07_3_Machine_Learning_Models

Yue Xiong, LMU, yue.xiong@stat.uni-muenchen.de

Machine Learning – Model Training and Evaluation

Introduction

In this tutorial, we'll discuss how to formulate a policy problem or a social science question in the machine learning framework; how to transform raw data into something that can be fed into a model; how to build, evaluate, compare, and select models; and how to reasonably and accurately interpret model results. You'll also get hands-on experience using the `caret` package in R.

This tutorial is based on chapter "Machine Learning" of Big Data and Social Science.

Setup

```
library('dbplyr')  # load the corresponding libraries
library('dplyr')
library('RSQLite')
library('glue')
library('caret')
```

```
# Establish a connection to the ncdoc.db database
database_path = "F:/hiwi_work_notebook/ncdoc.db"
conn = DBI::dbConnect(SQLite(), database_path)
src_dbi(conn)
```

```
## src:  sqlite 3.36.0 [F:\hiwi_work_notebook\ncdoc.db]
## tbls: feature_age_2008, feature_age_2013, feature_age_first_admit,
##   feature_agefirstadmit, feature_length_sentence_2000_2008,
##   feature_length_sentence_2000_2013, feature_num_admits_2000_2008,
##   feature_num_admits_2000_2013, features_2000_2008, features_2000_2013, inmate,
##   offender, recidivism_labels_2009_2013, recidivism_labels_2014_2018,
##   sentences, sentences_prep, test_matrix, train_matrix
```

Problem Formulation

Our Machine Learning Problem Of all prisoners released, we would like to predict who is likely to reenter jail within *5* years of the day we make our prediction. For instance, say it is Jan 1, 2009 and we want to identify which prisoners are likely to re-enter jail between now and end of 2013. We can run our predictive model and identify who is most likely at risk. The is an example of a *binary classification* problem.

Note the outcome window of 5 years is completely arbitrary. You could use a window of 5, 3, 1 years or 1 day.

In order to predict recidivism, we will be using data from the `inmate` and `sentences` table to create labels (predictors, or independent variables, or X variables) and features (dependent variables, or Y variables).

We need to munge our data into **labels** (1_Machine_Learning_Labels.rmd) and **features** (2_Machine_Learning_Features.rmd) before we can train and evaluate **machine learning models** (3_Machine_Learning_Models.rmd).

