IN4355
Machine Learning Samples

Yanbo Huang, Yu Liang and Bo Wang

Goal

Implement Mike’s blog samples using several different Machine Learning libraries

Libraries we used:

1. Scala - Mllib
2. R packages
3. Go - Golearn
4. Python - Scikit Learn
5. Python - Graphlab
6. Python - NLTK (Naive Bayes)
7. Python - Recommendation system

Project Structure:

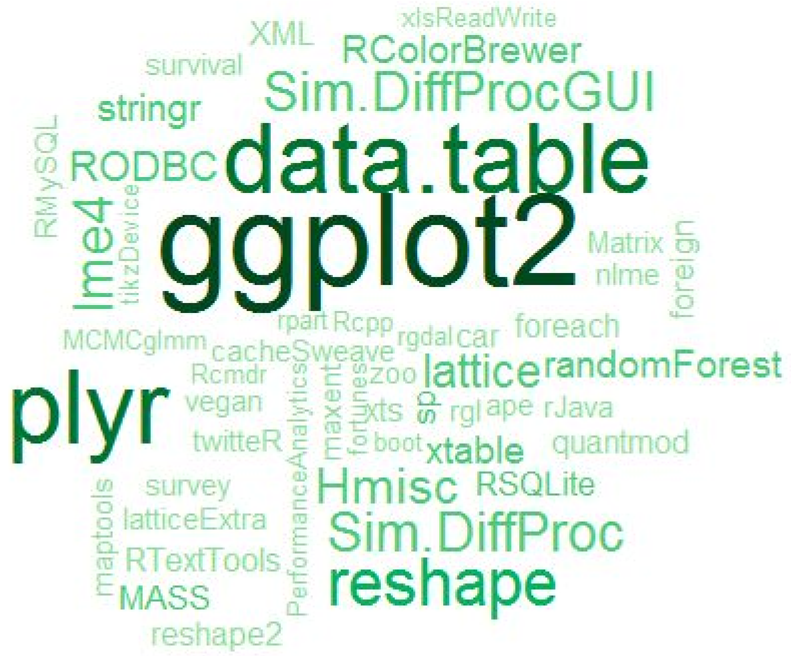
```
/library
    /img
    readme.md
    linear_regression.*
    logistic_regression.*
    knn.*
    svm.*
    …
```

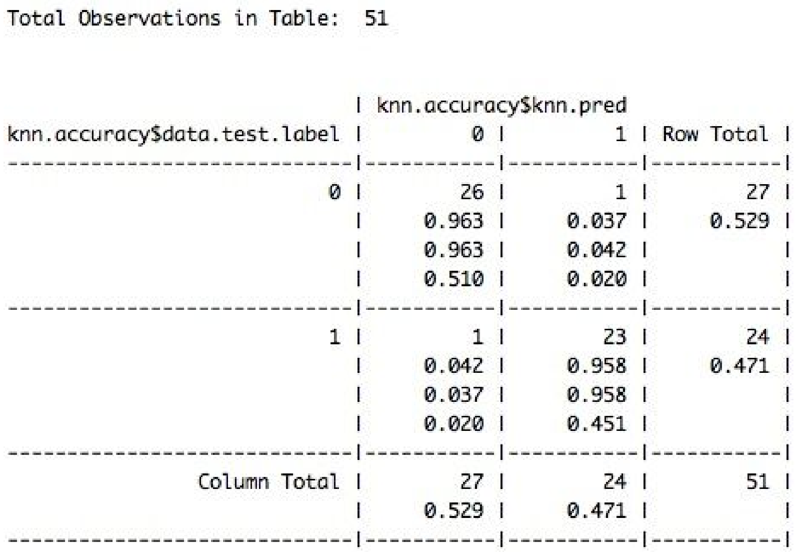
Readme Structure:

1. Overview of the library
2. Implement ML algorithms with this library
3. Model Evaluation
4. Combine with Functional Programming

