Машинное обучение в R

Пакет caret

CARET - Classification And REgression Training

install.packages(“caret”)

library(caret)

Книга о пакете caret:
https://topepo.github.io/caret/

Линейная регрессия

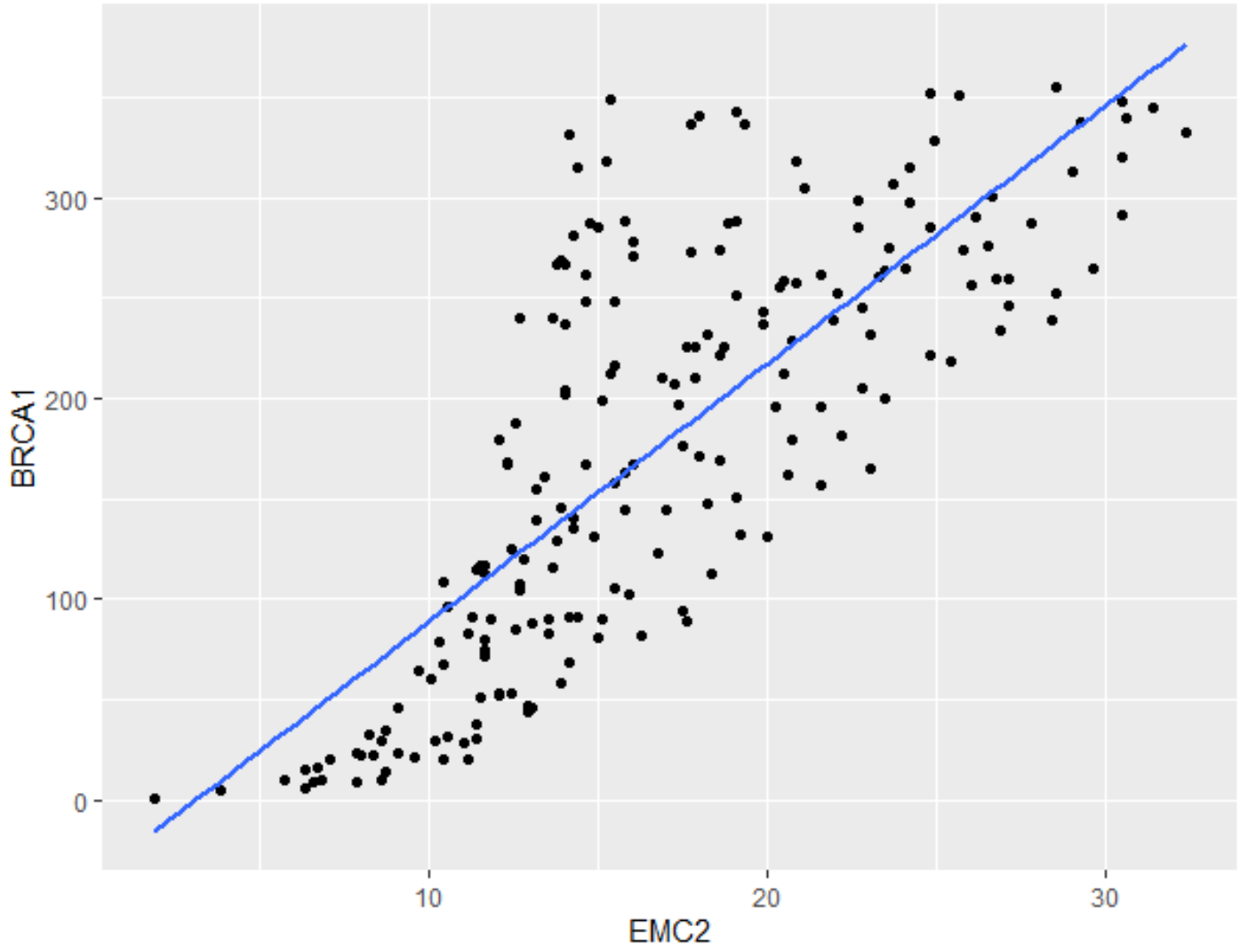
$Y_i = \beta_0 + \beta_1 x_i$

β_0 — сдвиг (пересечение с осью Y)

