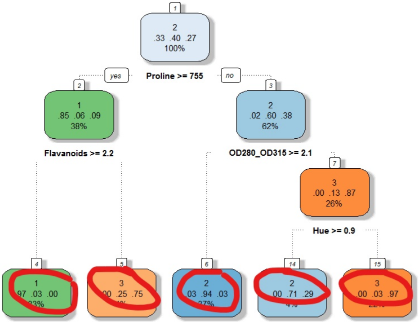
Use case

Problem

How do identify clients that should be in a different Service Model?

Idea

Fit decision tree and investigate terminal nodes for misclassified clients



Dataset

Generating multivariate tri-modal mixed distributions

Multivariate data with count and continouse variable with a pre-specified correlation matrix is generated. The count and continouous variable are assumed to have Poisson and normal marginals, respectively. The resulting mixture is

$$F(x) = \sum_{i=1}^{n} w_i P_i(x),$$

where $n = 3$; $w_1 = 0.6$, $w_2 = 0.3$, $w_3 = 0.1$ and P_i is the corresponding multivariate Poisson-Normal distribution.

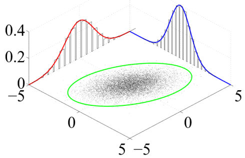


Figure: Example of sample points from a multivariate normal distribution with $\sigma = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $\Sigma = \begin{bmatrix} 1 & \frac{3}{5} \\ \frac{3}{5} & 2 \end{bmatrix}$, shown along with the 3-sigma ellipse.[3]

[3] en.wikipedia.org/wiki/Multivariate_normal_distribution

Data Summary

Syntethic Data

All data contained in this slides have been generated synthetically and not by Julius Bär. In no event shall the author or Julius Bär be liable for any direct, indirect, special or incidental damages resulting from, arising out of or in connection with, the use of the data contained herein.

```r
head(data)
```

```
##   Service.Model Nr.Trades Nr.CCY Nr.Positions
## 1         basic         1      2            1
## 2         basic         2      4            4
## 3      advanced         2      6           15
## 4      advanced         9      6           18
## 5      advanced         2      3            5
## 6       premium         0      3            7
##         AuM       Cash      Random
## 1 1198618.9  133102.70  0.46951209
## 2 1001045.8 -223904.30 -1.18726746
## 3 1629286.8  180392.06  0.54675242
## 4 3947500.7  234791.35  0.01287336
## 5  361907.7  -74267.38 -1.21565020
## 6 2760734.3  340989.71 -2.05284115
```

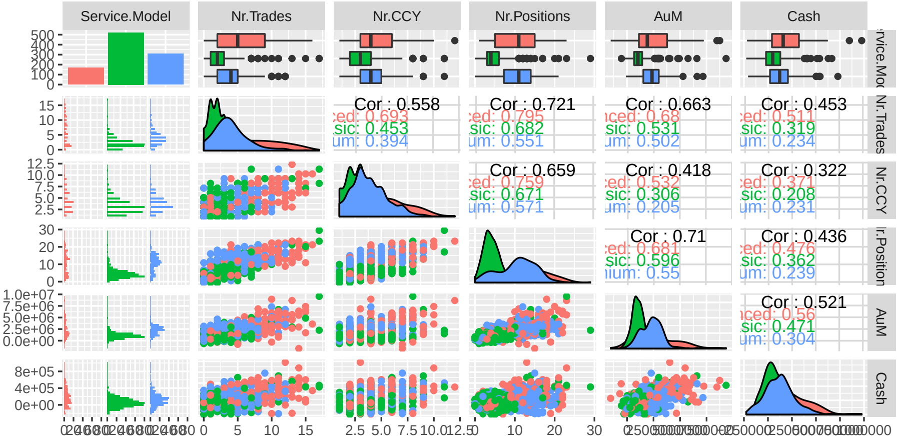
```r
summary(data[,1:6])
```

```
##   Service.Model     Nr.Trades
##  advanced:16955   Min.   : 0.000
##  basic   :52140   1st Qu.: 1.000
##  premium :30905   Median : 3.000
##                   Mean   : 3.394
##                   3rd Qu.: 4.000
##                   Max.   :23.000
##      Nr.CCY         Nr.Positions
##  Min.   : 1.000   Min.   : 0.000
##  1st Qu.: 2.000   1st Qu.: 3.000
##  Median : 3.000   Median : 6.000
##  Mean   : 3.587   Mean   : 7.478
##  3rd Qu.: 5.000   3rd Qu.:11.000
##  Max.   :18.000   Max.   :30.000
##       AuM                  Cash
##  Min.   :-4851052   Min.   :-464873
##  1st Qu.:  868542   1st Qu.:   3929
##  Median : 1385314   Median :  85995
##  Mean   : 1701572   Mean   : 105146
##  3rd Qu.: 2346875   3rd Qu.: 183797
##  Max.   :12001587   Max.   :1066861
```