This tutorial assumes that you have already worked through the `1_Machine_Learning_Labels` and `2_Machine_Learning_Features` tutorials. In the following chunk, we will recreate the two function `create_labels` and `create_features` again.

```
# create labels function
create_labels <- function(features_end, prediction_start, prediction_end, conn) {
  # @param features_end
  # @param prediction_start
  # @param prediction_end
  # @param conn: obj

  end_x_year = format(as.Date(features_end, format="%Y-%m-%d"),"%Y")
  start_y_year = format(as.Date(prediction_start, format="%Y-%m-%d"),"%Y")
  end_y_year = format(as.Date(prediction_end, format="%Y-%m-%d"),"%Y")

  drop_script_1 = "drop table if exists sentences_prep;"
  sql_script_1 = glue("create table sentences_prep as
select inmate_doc_number, cast(inmate_sentence_component as integer) as sentence_component, date([senter
date(actual_sentence_end_date) as sentence_end_date
from sentences;")

  drop_script_2 = glue("drop table if exists release_dates_2000_{end_x_year};")
  sql_script_2 = glue("create temp table release_dates_2000_{end_x_year} as
select inmate_doc_number, sentence_end_date
from sentences_prep where sentence_end_date >= '2000-01-01' and sentence_end_date <= '{features_end}';")

  drop_script_3 = glue("drop table if exists last_exit_2000_{end_x_year};")
  sql_script_3 = glue("create temp table last_exit_2000_{end_x_year} as select inmate_doc_number, max(s

  drop_script_4 = glue("drop table if exists admit_{start_y_year}_{end_y_year};")
  sql_script_4 = glue("create temp table admit_{start_y_year}_{end_y_year} as select inmate_doc_number,

  drop_script_5 = glue("drop table if exists recidivism_{start_y_year}_{end_y_year};")
  sql_script_5 = glue("create temp table recidivism_{start_y_year}_{end_y_year} as
select r.inmate_doc_number, r.sentence_end_date, a.sentence_begin_date, case when a.sentence_begin_date
left join admit_{start_y_year}_{end_y_year} a on r.inmate_doc_number = a.inmate_doc_number;")

  drop_script_6 = glue("drop table if exists recidivism_labels_{start_y_year}_{end_y_year};")
  sql_script_6 = glue("
create table recidivism_labels_{start_y_year}_{end_y_year} as
select distinct inmate_doc_number, recidivism
from recidivism_{start_y_year}_{end_y_year};")

  DBI::dbSendStatement(conn, drop_script_1)
  DBI::dbSendQuery(conn, sql_script_1)

  DBI::dbSendStatement(conn, drop_script_2)
  DBI::dbSendQuery(conn, sql_script_2)
```

```
  DBI::dbSendStatement(conn, drop_script_3)
  DBI::dbSendQuery(conn, sql_script_3)

  DBI::dbSendStatement(conn, drop_script_4)
  DBI::dbSendQuery(conn, sql_script_4)

  DBI::dbSendStatement(conn, drop_script_5)
  DBI::dbSendQuery(conn, sql_script_5)

  DBI::dbSendStatement(conn, drop_script_6)
  DBI::dbSendQuery(conn, sql_script_6)

  sql_query = glue("select * from recidivism_labels_{start_y_year}_{end_y_year}")
  df_label = data.frame(tbl(conn, sql(sql_query)))

  return(df_label)
}
```

```
# create features function
create_features <- function(features_end, prediction_start, prediction_end, conn) {
  # @param features_end
  # @param prediction_start
  # @param prediction_end
  # @param conn: obj
  end_x_year = format(as.Date(features_end, format="%Y-%m-%d"),"%Y")
  start_y_year = format(as.Date(prediction_start, format="%Y-%m-%d"),"%Y")
  end_y_year = format(as.Date(prediction_end, format="%Y-%m-%d"),"%Y")

  drop_script_1 = "drop table if exists sentences_prep;"
  sql_script_1 = glue("create table sentences_prep as
select inmate_doc_number,
cast(inmate_sentence_component as integer) as sentence_component,
date([sentence_begin_date_(for_max)]) as sentence_begin_date,
date(actual_sentence_end_date) as sentence_end_date
from sentences;")

  drop_script_2 = glue("drop table if exists feature_num_admits_2000_{end_x_year};")
  sql_script_2 = glue("create table feature_num_admits_2000_{end_x_year} as
select inmate_doc_number, count(*) num_admits
from sentences_prep
where inmate_doc_number in (select inmate_doc_number from recidivism_labels_{start_y_year}_{end_y_year}
and sentence_begin_date < '{features_end}' and sentence_component = 1