β_1 — наклон прямой Y

x_i — значение переменной X в i-м наблюдении

Задача – предсказать экспрессию BRCA1



Воспроизводимость результатов

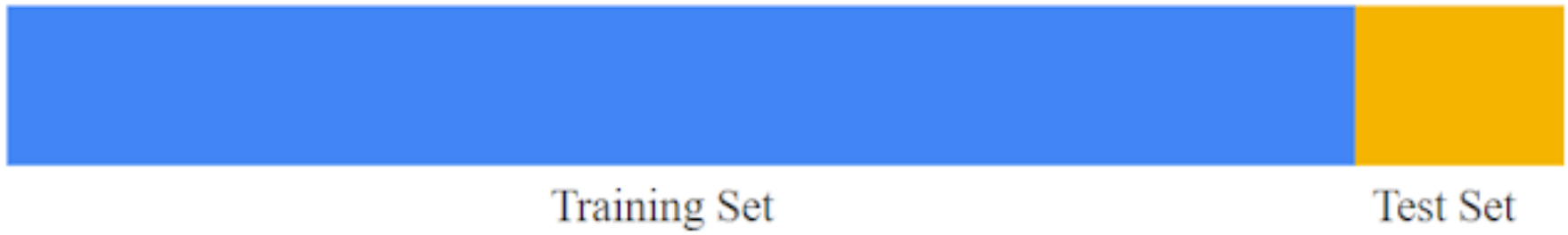
set.seed(123)

В практике статистического анализа данных часто приходится иметь дело с необходимостью генерации случайных чисел, подчинящихся тому или иному закону распределения вероятностей (например, при необходимости случайным образом отобрать небольшую выборку из массивной таблицы данных, при использовании бутстреп-методов, методов Монте-Карло, и т.п.).

Генератор псевдослучайных чисел начинает свою работу с определенной точки в пространстве возможных чисел. Эта точка называется *начальным числом* (англ. *seed*). В R имеется возможность зафиксировать это число так, что при повторном использовании ГПСЧ будет генерироваться точно та же последовательность чисел, что и в первый раз. Это может оказаться полезным в случаях, когда исследователь (по тем или иным причам) желает иметь точную воспроизводимость результатов, получаемых с задействованием ГПСЧ.

Разделение выборки на тренировочный и тестовый наборы

```
training.samples <- marketing$sales %>%
    createDataPartition(p = 0.8, list = FALSE)
train.data <- marketing[training.samples, ]
test.data <- marketing[-training.samples, ]
```



Модель

```
model <- train(BRCA1 ~ EMC2, data = train.data,
method = "lm")

summary(model)
```

Предсказание

```
predictions <- model %>% predict(test.data)
```

Качество модели

- Prediction error, RMSE - среднеквадратичная ошибка – чем меньше, тем лучше

RMSE(predictions, test.data$BRCA1)

