Database Data Mining: Practical R Enterprise and Oracle Advanced Analytics

Husnu Sensoy
husnu.sensoy@globalmaksimum.com

Global Maksimum Data & Information Technologies

October 2, 2012
Content

1. Introduction

2. Oracle Enterprise R in Practice
   - Data Visualization
   - A Bit of Probability and Information Theory
   - Optimization
   - Text Analysis & Decision Trees

3. Wrap up
Who am I?

✓ Founder at Global Maksimum Data & Information Technologies
✓ Oracle ACE in BI Domain
✓ Oracle Magazine DBA of the Year in 2009
✓

Husnu Sensoy husnu.sensoy@globalmaksimum.com
A bunch of people who know what they are doing mainly focused on data and the transformation of data into information.
Global Maksimum Data & Information Technologies

✓ A bunch of people who know what they are doing mainly focused on data and the transformation of data into information.
✓ Complex Event Processing
Global Maksimum Data & Information Technologies

✓ A bunch of people who know what they are doing mainly focused on data and the transformation of data into information.
✓ Complex Event Processing
  ✓ 1.2 Million Event in a second on 2x2 Socket Nehalem Blades
✓ Data Mining
Global Maksimum Data & Information Technologies

✓ A bunch of people who know what they are doing mainly focused on data and the transformation of data into information.
✓ Complex Event Processing
  ✓ 1.2 Million Event in a second on 2x2 Socket Nehalem Blades
✓ Data Mining
  ✓ Churn Prediction Models for Telcos
  ✓ Marketing Target Selection Models
✓ Large Scale Database Management System Projects

Husnu Sensoy husnu.sensoy@globalmaksimum.com Global Maksimum Data & Information Technologies

Database Data Mining: Practical R Enterprise and Oracle Advanced Analytics
Global Maksimum Data & Information Technologies

✓ A bunch of people who know what they are doing mainly focused on data and the transformation of data into information.
✓ Complex Event Processing
  ✓ 1.2 Million Event in a second on 2x2 Socket Nehalem Blades
✓ Data Mining
  ✓ Churn Prediction Models for Telcos
  ✓ Marketing Target Selection Models
✓ Large Scale Database Management System Projects
  ✓ 120+ TB Exadata migration from UNIX systems.
  ✓ Exadata Master Class all over the EMEA region for Exadata customers, Oracle partners, and Oracle staff at the region.
Advanced Analytics

✓ For the first version of BI we just filter rows, project columns, aggregate them using some functions, and give only what customer asks for.
Advanced Analytics

✓ For the first version of BI we just filter rows, project columns, aggregate them using some functions, and give only what customer asks for.

✓ After we have focused on *machine generated data*, or Big Data dealing the data as we did before becomes more and more fruitless.
Advanced Analytics

✓ For the first version of BI we just filter rows, project columns, aggregate them using some functions, and give only what customer asks for.

✓ After we have focused on machine generated data, or Big Data dealing the data as we did before becomes more and more fruitless.

✓ That’s mainly because of the fact that there is only tiny amount of information available in this pile of data.
Advanced Analytics

✓ For the first version of BI we just filter rows, project columns, aggregate them using some functions, and give only what customer asks for.

✓ After we have focused on *machine generated data*, or Big Data dealing the data as we did before becomes more and more fruitless.

✓ That’s mainly because of the fact that there is only tiny amount of information available in this pile of data.

✓ So it requires better tricks, automation, and post-analysis capabilities.
In-database Advanced Analytics

✓ 80% of data mining activity for enterprise means *feature engineering*. 
80% of data mining activity for enterprise means feature engineering.

Feature Engineering requires an iterative process of

- Filtering data (WHERE)
- Aggregating data (GROUP BY)
- Transforming data (CASE, DECODE, COALESCE, etc.)
In-database Advanced Analytics

✓ 80% of data mining activity for enterprise means feature engineering.
✓ Feature Engineering requires an iterative process of
  ✓ Filtering data (WHERE)
  ✓ Aggregating data (GROUP BY)
  ✓ Transforming data (CASE, DECODE, COALESCE, etc.)
In-database Advanced Analytics

✓ 80% of data mining activity for enterprise means *feature engineering*.