Machine Learning with R

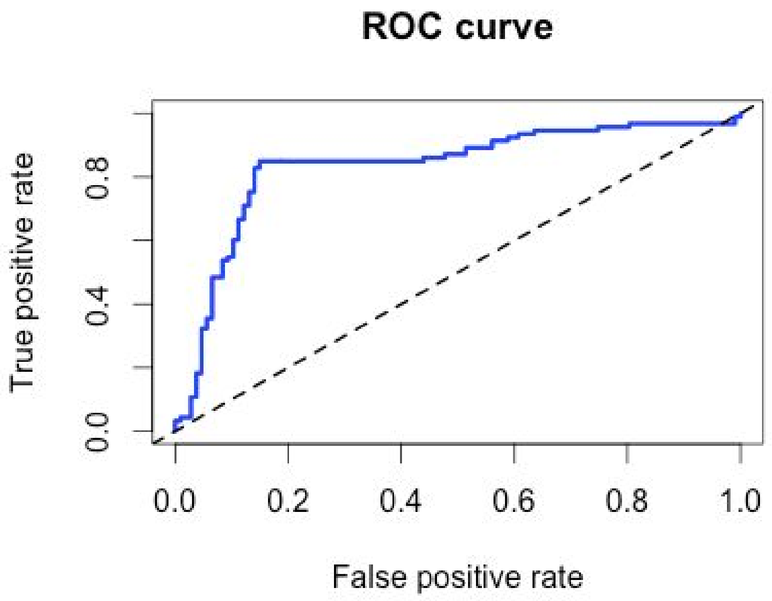
Data Mining and
Machine Learning


xlsReadWrite
XML
RColorBrewer
survival
Sim.DiffProcGUI
stringr
RMySQL
RODBC
data.table
lme4
tikzDevice
ggplot2
Matrix
foreign
nlme
rpart
Rcpp
rgdal
car
foreach
MCMCglmm
cacheSweave
lattice
randomForest
plyr
Rcmdr
zoo
vegan
PerformanceAnalytics
maxent
fortunes
xts
sp
rgl
ape
rJava
twitteR
boot
xtable
quantmod
survey
Hmisc
RSQLite
maptools
latticeExtra
Sim.DiffProc
RTextTools
reshape
MASS
reshape2



```
Total Observations in Table:  51


                             | knn.accuracy$knn.pred
knn.accuracy$data.test.label |         0 |         1 | Row Total |
-----------------------------|-----------|-----------|-----------|
                           0 |        26 |         1 |        27 |
                             |     0.963 |     0.037 |     0.529 |
                             |     0.963 |     0.042 |           |
                             |     0.510 |     0.020 |           |
-----------------------------|-----------|-----------|-----------|
                           1 |         1 |        23 |        24 |
                             |     0.042 |     0.958 |     0.471 |
                             |     0.037 |     0.958 |           |
                             |     0.020 |     0.451 |           |
-----------------------------|-----------|-----------|-----------|
                Column Total |        27 |        24 |        51 |
                             |     0.529 |     0.471 |           |
-----------------------------|-----------|-----------|-----------|
```

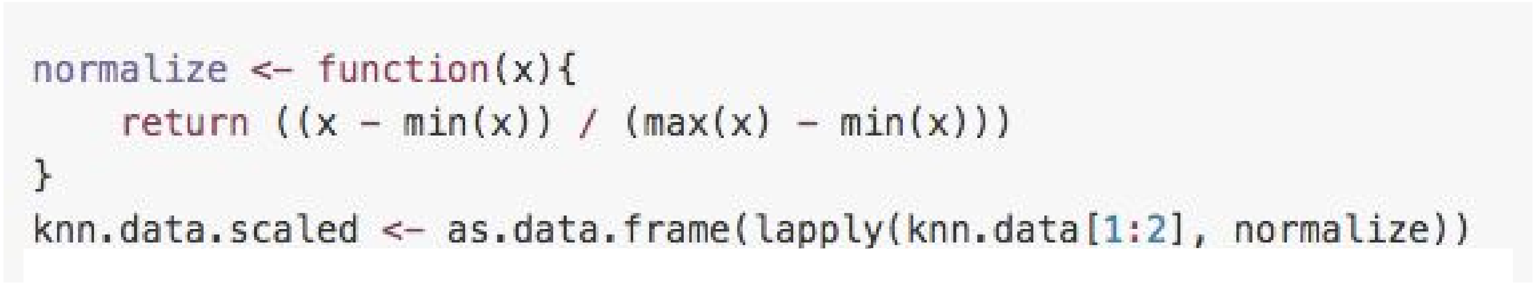


"R, at its heart, is a functional programming (FP) language"

