Silesian University of Technology

Programming in R and Python

Lecture 6 - R for high-dimensional data analysis - part 2

Anna Papież
Department of Data Science and Engineering

Homework practice

Practice

1) There are 10 multiple choice questions at an exam with only one correct answer out of four. You need to score at least 5 points to pass. What is the probability of failing if you choose all the answers at random? (Calculate using the binomial distribution.)
2) In the Auto dataset (ISLR package), check if there is a significant difference in mileage between Dodges and Toyotas.

Practice

Load the Auto dataset.

1) Build a regression model of mpg as a function of horsepower, dividing the dataset 50:50 into a training and test set. Calculate the MSE.
2) Perform LOO crossvalidation on the dataset. Use the `glm` function for building the model and the `cv.glm` function from the `boot` package for obtaining estimates of the prediction error.
3) Perform 10-fold crossvalidation on the dataset. Estimate the prediction error as in (2).

Classification:
prediction of categorical response

Classification

Regression involves predicting continuous-valued response, like tumor size.

Classification involves predicting categorical response:

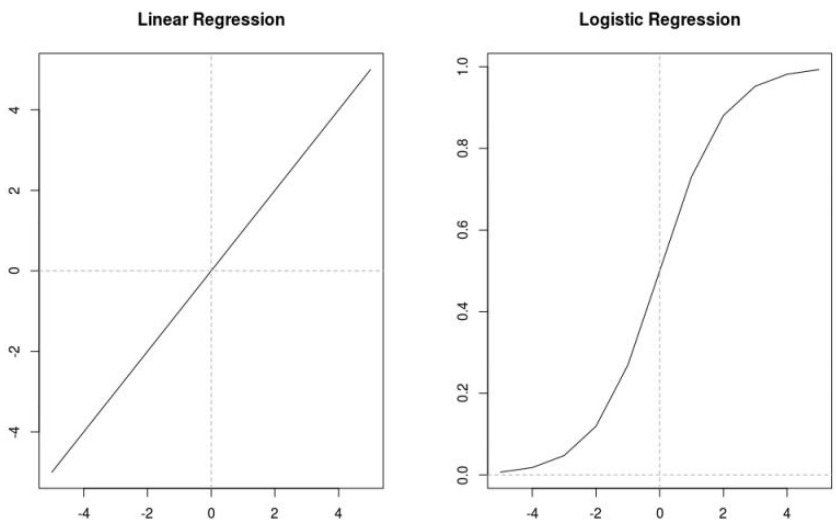
- Cancer versus Normal
- Tumor Type 1 versus Tumor Type 2.

Logistic regression

Straightforward extension of linear regression to the classification setting, for simplicity, suppose a two-class problem.

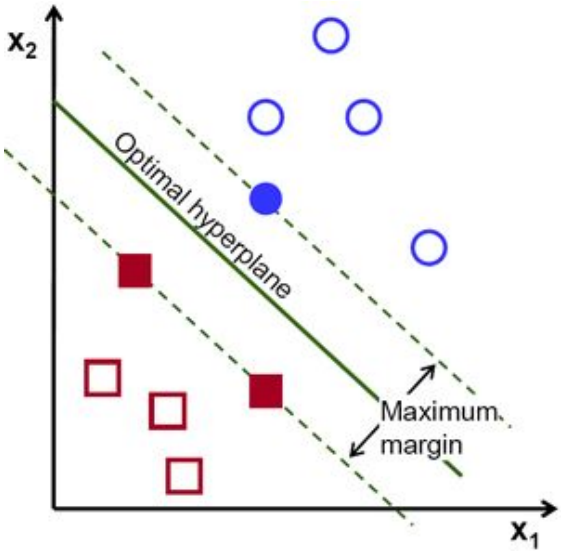
Model fit with maximum likelihood.