- R-square – коэффициент детерминации - чем ближе к 1, тем лучше

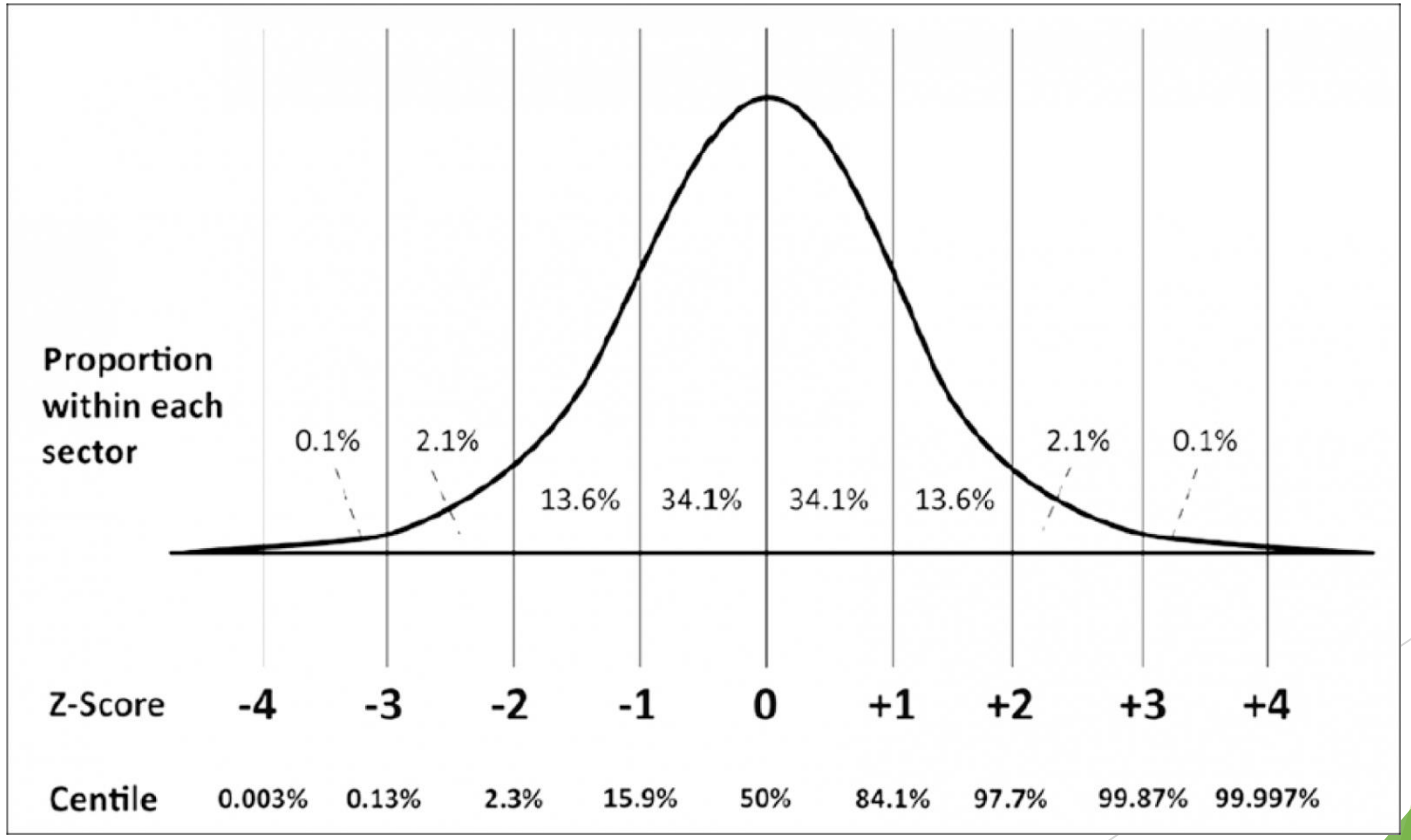
R2(predictions, test.data$BRCA1)

- Mean absolute error, MAE - чем меньше, тем лучше

MAE(predictions, test.data$BRCA1)

Препроцессинг - нормализация

```
model <- train(BRCA1 ~ EMC2, data = train.data, method
= "lm", preProcess = c('scale', 'center'))
```



Cross validation

```
fitControl <- trainControl(method = "repeatedcv",
                           number = 10,     # number of folds
                           repeats = 10)

model <- train(BRCA1 ~ EMC2,
               data = train.data,
               method = "lm",
               trControl = fitControl,
               preProcess = c('scale', 'center'))
```