group by inmate_doc_number;")

  drop_script_3 = glue("drop table if exists feature_length_sentence_2000_{end_x_year};")
  sql_script_3 = glue("create table feature_length_sentence_2000_{end_x_year} as
select inmate_doc_number, sentence_component, cast(julianday(sentence_end_date) - julianday(sentence_beg
from sentences_prep
where inmate_doc_number in (select inmate_doc_number from recidivism_labels_{start_y_year}_{end_y_year}
and sentence_begin_date < '{features_end}' and sentence_component = 1
and sentence_begin_date > '0001-01-01' and sentence_end_date > '0001-01-01' and sentence_end_date > sen

  drop_script_4 = glue("drop table if exists feature_length_long_sentence_2000_{end_x_year};")
```

```
  sql_script_4 = glue("create temp table feature_length_long_sentence_2000_{end_x_year} as
select inmate_doc_number, max(length_sentence) length_longest_sentence
from feature_length_sentence_2000_{end_x_year}
group by inmate_doc_number;")

  drop_script_5 = "drop table if exists docnbr_admityr;"
  sql_script_5 = "create temp table docnbr_admityr as
select inmate_doc_number, min(sentence_begin_date) min_admityr
from sentences_prep
where sentence_begin_date > '0001-01-01'
group by inmate_doc_number;"

  drop_script_6 = "drop table if exists age_first_admit_birth_year;"
  sql_script_6 = 'create temp table age_first_admit_birth_year as
select da.inmate_doc_number,
cast(strftime("%Y", da.min_admityr) as integer) min_admityr,
cast(strftime("%Y", p.inmate_birth_date) as integer) inmate_birth_date
from docnbr_admityr da
left join inmate p on da.inmate_doc_number = p.inmate_doc_number;'

  drop_script_7 = "drop table if exists feature_age_first_admit;"
  sql_script_7 = "create table feature_age_first_admit as
select inmate_doc_number, (min_admityr - inmate_birth_date) age_first_admit
from age_first_admit_birth_year;"

  drop_script_8 = "drop table if exists feature_agefirstadmit;"
  sql_script_8 = glue("create table feature_agefirstadmit as
select inmate_doc_number, age_first_admit
from feature_age_first_admit
where inmate_doc_number in (select inmate_doc_number from recidivism_labels_{start_y_year}_{end_y_year}

  drop_script_9 = glue("drop table if exists feature_age_{end_x_year};")
  sql_script_9 = glue('create table feature_age_{end_x_year} as
select inmate_doc_number, ({end_x_year} - cast(strftime("%Y", inmate_birth_date) as integer)) age
from inmate
where inmate_doc_number in (select inmate_doc_number from recidivism_labels_{start_y_year}_{end_y_year}

  drop_script_10 = glue("drop table if exists features_2000_{end_x_year};")
  sql_script_10 = glue('create table features_2000_{end_x_year} as
select f1.inmate_doc_number, f1.num_admits, f2.length_longest_sentence, f3.age_first_admit, f4.age
from feature_num_admits_2000_{end_x_year} f1
left join feature_length_long_sentence_2000_{end_x_year} f2 on f1.inmate_doc_number = f2.inmate_doc_numb
left join feature_agefirstadmit f3 on f1.inmate_doc_number = f3.inmate_doc_number
left join feature_age_{end_x_year} f4 on f1.inmate_doc_number = f4.inmate_doc_number;')

  DBI::dbSendStatement(conn, drop_script_1)
  DBI::dbSendStatement(conn, sql_script_1)

  DBI::dbSendStatement(conn, drop_script_2)
  DBI::dbSendStatement(conn, sql_script_2)

  DBI::dbSendStatement(conn, drop_script_3)
  DBI::dbSendStatement(conn, sql_script_3)
```

```
  DBI::dbSendStatement(conn, drop_script_4)
  DBI::dbSendStatement(conn, sql_script_4)

  DBI::dbSendStatement(conn, drop_script_5)
  DBI::dbSendStatement(conn, sql_script_5)

  DBI::dbSendStatement(conn, drop_script_6)
  DBI::dbSendStatement(conn, sql_script_6)

  DBI::dbSendStatement(conn, drop_script_7)
  DBI::dbSendStatement(conn, sql_script_7)

  DBI::dbSendStatement(conn, drop_script_8)
  DBI::dbSendStatement(conn, sql_script_8)

  DBI::dbSendStatement(conn, drop_script_9)
  DBI::dbSendStatement(conn, sql_script_9)

  DBI::dbSendStatement(conn, drop_script_10)
  DBI::dbSendStatement(conn, sql_script_10)

  sql_query = glue("select * from features_2000_{end_x_year}")
  df_features = data.frame(tbl(conn, sql(sql_query)))

  return(df_features)}
```