✓ *Feature Engineering* requires an iterative process of
  ✓ Filtering data (WHERE)
  ✓ Aggregating data (GROUP BY)
  ✓ Transforming data (CASE, DECODE, COALESCE, etc.)

✓ It is almost impossible to maintain an integrated mining environment (Scripts, files, metafiles, etc.) out of database
Oracle Advanced Analytics Toolkit

✓ SQL-2003 & Extensions
✓ Oracle Data Mining
✓ Oracle Spatial Extensions
✓ Flow based mining with SQL Developer
✓ Oracle Enterprise R
✓ R is a free software environment for statistical computing and graphics.
✓ Majority of newbies (young data scientists) recently graduate or to be graduated from top universities use R.
✓ Batteries are included.
✓ Runs on all modern platforms.
Oracle R Enterprise

✓ Data you can process with standard R is limited with the amount of memory available on the server running R.
Oracle R Enterprise

✓ Data you can process with standard R is limited with the amount of memory available on the server running R.
✓ In order to bypass this problem people implement their own solutions in order to off-store data or utilize data sampling techniques.
Oracle R Enterprise

✓ Data you can process with standard R is limited with the amount of memory available on the server running R.
✓ In order to bypass this problem people implement their own solutions in order to off-store data or utilize data sampling techniques.
✓ ORE is an extension to standard R adding Oracle steroids into it.
Oracle R Enterprise

✓ Data you can process with standard R is limited with the amount of memory available on the server running R.

✓ In order to bypass this problem people implement their own solutions in order to off-store data or utilize data sampling techniques.

✓ ORE is an extension to standard R adding Oracle steroids into it.

✓ The basic idea is to off-load R commands seemless to Oracle Database or Oracle Big Data Appliance.
This session is not a R tutorial session but rather a fly over some possible solutions to real life scenarios using R. If you need some R tutorial please refer to:

✓ Oracle R Enterprise Training 2 - Introduction to R
✓ R Studio
Data Visualization

✓ Advance data analysis usually starts and ends with data visualization.
  ✓ Before modeling anything data scientists use graphs & charts to figure out behaviour of data
  ✓ After modeling in order to report the results they again refer to charts.

✓ R supports tens of different charting & graphing packages. Just to mention two of them
  lattice is used to generate conditioned graphs (a.k.a. trellis graphs)
  ggplot2 is used to make graph generation more consistent in R.
Histogram

✓ Do you see any significant pattern in distribution?
✓ Do you like the way histogram is represented?

```r
source("~/r-snippets/oow2012/mydata.r", local=TRUE)
dataset = generateCustomer()

h = hist(dataset$BillPerPeriod, freq=TRUE,
          ylab="Number of Customers",
          xlab="Bill Amount",
          main="Bill Amount Distribution")
```

![Histogram of Bill Amount Distribution](image)
Data Visualization

Remove the Outliers

Do you see any significant pattern in distribution?

```r
source("~/r-snippets/oow2012/mydata.r", local=TRUE)

dataset = generateCustomer()

nooutlier = function(data, column, q=0.99, inc=TRUE) {
  q = quantile(data[,column], na.rm=TRUE, probs = quantile, names=FALSE)
  if (inclusive) {
    pruned = subset(data, data[,column] <= q)
  } else {
    pruned = subset(data, data[,column] < q)
  }
  pruned
}

pruned = nooutlier(dataset, "Bill per Period", 0.99)

h = hist(pruned$Bill per Period, freq=TRUE,
         ylab="Number of Customers",
         xlab="Bill Amount",
         main="Bill Amount Distribution")
```

Husnu Sensoy husnu.sensoy@globalmaksimum.com
Global Maksimum Data & Information Technologies

Database Data Mining: Practical R Enterprise and Oracle Advanced Analytics
Data Visualization

