Statistical Learning, Homework #1

Jacopo Manenti, id: 247279

2024-03-27

Introduction and Data Exploration

In this paper, we explore factors influencing pregnant women's decisions regarding breastfeeding their babies. Data from a study at a UK hospital involving 139 participants are stored in "bf.csv". We will conduct logistic regression and k-nearest neighbors analysis to understand factors affecting the decision of pregnant women to breastfeed their babies. This analysis might be useful in targeting breastfeeding promotions effectively.

```
df <- read.csv("bf.csv") # import data
dim(df) # get dimension
```

```
[1] 139  10
```

In Table 1 it has been reported a glimpse of the data set and in Table 2 the variables' data structure. The **breast** variable is the response variable, categorizing feeding practices into breastfeeding (Breast) and exclusive bottle feeding (Bottle). Since logistic regression model requires a Bernoulli variable, we will encode it into binary values (1 and 0, respectively). Furthermore, most variables are categorical, so to facilitate the modeling process, we will organize them into factors. Additionally, we look at missing values reported in Table 3, which mostly concern age and education level, notably among bottle-feeding and white mothers, indicating possible data collection issues. Thus, we remove these records, mindful of potential impacts on class balance. Finally, Table 4 displays the recoded and cleaned version of the data set, with fewer rows resulting from the removal of NA values (135). Following this data preparation step, we will perform basic statistical analyses and generate relevant plots.

Table 1: shows the first 6 rows in the data set.

breast	pregnancy	howfed	howfedfr	partner	smokenow	smokebf	age	educat	ethnic
Breast	Beginning	Breast	Breast	Partner	No	No	24	19	Non-white
Breast	Beginning	Bottle	Breast	Partner	No	No	27	18	White
Bottle	Beginning	Breast	Breast	Partner	No	No	39	16	White
Bottle	Beginning	Breast	Breast	Partner	Yes	Yes	29	16	White
Breast	Beginning	Breast	Breast	Partner	No	No	21	21	White
Bottle	Beginning	Breast	Bottle	Partner	No	No	NA	28	White

Table 2: shows the data type of each variable of the data set.

breast	pregnancy	howfed	howfedfr	partner	smokenow	smokebf	age	educat	ethnic
character	character	character	character	character	character	character	integer	integer	character

Table 3: shows the rows containing NA values.

breast	pregnancy	howfed	howfedfr	partner	smokenow	smokebf	age	educat	ethnic
Bottle	Beginning	Breast	Bottle	Partner	No	No	NA	28	White
Bottle	End	Breast	Bottle	Partner	No	No	31	NA	Non-white
Bottle	End	Bottle	Bottle	Partner	No	No	38	NA	White
Breast	End	Breast	Bottle	Partner	No	No	NA	16	White

Table 4: shows the cleaned version of the dataset with 135 rows (observations)

breast	pregnancy	howfed	howfedfr	partner	smokenow	smokebf	age	educat	ethnic
1	Beginning	Breast	Breast	Partner	No	No	24	19	Non-white
1	Beginning	Bottle	Breast	Partner	No	No	27	18	White
0	Beginning	Breast	Breast	Partner	No	No	39	16	White
0	Beginning	Breast	Breast	Partner	Yes	Yes	29	16	White

In Table 5 it has been reported the statistical summary of the dataset. It can be clearly seen that the *response variable* has a significant imbalance between classes, with 99 out of 135 instances reporting breastfeeding (gaphical representation at Figure 1). Also, among the 135 participants, only 21 are single, and a mere 32 are current smokers. The data set comprises patients aged between 17 and 40 years, having studed up to the age of 18 on average. Interviews were conducted balancedly across both races.

Table 5: summary statistics of response variable and demographic factors

(a) breast

breast
0:36
1:99

(b) partner

partner
Partner:114
Single : 21

(c) age

age
Min. :17.00
1st Qu.:25.00
Median :28.00
Mean :28.17
3rd Qu.:32.00
Max. :40.00

(d) smokenow

smokenow
No :103
Yes: 32

(e) educat

educat
Min. :14.00
1st Qu.:16.00
Median :17.00
Mean :18.09
3rd Qu.:19.00
Max. :38.00

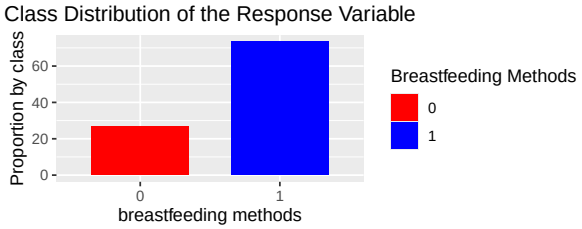

Figure 1: Illustrates breastfeeding (blue) vs. bottle-feeding (red) proportions.