Create Training and Test Sets

Our Training Set

We create a training set that takes people at the beginning of 2009 and defines the outcome based on data from 2009-2013 (`recidivism_labels_2009_2013`). The features for each person are based on data up to the end of 2008 (`features_2000_2008`).

Note: It is important to segregate your data based on time when creating features. Otherwise there can be "leakage", where you accidentally use information that you would not have known at the time.

```
sql_string = "drop table if exists train_matrix;"
DBI::dbSendStatement(conn, sql_string)
```

```
## <SQLiteResult>
##   SQL  drop table if exists train_matrix;
##   ROWS Fetched: 0 [complete]
##        Changed: 0
```

```
sql_string = "
create table train_matrix as
select l.inmate_doc_number, l.recidivism, f.num_admits, f.length_longest_sentence, f.age_first_admit, f
from recidivism_labels_2009_2013 l
left join features_2000_2008 f on f.inmate_doc_number = l.inmate_doc_number;"
DBI::dbSendQuery(conn, sql_string)
```

```
## Warning: Closing open result set, pending rows
```

```
## <SQLiteResult>
##   SQL
## create table train_matrix as
## select l.inmate_doc_number, l.recidivism, f.num_admits, f.length_longest_sentence, f.age_first_admit
## from recidivism_labels_2009_2013 l
## left join features_2000_2008 f on f.inmate_doc_number = l.inmate_doc_number;
##   ROWS Fetched: 0 [complete]
##        Changed: 0
```

We then load the training data into `df_training`.

```
sql_string = "SELECT * FROM train_matrix"

df_training = data.frame(tbl(conn, sql(sql_string)))
```

```
## Warning: Closing open result set, pending rows
```

```
head(df_training, n=5)
```

```
##   INMATE_DOC_NUMBER recidivism num_admits length_longest_sentence
## 1           0000028          0          3                     1097
## 2           0000062          0          2                      273
## 3           0000114          0          2                      256
## 4           0000133          0          1                     6209
## 5           0000152          1          1                       76
##   age_first_admit age
## 1              19  36
## 2              36  51
## 3              49  56
## 4              23  39
## 5              37  43
```

Our Test (Validation) Set

In the machine learning process, we want to build models on the training set and evaluate them on the test set. Our test set will use labels from 2014-2018 (`recidivism_labels_2014_2018`), and our features will be based on data up to the end of 2013 (`features_2000_2013`).

```
sql_string = "drop table if exists test_matrix;"
DBI::dbSendStatement(conn, sql_string)
```

```
## <SQLiteResult>
##   SQL  drop table if exists test_matrix;
##   ROWS Fetched: 0 [complete]
##        Changed: 0
```

```
sql_string = "
create table test_matrix as
select l.inmate_doc_number, l.recidivism, f.num_admits, f.length_longest_sentence, f.age_first_admit, f
from recidivism_labels_2014_2018 l
left join features_2000_2013 f on f.inmate_doc_number = l.inmate_doc_number;"
DBI::dbSendQuery(conn, sql_string)
```

```
## Warning: Closing open result set, pending rows
```

```
## <SQLiteResult>
##   SQL
## create table test_matrix as
## select l.inmate_doc_number, l.recidivism, f.num_admits, f.length_longest_sentence, f.age_first_admit
## from recidivism_labels_2014_2018 l
## left join features_2000_2013 f on f.inmate_doc_number = l.inmate_doc_number;
##   ROWS Fetched: 0 [complete]
##        Changed: 0
```

We load the test data into `df_test`.

```
sql_string = "SELECT * FROM test_matrix"

df_test = data.frame(tbl(conn, sql(sql_string)))
```

```
## Warning: Closing open result set, pending rows
```

```
head(df_test, n=5)
```

```
##   INMATE_DOC_NUMBER recidivism num_admits length_longest_sentence
## 1           0000028          0          3                     1097
## 2           0000033          0          5                     5186
## 3           0000035          0          3                     3969
## 4           0000037          0          1                      137
## 5           0000062          0          2                      273
##   age_first_admit age
## 1              19  41
## 2              18  53
## 3              25  47
## 4              34  38
## 5              36  56
```