Conditional Histograms

```r
source("~/r-snippets/ow2012/mydata.r", local=TRUE)
source("~/r-snippets/ow2012/commons.r", local=TRUE)
dataset = generateCustomer()
pruned = nooutlier(dataset, "BillperPeriod", 0.99)
library(lattice)
histogram(~BillperPeriod | UsingServiceX, data=pruned)
```
Too Many Columns to Visualize

```r
source("~/r-snippets/oow2012/mydata.r", local=TRUE)
source("~/r-snippets/oow2012/commons.r", local=TRUE)

dataset = generateCustomer()
head(dataset)

pruned = nooutlier(dataset, "BilliperPeriod", 0.99)

library(lattice)
histogram(~BilliperPeriod | CarBrand, data=pruned)
```
A Bit of Probability and Information Theory

Comparing Histograms

✓ We need a way to calculate similarity between those histograms.
✓ A strong tool from information theory Kullback—Leibler Divergence allows us to define a distance metric between two distributions.

```r
equidwidth = function(data, col, n=10, sf=1e-6){
  qlist = quantile(data[, col], na.rm=TRUE,
                  probs = seq(0.1,1.0, by=1./n),
                  names=FALSE)

  result=c()
  for (quantile in qlist){
    result = c(result ,
                (nrow(subset(data, data[, col] <= quantile))/nrow(data)))
  }

  result[1:n]=c(0, result[1:(n-1)]) + rep(sf, n)
}
```
KL Divergence & Symmetry

- \( D_{KL}(P \parallel Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)} \)
- Notice that \( D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P) \)
- So we simply take the average of two to obtain a symmetric metric.

```r
kl_distance = function(dist1, dist2) {
    kl1 = 0.0
    for (i in 1:length(dist1)) {
        kl1 = kl1 + dist1[i] * log(dist1[i]/dist2[i], 2)
    }
    kl2 = 0.0
    for (i in 1:length(dist1)) {
        kl2 = kl2 + dist2[i] * log(dist2[i]/dist1[i], 2)
    }
    (kl1+kl2)/2
}
```
A Bit of Probability and Information Theory

Top 5 Car Brands whose Owners Diverge from Baseline

<table>
<thead>
<tr>
<th>Brand</th>
<th>KL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lancia</td>
<td>8.969125</td>
</tr>
<tr>
<td>Lincoln</td>
<td>8.969125</td>
</tr>
<tr>
<td>Proton</td>
<td>7.572549</td>
</tr>
<tr>
<td>Daewoo</td>
<td>7.572549</td>
</tr>
<tr>
<td>Pontiac</td>
<td>6.421267</td>
</tr>
</tbody>
</table>

```
ddf = NULL
baseline = equiwidth(pruned, "BillperPeriod")
for (brand in dataset[!duplicated(dataset[,c('CarBrand')]),1]){
    brandDist = equiwidth(subset(pruned, pruned[, 'CarBrand'] == brand), "BillperPeriod")
    ddf = rbind(ddf, data.frame(carbrand=brand, kl=kl_distance(baseline, brandDist)))
}

head(ddf[order(ddf$kl, decreasing=TRUE),])
```
Problem Definition

✓ We have a terrain covered by several stations and each point on the terrain has one of the following status:

   **GREEN** Region is in the LoS of at least one station.

   **YELLOW** Region is in the LoS of at least one station but far away.

   **RED** Region is out of LoS.

✓ For a fixed number of stations we need to cover as much as we can.
Model Sketch Up

1. Define a function to calculate the ratio of green zones on terrain.

```r
# Compute merged status of all observers
mergedstatus <- rep("red", length(terrain$height))
for (i in seq(1:dim(m)[1])){
  terrain$dist2observer = distance(terrain, c(m[i,], 7))
  status = LoS(terrain, c(m[i,], 7), maxDist)
  mergedstatus = updatestatus(mergedstatus, status)
}
sum(mergedstatus=="green")
```

2. Give this function to one of optimization modules of R (Nelder — Mead Technique) which can handle non-smooth target functions.

```r
optim <- optim(observers, targetfunc, control=list(fnscale=-1, trace=5, REPORT=1))
```

3. Get the optimal station distribution.

---

1Refer to LoS Analysis (Part 4)
1 Station (54%)
3 Stations (83%)
6 Stations (99%)
Problem Definition

✓ For a given string which is written intentionally or erroneously wrong by subscribers, how can we build a model which can deduce the most probable string among 3 possibilities (or chose to not making any decision).
Problem Definition

✓ For a given string which is written intentionally or erroneously wrong by subscribers, how can we build a model which can deduce the most probable string among 3 possibilities (or chose to not making any decision).