We proceedexploring the relationship between numerical features (**age** and **educat**) and breastfeeding practices. Figure 2 (A) shows scatterplot, where breastfeeding individuals are in blue, non-breastfeeding in orange. Notably, a pattern emerges suggesting that the first groups tended to have a higher educational attainment compared to the other. This is reinforced by the boxplots, B for age and C for education, as *C* reports a revelant difference between the two conditionated classes (Bottle-feeding distribution is concentrated at lower education levels).

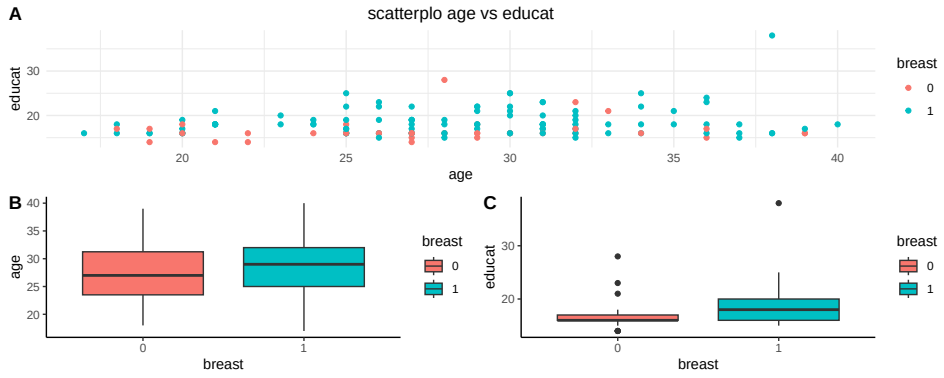

Figure 2: *Top*: Scatterplot (A) of mothers' age and education, red for breastfeeding, blue for non. *Center*: Boxplots of age (B) and education (C) by response status.

Table 6: Shows correlation matrix between age and educational attainment, revealing no relevant association.

	age	educat
age	1.0000000	0.2001478
educat	0.2001478	1.0000000

For categorical features, bar plots aid in similar analysis: consistent disparities in bar heights may suggest an association between predictor and response. In Figure 3, this behavior is seen in **howfdfr**, **smokenow**, **smokebf**, and **ethnic**. We confirm this with a chi-square test; Table 7 shows significance only for these predictors, indicating correlation with baby-feeding methodology.

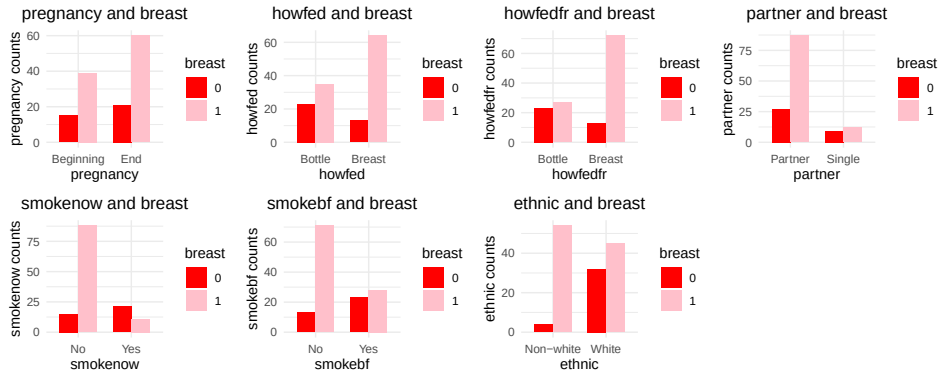

Figure 3: Bar plots for predictors and response (pink: breastfeeding, red: bottle-feeding).