$$P(y=1|X) = \frac{exp(X^T\beta)}{1+exp(X^T\beta)}$$



Support Vector Machine

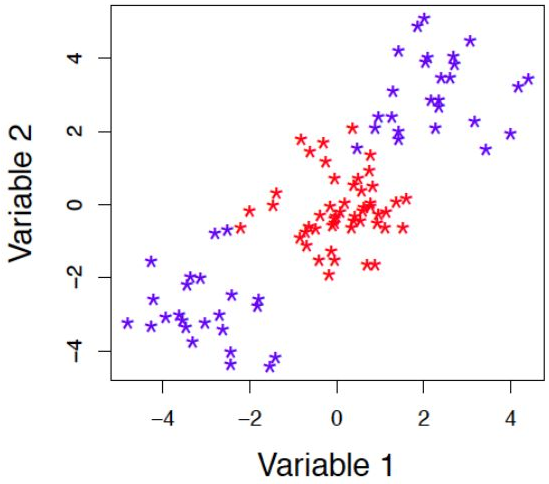
Find a separating hyperplane



Support Vector Machine

If a linear separating hyperplane doesn't exist we may:

1. allow for violations
2. use non-linear kernel



Practice

Load the Smarket dataset.

1) Fit a logistic regression model using all the `Lag` variables and `Volume`.
2) Estimate the model accuracy using the `predict` function.
3) Now use all the observations from 2005 as a test set. Fit the model again and see how the prediction works this time.

Repeat the same using a Support Vector Machine model. Experiment with different cost values.

Clustering analysis

Clustering analysis

Finding homogeneous subgroups among observations - objects in one cluster are more similar to each other than objects in other clusters.

What does similar mean?

Dissimilarity measures

Euclidean

$$\sqrt{\sum_{k=1}^{p}(X_{ik} - X_{jk})^2)}$$

Manhattan

$$\sum_{k=1}^{p}|X_{ik} - X_{jk}|$$

Mahalanobis

$$(X_i - X_j)^T \sum^{-1}(X_i - X_j)$$

...

Similarity measures

Correlation coefficients:
Pearson's

$$r = r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Spearman's
Kendall's
...

Hierarchical clustering

Hierarchical clustering results in a sequence of solutions (nested clusters), organized in a hierarchical tree structure, called the dendrogram

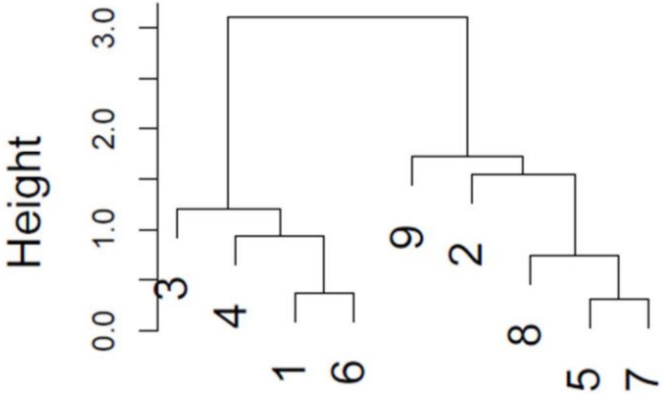
Bottom-Up:

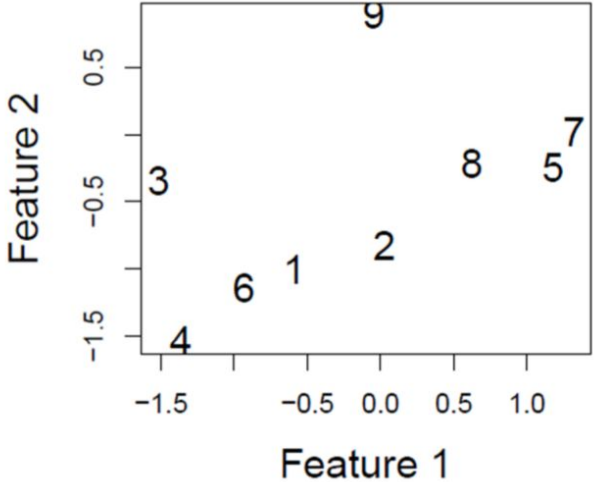
- Start from n individual clusters
- At each step, merge the closest pair of clusters until all objects form a single cluster