Множественная линейная регрессия

$$Y_i = \beta_0 + \beta_1 x_i$$

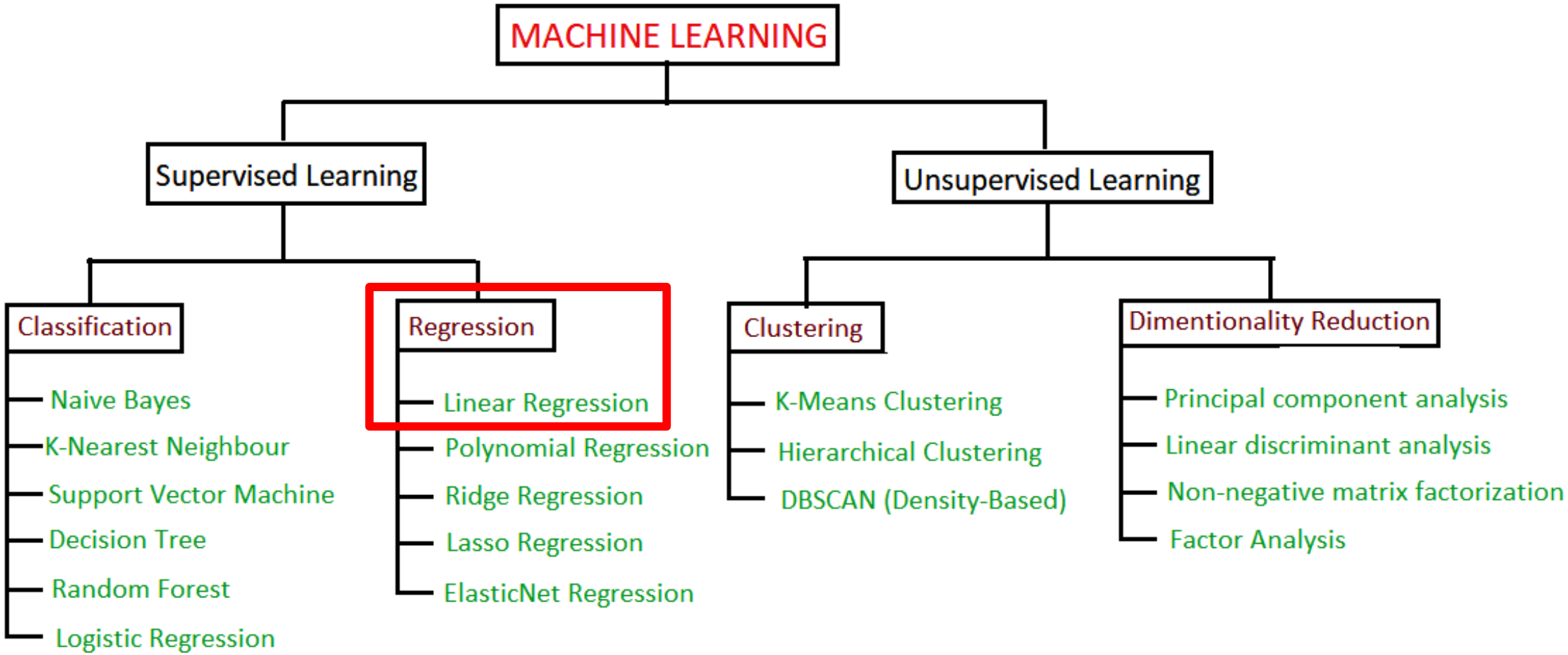
k *предикторов*

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_{\boldsymbol{k}} x_{i\boldsymbol{k}}$$

Множественная линейная регрессия

```
model_mult <- train(BRCA1 ~ .,
                    data = train.data,
                    method = "lm",
                    trControl = fitControl,
                    preProcess = c('scale', 'center'))
```

BRCA1 ~ EMC2 + NBN + BRCA2


MACHINE LEARNING
Supervised Learning
Unsupervised Learning
Classification
Regression
Clustering
Dimentionality Reduction
Naive Bayes
K-Nearest Neighbour
Support Vector Machine
Decision Tree
Random Forest
Logistic Regression
Linear Regression
Polynomial Regression
Ridge Regression
Lasso Regression
ElasticNet Regression
K-Means Clustering
Hierarchical Clustering
DBSCAN (Density-Based)
Principal component analysis
Linear discriminant analysis
Non-negative matrix factorization
Factor Analysis