Table 7: shows results of chisq-test between pairs of each categorical and breast variable

Variable	Chi_squared	P_value
pregnancy	0.0015783	0.968310
howfed	7.6465810	0.005688
howfedfr	13.6488971	0.000220
partner	2.4251324	0.119403
smokenow	29.9931994	0.000000
smokebf	12.7642165	0.000353
ethnic	18.5906330	0.000016

Splitting the data set into training and testing sets

Before proceeding with the classification task, we split the dataset into training (df_train; 100 samples) and testing sets (df_test; 35 samples), maintaining the observed class imbalance.

```
set.seed(5) # Set the seed for reproducible results
# get proportion of class
prop_breast<- round(prop.table(table(breast)) *100,2)
# Obtain indices
train_indices <- createDataPartition(breast,p=prop_breast[["1"]]/100, list = FALSE)
#create a train set and a vector of its labels
df_train <- df[train_indices,]
y_tr <-  breast[train_indices]
#create a test set and a vector of its labels
df_test  <- df[-train_indices, ]
y_ts <- breast[-train_indices]
```

Fit a Logistic Regression model and a k-NN

We can now start modelling a logistic regression (using the training set) for the purpose of estimating the probability of a mother breastfeeding her baby (coded as 1) given specific factor configurations, denoted as $P(breast = 1|factors)$. We estimate the values of the coefficients by the following approach:

$$\begin{aligned}\text{logit}(E(\text{breast})) = \beta_0 + \beta_1\text{pregnancy} + \beta_2\text{howfed} + \beta_3\text{howfedfr} + \beta_4\text{partner} \\ +\beta_5\text{age} + \beta_6\text{educat} + \beta_7\text{ethnic} \quad + \beta_8\text{smokenow} + \beta_9\text{smokebf}\end{aligned}$$

Table 8 presents the estimated coefficients and associated information of the fitted model. We conclude that, apart from **howfedfr**, **smokenow** and **ethnic**, there is no significant association between the predictors and the probability. The value of each coefficients means a variation in the probability of breastfeeding. For example, the coefficient of **smokenow** is negative, indicating that mothers that smoke are less likely breastfeed than non-smokers. To be more specific, holding other variables constant, smoking is associated with a decrease in the log odds of breastfeeding by 3.51100 units, and subsequently with a decrease in probability.

```
mod1 <- glm(breast~., data = df_train, family=binomial) # fit the model
mod1_sum <- summary(mod1) # retrieve associated information
```

Table 8: shows fitted logistic regression coefficients for predicting breastfeeding likelihood.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.4642895	3.0780088	-0.8006116	0.4233565
pregnancyEnd	0.8729899	0.6842648	1.2758070	0.2020237
howfedBreast	0.5368425	0.6604820	0.8128041	0.4163304
howfedfrBreast	1.6489413	0.6755578	2.4408588	0.0146524
partnerSingle	-0.9265348	0.7586623	-1.2212744	0.2219822
smokenowYes	-3.2118265	1.2439968	-2.5818608	0.0098269
smokebfYes	2.0509659	1.2159799	1.6866775	0.0916654
age	-0.0016593	0.0580734	-0.0285726	0.9772055
educat	0.1837402	0.1445962	1.2707126	0.2038309
ethnicWhite	-1.6617704	0.7746736	-2.1451232	0.0319430

After fitting the logistic regression, we proceed to the k-nearest neighbors (k-NN) classifier. To determine the optimal value for k, we compare values from 1 to 100 to find the one with the lowest error rate, indicating the fraction of misclassified test observations. Choosing the value of k that minimizes this error helps obtain a more accurate model that generalizes more effectively. From Figure 4, the optimal value is $k = 9$.