Top-Down:

- Start from 1 cluster
- At each step, split the most heterogeneous cluster until every cluster has only one member

Dendrogram


Height
0.0
1.0
2.0
3.0
3
4
1
6
9
2
8
5
7


Feature 2
−1.5
−0.5
0.5
Feature 1
−1.5
−0.5
0.0
0.5
1.0
9
7
3
8
5
2
1
6
4

Inter-cluster similarity

Linkage:
Single - minimum distance between points in two clusters is used to determine which two clusters should be merged
Complete - maximum distance between points in two clusters is used to determine which two clusters should be merged
Average - the average distance between points in two clusters is used to determine which two clusters should be merged

RStudio®

dplyr

knitr

ggplot2

rmarkdown

Shiny

K-means clustering

Partition-based method - minimizing within cluster variation

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$

K-means clustering

Finding exact solutions (global minimum) is not tractable.
However, we can efficiently find good approximate solutions for this problem (local minimum) using the following algorithm:

1. Randomly assign each observation to one of K clusters.
2. Iterate until the cluster assignments don't change:
   (a) For each of the K clusters, compute the cluster centroid, i.e. the mean of the observations assigned to the each cluster. This is a vector of length p (for p features).
   (b) Assign each observation to the cluster with closest centroid (based on Euclidean distance).

K-means clustering

Evaluating the Quality of a Clustering

Cluster homogeneity

Within sum of squares (WSS)

For each object the "error" is the distance to its cluster centroid:

$$\sum_{k=1}^{K} \sum_{i \epsilon C_k} d^2(m_k; X_i)$$

Cluster separation

Between sum of squares (BSS)

For each cluster the "error" is the distance between the cluster centroid and the grand mean:

$$\sum_{k=1}^{K} d^2(m_k; m)$$

RStudio®

dplyr

knitr

ggplot2

rmarkdown

Shiny

Practice

Use the following commands to simulate a dataset:

```
x=matrix(rnorm(50*2), ncol=2)
x[1:25,1]=x[1:25,1]+3
x[1:25,2]=x[1:25,2]-4
```


1) Perform k-means clustering on the dataset with two and three clusters. Experiment with the `nstart` parameter. Visualize the data on a scatterplot.
2) Perform hierarchical clustering with all linkage methods. Plot the resulting dendrograms. Use the `cutree` function to experiment with cutoff thresholds.

R or Python?

Analysis Tool	Similar Superhero	Super Powers in Common
R	Batman	• Detective Work • Intelligence • Cunning • Usage of Tools • More Brain than Muscles
Python	Superman	• Muscle Power • Super Strength • Elegance • Wide Range • More Muscles than Brain

Choice's up to you

But remember:

Parallelization

mclapply {parallel}

Parallel version of lapply. Applies a function to each list element, returns list

```
mclapply(L, FUN)
L              list
FUN            function
```

mclapply

```
library(parallel)
L <- list(a = 1, b = 1:3, c = 10:100)
mclapply(L, length)
   $a
   [1] 1
   $b
   [1] 3
   $c
   [1] 91
```

Code parallelization in R

One can use the `%dopar%` function to parallelize for loops. The result returned is a list:

```
library(doParallel)
cl <- makeCluster(2)
registerDoParallel(cl)
foreach(i=1:3) %dopar% sqrt(i)
stopCluster(cl)
```

RStudio®

dplyr

knitr

ggplot2

rmarkdown

Shiny

Practice

Load the ChickWeight dataset.

1) Use a grouping function to determine which variables could serve as grouping variables (hint: use the unique function).
2) Use these grouping variables to summarise the basic statistics of chick weight in corresponding groups.
3) Use the weight and diet variables to construct multiple logistic regression models. Perform 10000 trials sampling 300 observations out of all possible. Compare the runtimes of a **`for`** loop, **`lapply`** function and a two-core parallel run with **`%dopar%`**.

Silesian
University
of Technology

I APPRECIATE YOUR ATTENTION

Have a good Easter time!