Exercises Linear Classification

1 Logistic Regression

We consider a research study on the behaviour of customers of a web shop and the relationship between sales and appreciation of the website. A number of visitors to the website were asked to mark their appreciation of the website on a 5 point Likert scale, ranging from 1 (very bad) to 5 (very good). For these visitors was also recorded whether they actually bought something on the website.

The results are summarized here

Appreciation	Buy(no)	Buy(yes)	Total
1	6	2	8
2	5	2	7
3	7	6	13
4	3	7	10
5	1	8	9
Total	22	25	47

For example, among 8 people with appreciation score 1, only two bought something. We introduce the variable Y to denote whether a customer has bought something ($Y = 1$) or not ($Y = 0$). The variable X is introduced to denote the appreciation value for a customer $X = 1, ..., 5$.

1.1 Exercise

Consider an arbitrary visitor of the website and assume that you do not have any information about how much this customer appreciates the website. How large is the probability that this customer has actually bought something?

1.2 Exercise

Consider the group of 8 visitors with a website appreciation score equal to 1

1. What is the probability that a customer from this group has actually bought some thing on this website?
2. What are the odds that a customer from this group has bought something against not having bought something?
3. What is the (natural) logarithm of the odds ratio for a customer of this group.
4. Repeat the calculations for each group of website visitors corresponding to the different levels of the appreciation variable

1.3 Exercise

Sketch the following plots

1. The probabilities (as calculated in Exercise 2) as a function of the value of the appreciation variable.
2. The (natural) logarithm of the odds, as a function of the value of the appreciation variable.

1.4 Exercise

In a logistic regression analysis the probability of a buy (Y=1), as a function of the appreciation value X is modeled by

$$P(Y = 1 \mid X = x) = \frac{exp(\theta_0 + \theta_1 x)}{1 + exp(\theta_0 + \theta_1 x)}$$

1. If we decide not to use the appreciation value X we set $\theta_1 = 0$. Suppose the maximum likelihood solution is $\theta_0 = 0.128$. Explain why this is indeed a good value.
2. Calculate the log likelihood of the data set, for the logistic model with $\theta_0 = 0.128$.

1.5 Exercise

We now fit the complete model and obtain $\theta_0 = -2.3190$ and $\theta_1 = 0.795$.

1. calculate for each value of X the estimated probabilities $P(Y = 1 \mid X = x)$
2. calculate the log likelihood of the data under this model.