Biomedical Data Science & AI

Exercise sheet 5

SoSe2025

Prof. Dr. Holger Fröhlich, Jonas Botz, Dr. Shammi More, Diego Valderrama

Due date: May13th

Exercise 1 - Logistic Regression - Theory (6 points)

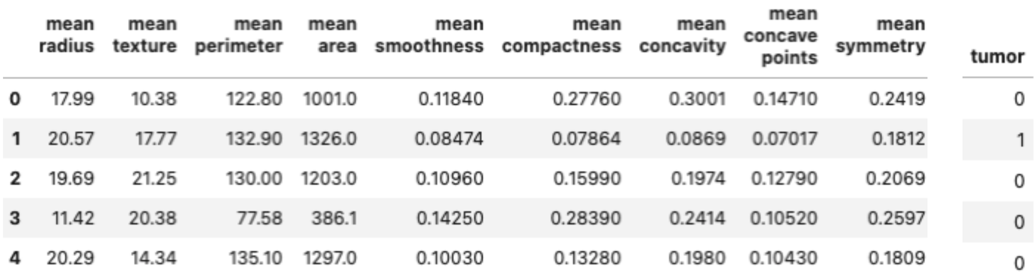
You want to model the probability of lung cancer given a patient's years of smoking as well as age.

1. Considering the equation of the logistic regression:

$$\mathrm{P}(Y=1 \mid X=x_i) = \mathrm{P}(Y_i=1) = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^\top \boldsymbol{\beta})} = \frac{1}{1+\exp(-\mathbf{x}_i^\top \boldsymbol{\beta})},$$

   a. What are X and Y in this case? **(1 point)**
   b. What do the outputs of the logistic model, and the logistic function represent in general? What are their ranges? **(1 point)**

2. Consider the following dataset on breast cancer. The target variable 'tumor' defines whether it is a benign (0) tumor or a malignant (1) tumor.



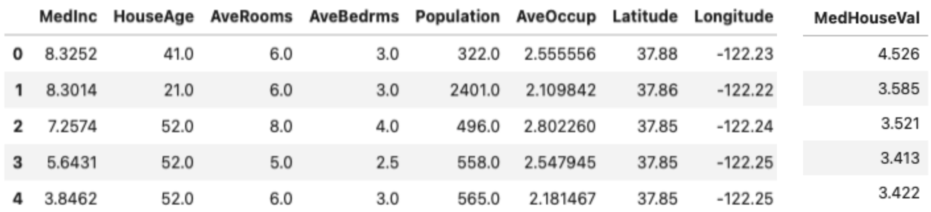
	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	tumor
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419	0
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812	1
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069	0
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597	0
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809	0

We want to assess the statistical significance of the predictor mean radius. Let us say we have two different models to estimate the target variable:

- Model 1 has all nine predictor variables.
- Model 2 has eight predictor variables, all but the 'mean radius'.

   a. What is our null hypothesis here? Which statistical test would you apply to compare the fit of the two models? **(1 point)**
   b. Which result of the statistical test would let you conclude whether the predictor variable 'mean radius' is statistically significant or not? (1 **point)**

3. Considering this example housing dataset. The input variables (MedInc, HouseAge, … etc.) as well as the target variable (MedHouseVal) are shown. Which assumptions must be fulfilled to apply logistic regression? Are they fulfilled in this example? **(2 points)**



	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude
0	8.3252	41.0	6.0	3.0	322.0	2.555556	37.88	-122.23
1	8.3014	21.0	6.0	3.0	2401.0	2.109842	37.86	-122.22
2	7.2574	52.0	8.0	4.0	496.0	2.802260	37.85	-122.24
3	5.6431	52.0	5.0	2.5	558.0	2.547945	37.85	-122.25
4	3.8462	52.0	6.0	3.0	565.0	2.181467	37.85	-122.25

MedHouseVal
4.526
3.585
3.521
3.413
3.422

Exercise 2 - ANOVA F-test and Hierarchical Clustering (12 points)

Load the leukemia dataset. It contains gene expression data of 1397 genes from 38 tumor mRNA samples. The expression data is organized in a matrix where rows correspond to genes and columns to samples. The tumor class of the columns is given in the file “golub.cl”.

1. ANOVA F-test
   a. What are the assumptions of the ANOVA F-test? **(1 point)**
   b. For each gene in the dataset, perform the ANOVA F-test (assumptions are already met) to see whether the gene is significantly differentially expressed between the two types of Leukemia. **(2 point)**
   c. Due to our analysis, we now know which genes are significantly differentially expressed between groups. These will be the best features to use in order to get good cluster separation. Subset only the rows which represent the top 100 most significant genes. **(1 point)**
2. Plot 2 dendrograms using the 100 selected genes:
   a. One for a single linkage approach and another one for ward approach. **(2 point)**
   b. Which method would you recommend based on the dendrograms for a clustering? Why? **(1 point)**
   c. Familiarize yourself with Cophenetic correlation coefficient and calculate the cophenetic correlation distance for both single linkage as well as ward. **(2 point)**
   d. Based on the cophenetic correlation distance, which clustering method performed better? **(1 point)**
3. Apply two Agglomerative Clustering.
   a. One using single linkage and one using ward method. **(2 point)**

Exercise 3 - PCA (7 Points)

Load the dataset (g*olub.csv*). Start again with all genes and generate a feature matrix (transposed leukemia dataset) and a class label variable (*golub.cl.csv*).

1. Perform a PCA on the feature matrix and answer the following:
    a. Create a combined plot displaying both the scree plot and the cumulative explained variance to illustrate how many principal components are required to explain at least 95% of the total variance? **(2 points)**
    b. Make a scatterplot of the projections on the first two PC‘s. Color the plot according to the class labels. **(1 points)**
    c. Inform yourself about the UMAP, obtain the projection plot and compare with PCA results **(2 points)**
    d. Based on the scatterplot, answer the following questions **(2 points)**
        i. Given the different plots, which of the previous techniques do you think would be the better choice?
        ii. Do you think n=2 components are a good choice? Why?