Exploratory Data Analysis

Syntethic Data

All data contained in this slides have been generated synthetically and not by Julius Bär. In no event shall the author or Julius Bär be liable for any direct, indirect, special or incidental damages resulting from, arising out of or in connection with, the use of the data contained herein.



Label Noise

Syntethic Data

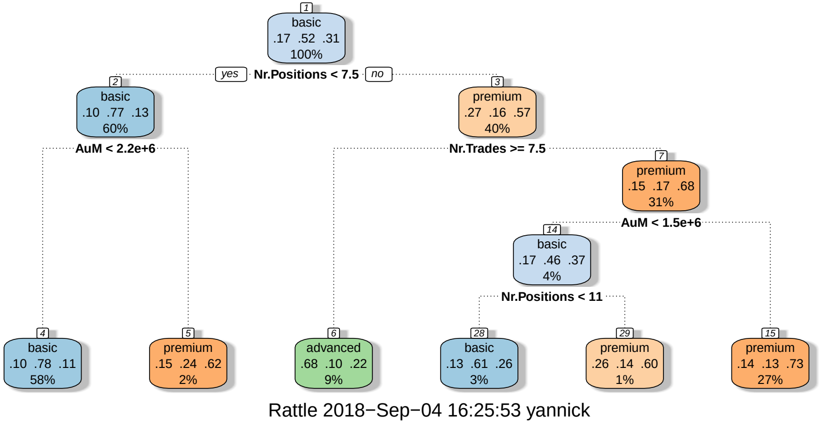
All data contained in this slides have been generated synthetically and not by Julius Bär. In no event shall the author or Julius Bär be liable for any direct, indirect, special or incidental damages resulting from, arising out of or in connection with, the use of the data contained herein.

```
noisy_data <- clean_data
leng <- nrow(noisy_data)
labelnoise <- 20
resample <- sample.int(leng, leng/100*labelnoise)
mylabels <- unique(clean_data$Service.Model)
for(k in resample){
  myset <- noisy_data[k,]
  noisy_data[k,1] <- sample(mylabels[(myset$Service.Model != mylabels)],1)
}
```

Decision Tree

Syntethic Data

All data contained in this slides have been generated synthetically and not by Julius Bär. In no event shall the author or Julius Bär be liable for any direct, indirect, special or incidental damages resulting from, arising out of or in connection with, the use of the data contained herein.



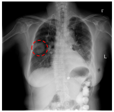
After Math

Noisy lables - sources and effects[4]

Sources of noise

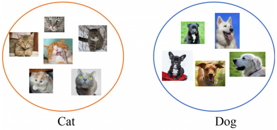
- insufficient information provided to the expert
- errors in the expert labelling itself
- subjectivity of the labelling task
- communication/enconding problems

X	Y	Z	Label
0.01	-0.01	-0.02	Resting
4.04	-10.2	7.66	Running
1.23	1.73	0.02	Walking
0.03	-0.07	0.09	Resting
15.72	-25.76	12.23	Running
1.45	0.33	0.43	Walking



Effects of noise

- decrease the classification performances
- increase/decrease the complexity of learned models
- pose a threat to tasks like e.g. feature selection



[4] https://labelnoise2017.loria.fr/wp-content/uploads/2017/11/présentation-LABELNOISE17-Frénay.pdf

Dealing with Noise

remove

robust MLS (SVM soft margin) ensembles

Noise Sensitivity of ML algorithms