Our legitimate strings are *mom, dad, and brother*. And we have

- *brothe* → brother
- *bro* → brother
- *brother1* → brother
- *p* → ?
- *1234* → ?
- *mom.i.came.home* → mom
- *mmomyy* → mom
- *dad[atwork]* → dad
- *dod* → dad
Model Sketch Up

1. Do some feature engineering
Model Sketch Up

1. Do some feature engineering
   - ✔ Length of the string

Husnu Sensoy husnu.sensoy@globalmaksimum.com
Global Maksimum Data & Information Technologies

Database Data Mining: Practical R Enterprise and Oracle Advanced Analytics
Model Sketch Up

1. Do some feature engineering
   - ✓ Length of the string
   - ✓ Prefix flag (3 attributes for each)
Model Sketch Up

1. Do some feature engineering
   - ✓ Length of the string
   - ✓ Prefix flag (3 attributes for each)
   - ✓ Contains flag (3 attributes for each)
Model Sketch Up

1. Do some feature engineering
   - ✓ Length of the string
   - ✓ Prefix flag (3 attributes for each)
   - ✓ Contains flag (3 attributes for each)
   - ✓ Anything else?

2. Build a classifier to classify those texts based on those features.

3. Evaluate your classifier
Text Analysis & Decision Trees

First Model

```r
source("~/r-snippets/oow2012/mydata.r", local=TRUE)
df = generateText()

library(rpart)

# grow tree
fit <- rpart(corrected ~ length+prefixBrother+prefixDad+prefixMom+instrBrother+
             instrDad+instrMom,
             method="class", data=df)
table(pred = predict(fit, df, type="class"),
      true = df$corrected)
```

<table>
<thead>
<tr>
<th>true</th>
<th>pred</th>
<th>brother</th>
<th>dad</th>
<th>mom</th>
</tr>
</thead>
<tbody>
<tr>
<td>?</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>brother</td>
<td>0</td>
<td>30</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>dad</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>mom</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20</td>
</tr>
</tbody>
</table>
Text Analysis & Decision Trees

More Feature Engineering using Jaro-Winkler Algorithm

Jaro-Winkler distance is a distance metric between strings which can be used as a fuzzy string matching algorithm resilient to typo errors.

```
library(RecordLinkage)

enhanced = data.frame(df,
  momScore = jaro.winkler("mom", df$original),
  dadScore = jaro.winkler("dad", df$original),
  brotherScore = jaro.winkler("brother", df$original))
```

<table>
<thead>
<tr>
<th></th>
<th>true</th>
<th>pred</th>
<th>brother</th>
<th>dad</th>
<th>mom</th>
</tr>
</thead>
<tbody>
<tr>
<td>?</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>brother</td>
<td>0</td>
<td>30</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>dad</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>mom</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>0</td>
</tr>
</tbody>
</table>
R contains lots of libraries to help you model a physical phenomenon in anyway you like and visualize it.
R contains lots of libraries to help you model a physical phenomenon in anyway you like and visualize it.

Oracle Enterprise R makes it possible to handle large volumes of data without changing your R environment basics.
Conclusion

✓ R contains lots of libraries to help you model a physical phenomenon in anyway you like and visualize it.
✓ Oracle Enterprise R makes it possible to handle large volumes of data without changing your R environment basics.
✓ Don’t take ODM and Oracle Enterprise R as alternatives of each other but rather complimentary solutions of the same problem.
Question & Answer
Stay in Touch

husnu.sensoy@gmail.com

husnu.sensoy@globalmaksimum.com

http://husnusensoy.wordpress.com

@husnusensoy