——Hadley Wickham

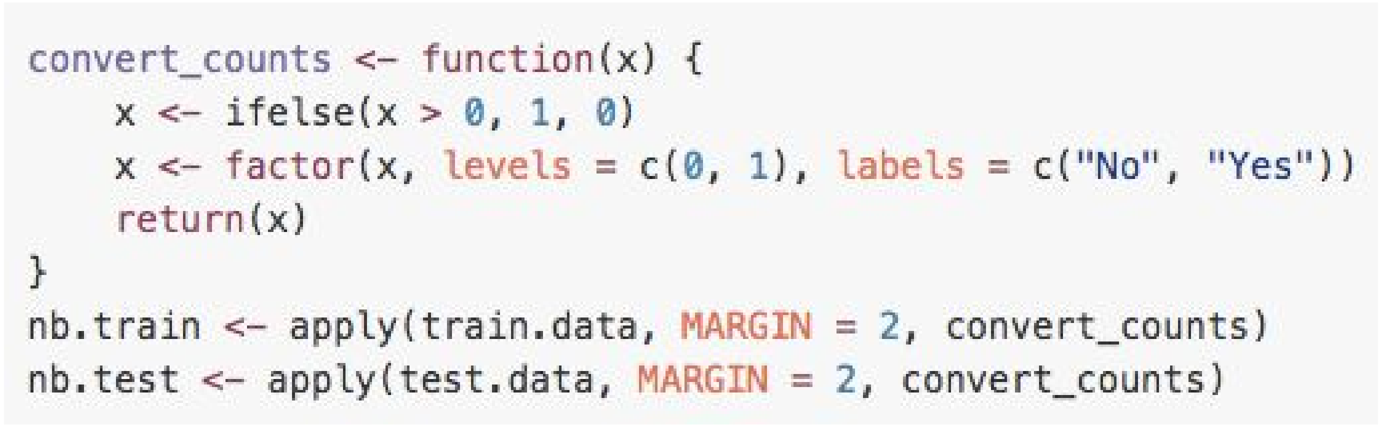
R	Haskell
`sum(1:5)`	`sum[1..5]`
`head(c(1,2,3),1)`	`head [1,2,3]`
`c(1,2,3)[1:3]`	`take(3) [1,2,3]`
`c(1,2,3)[-c(1,2)]`	`drop(2) [1,2,3]`
`c(l1, l2)`	`l1 ++ l2`
`rev(1:5)`	`reverse [1..5]`

In Knn, we did feature scaling with this line of code:



```r
normalize <- function(x){
    return ((x - min(x)) / (max(x) - min(x)))
}
knn.data.scaled <- as.data.frame(lapply(knn.data[1:2], normalize))
```

In Naive Bayes, we need to transform all factor values from 0, 1 to No, Yes:

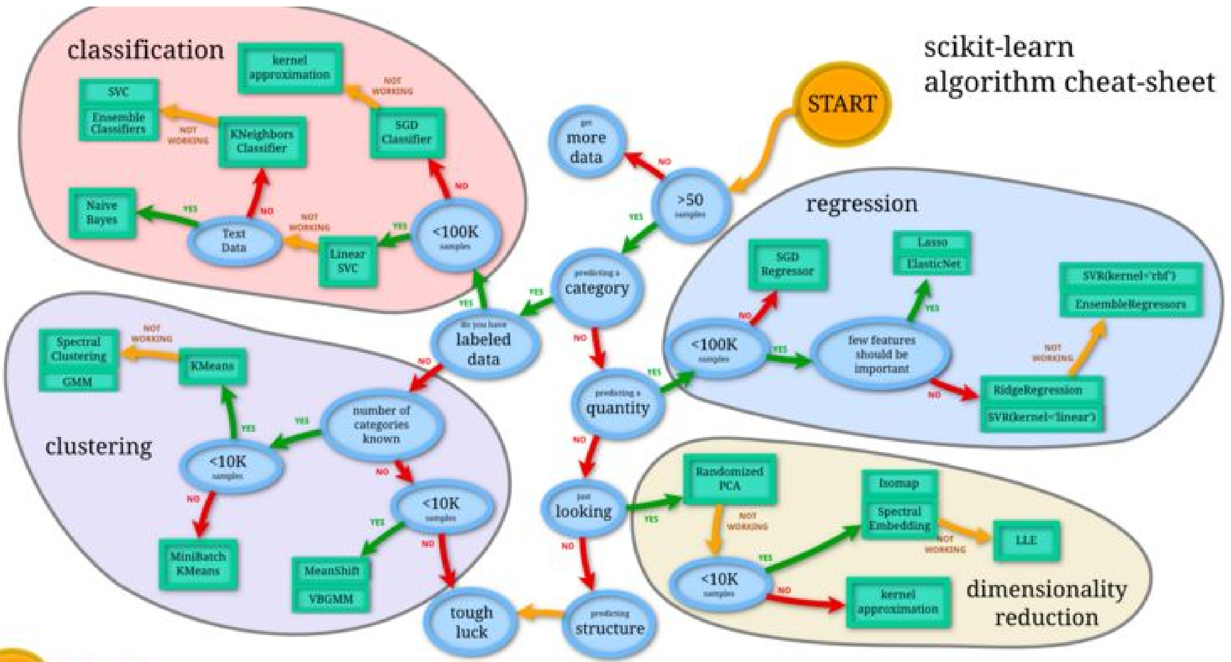


```r
convert_counts <- function(x) {
    x <- ifelse(x > 0, 1, 0)
    x <- factor(x, levels = c(0, 1), labels = c("No", "Yes"))
    return(x)
}
nb.train <- apply(train.data, MARGIN = 2, convert_counts)
nb.test <- apply(test.data, MARGIN = 2, convert_counts)
```