Data Cleaning

Before we proceed to model training, we need to clean our training data. First, we check the percentage of missing values.

```
# count the number of missing values row wise
isnan_training_rows = sum(!complete.cases(df_training))

nrows_training = nrow(df_training)
```

```
# count the missing value percentage
missing_percent = isnan_training_rows/nrows_training
missing = glue("missing rows ratio:{missing_percent}")
cat(missing)
```

```
## missing rows ratio:0.0095743866009651
```

We see that about 1% of the rows in our data have missing values. In the following, we will drop rows with missing values. Note, however, that better ways for dealing with missings exist, e.g., general data imputation methods.

```
df_training = na.omit(df_training)
cleaned_training = nrow(df_training) # number of rows after cleaning
```

```
c(nrows_training, cleaned_training) # checking whether the missing data are dropped
```

```
## [1] 156668 155168
```

Let's check if the values of the ages at first admit are reasonable.

```
length(unique(df_training$age_first_admit))
```

```
## [1] 77
```

```
length(df_training$age_first_admit)
```

```
## [1] 155168
```

Looks like this needs some cleaning. We will drop any rows that have age < 14 and > 99.

```
df_training <- df_training[ which(df_training$age_first_admit >= 14
& df_training$age_first_admit <= 99), ]
```

Let's check how much data we still have and how many examples of recidivism are in our training dataset. When it comes to model evaluation, it is good to know what the "baseline" is in our dataset.

```
row_after_cleaning = glue("Number of rows in training dataset: {nrow(df_training)}")
cat(row_after_cleaning)
```

```
## Number of rows in training dataset: 155159
```

```
count(df_training, df_training$recidivism, sort=TRUE)
```

```
##   df_training$recidivism      n
## 1                      0 117386
## 2                      1  37773
```

We have about 155,000 examples, and about 25% (37773/155159) of those are positive examples (recidivist), which is what we're trying to identify. About 75% (117386/155159) of the examples are negative examples (non-recidivist).

In the next step, let's take a look at the test set.

```
# count the number of missing values row wise
isnan_test_rows = sum(!complete.cases(df_test))

nrows_test = nrow(df_test)

# count the missing value percentage
missing_percent_test = isnan_test_rows/nrows_test
missing_test = glue("missing rows in test ratio:{missing_percent_test}")
cat(missing_test)
```

```
## missing rows in test ratio:0.00949937563855148
```

We see that about 1% of the rows in our test set have missing values. This matches what we'd expect based on what we saw in the training set.

Similarly, we drop cases with age < 14 and > 99 in the test set.

```
df_test <- na.omit(df_test)
df_test <- df_test[ which(df_test$age_first_admit >= 14
& df_test$age_first_admit <= 99), ]
```

We also check the number of observations and the outcome distribution for our test data.

```
rows_after_cleaning = glue("Number of rows in test dataset: {nrow(df_test)}")
cat(row_after_cleaning)
```

```
## Number of rows in training dataset: 155159
```

```
count(df_test, df_test$recidivism, sort=TRUE)
```

```
##   df_test$recidivism      n
## 1                  0 184318
## 2                  1  33805
```

Split into features and labels

Here we select our features and outcome variable.

```
sel_features = c('num_admits', 'length_longest_sentence', 'age_first_admit', 'age')
sel_label = 'recidivism'
```

We can now create an X- and y-training and X- and y-test object to train and evaluate prediction models with `caret`.

```
X_train = df_training[sel_features]
y_train = df_training[sel_label]
X_test = df_test[sel_features]
y_test = df_test[sel_label]
```

Model Training

On this basis, we can now build a prediction model that learns the relationship between our predictors (`X_train`) and recidivism (`y_train`) in the training data. We start with using logistic regression as our first model.

```
DBI::dbDisconnect(conn)
```