Machine Learning with Scikit-learn

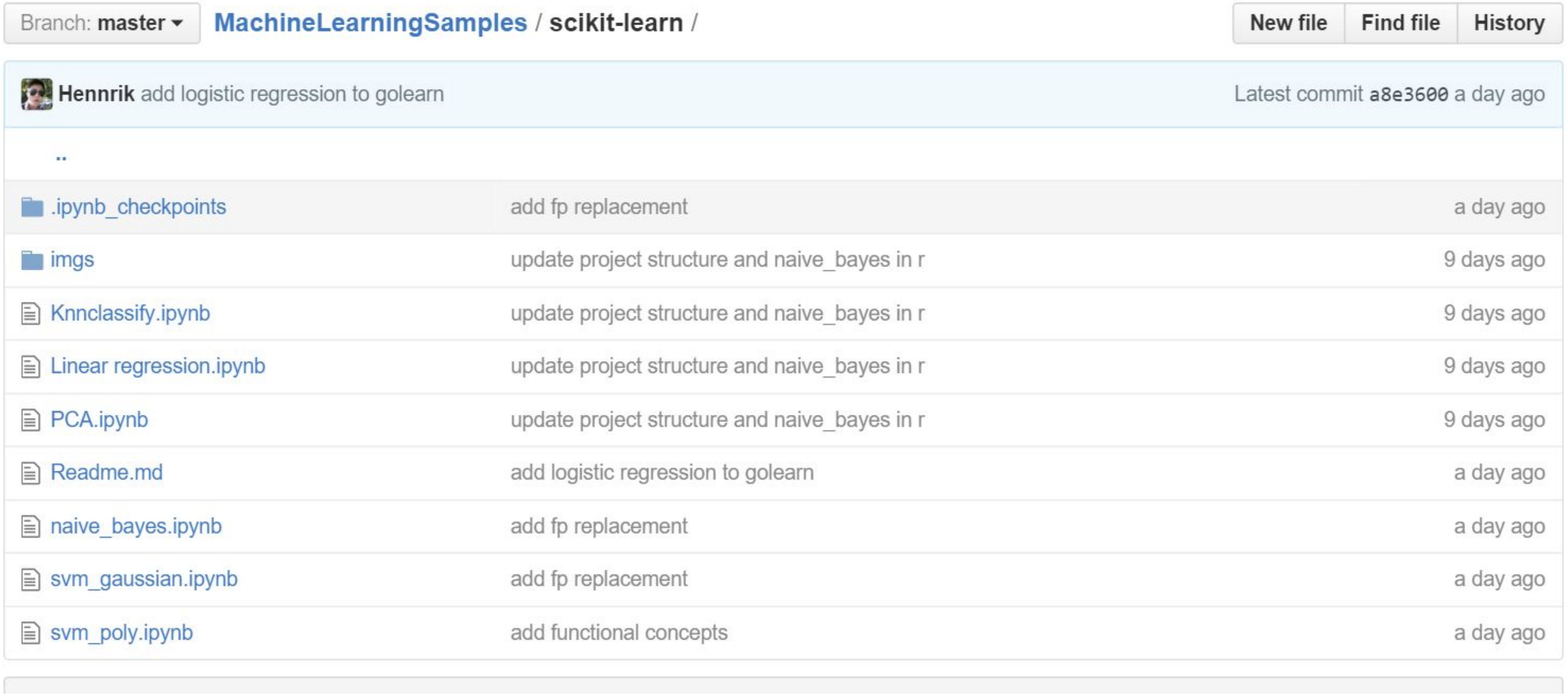
scikit learn

TU Delft


scikit-learn
algorithm cheat-sheet
START
>50 samples
get more data
predicting a category
do you have labeled data
predicting a quantity
just looking
predicting structure
tough luck
classification
<100K samples
SGD Classifier
kernel approximation
Linear SVC
Text Data
Naive Bayes
KNeighbors Classifier
SVC
Ensemble Classifiers
clustering
number of categories known
<10K samples
KMeans
Spectral Clustering
GMM
MiniBatch KMeans
MeanShift
VBGMM
regression
<100K samples
SGD Regressor
few features should be important
Lasso
ElasticNet
RidgeRegression
SVR(kernel='linear')
SVR(kernel='rbf')
EnsembleRegressors
dimensionality reduction
Randomized PCA
<10K samples
Isomap
Spectral Embedding
LLE
kernel approximation
YES
NO
NOT WORKING

Back

Example library: Scikit Learn



Branch: master ▾ MachineLearningSamples / scikit-learn /

New file | Find file | History

Hennrik add logistic regression to golearn — Latest commit a8e3600 a day ago

..

.ipynb_checkpoints	add fp replacement	a day ago
imgs	update project structure and naive_bayes in r	9 days ago
Knnclassify.ipynb	update project structure and naive_bayes in r	9 days ago
Linear regression.ipynb	update project structure and naive_bayes in r	9 days ago
PCA.ipynb	update project structure and naive_bayes in r	9 days ago
Readme.md	add logistic regression to golearn	a day ago
naive_bayes.ipynb	add fp replacement	a day ago
svm_gaussian.ipynb	add fp replacement	a day ago
svm_poly.ipynb	add functional concepts	a day ago

TU Delft

Example library: Scikit Learn

Map:

```
gammas = map(lambda x: 1.0/(2.0*x**2), sigmas)
```

filter:

```
dir_clean = filter(lambda x: (".DS_Store" not in x) and
("cmds" not in x), dir_content)
```

reduce:

```
FCR_min = reduce(lambda a,b: a if (a < b) else b, FCR)
```

list comprehension:

```
gammas = [ 1.0/(2.0*x**2) for x in sigmas ]
```

Problems we encountered

1. Library limitation.

2. Programming language challenge

3. Data pre-processing

TU Delft

We will use more functional programming in the future !