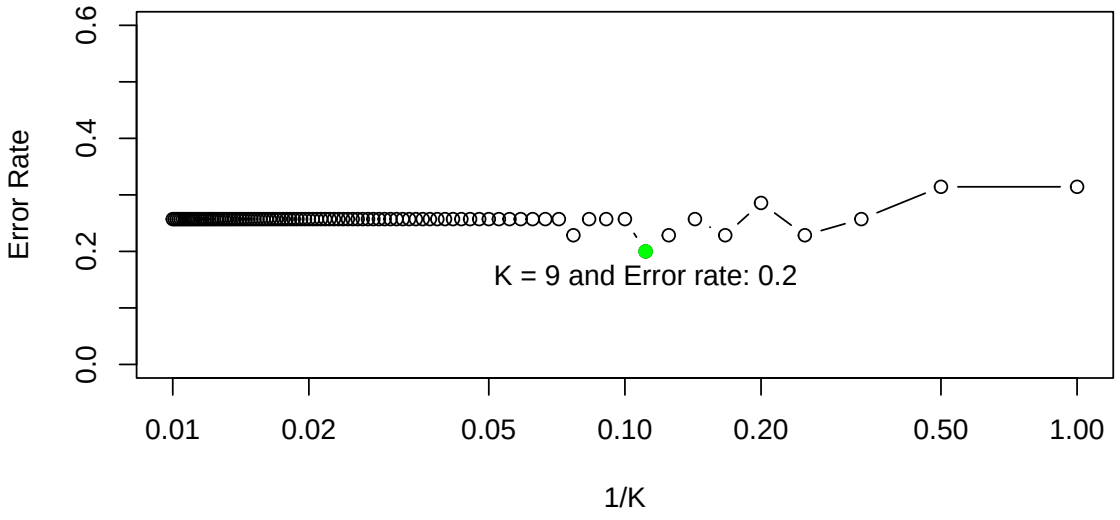

Figure 4: The KNN test error rate as the number of neighbors K decreases.

Predictions with the two models

We are now ready to take predictions on test set using both models. For the *k-NN*: given an observation, we will assign the majority class among k=9 neighbors. For *logistic regression*: given an observation, we will compute probabilities using the estimated coefficients and compare them with the threshold value of 0.5. If the estimated probability exceeds 0.5, the mother's breastfeeding behavior will be predicted as 1 (breastfeeding); otherwise, it will be classified as 0 (bottle).

```
#vector of predictions with k-nn where k = 9
knn_pred <- knn(x_train, x_test, y_tr, k = 9)
#vector of predictions with fitted logistic regression
mod1_predict <- predict(mod1, df_test, type= "response")
#classify using our 0.5 threshold:
glm_pred <- rep(0, nrow(df_test))
glm_pred[mod1_predict > 0.5] <- 1
```

Next, to assess performance evaluation, we utilize the confusion matrix and common evaluation metrics including **Accuracy** (proportion of correctly classified cases among all cases), **sensitivity** (proportion of true positives among all actual positive cases), **specificity** (proportion of true negatives among all actual negative cases) and **false positive rate** (proportion of negative cases mistakenly classified as positive among all actual negative cases).

Performance evaluation, conclusion and limitation

Looking at Table 9 (a) and (b), both algorithms show similar numbers of correctly classified cases (sum of elements in the main diag), implying similar accuracy levels. However, Table 9 (c) reveals some differences: **k-NN model** achieves perfect sensitivity (1), indicating precise identification of all positive cases. However, its specificity is notably low at 0.22 (high false positive rate). This suggests difficulties in accurately classifying negative cases, as it tends to classify most instances into the majority class (positive). Conversely, despite having lower sensitivity (0.84), **Logistic regression model** outperforms k-NN in specificity with a score of 0.88, resulting in a markedly lower false positive rate (0.11) and demonstrating a superior classification ability.

This observation is crucial for our study's objective of optimizing the targeting of breastfeeding promotions toward less inclined women. For instance, let's consider estimating the probability of a white, non-single, smoker mother (but not before), beginning her pregnancy, with friends who do not breastfeed, aged 24, and who discontinued schooling at 18, opting to breastfeed her baby despite not having been breastfed herself. We can make the predictions by using estimates for the regression coefficients from Table 8 as follows:

Table 9: Performance evaluation

(a) k-NN (k=9) predictions compared to true breastfeeding behavior for 35 test set observations.

	true 0	true 1
pred 0	1	0
pred 1	8	26

(b) Confusion matrix for logistic regression (threshold = 0.5) compared to true breastfeeding behavior for 35 test set observations.

	true 0	true 1
pred 0	8	4
pred 1	1	22

(c) Accurancy, sensitivity, specificity, false positive rate for both models

	k-nn	log.model
acc	0.7714286	0.8571429
sens	1.0000000	0.8461538
spec	0.1111111	0.8888889
fpr	0.8888889	0.1111111

```
#probability estimation of the observation
predict_example <- predict(mod1, newdata = list(pregnancy="Beginning",
                                                howfed = "Bottle",
                                                howfedfr = "Bottle",
                                                partner="Partner",
                                                smokenow="Yes",
                                                smokebf="No",
                                                age = 24,
                                                educat = 18,
                                                ethnic="White" ),
                           type = "response")
```

$$\hat{p}(1|X) = \frac{e^{-2.464-0.002(\text{age=24})+0.183(\text{educat=18})-1.662(\text{ethnic=}white)-3.212(\text{smokenow=}Yes)}}{1+e^{-2.464-0.002(\text{age=24})+0.183(\text{educat=18})-1.662(\text{ethnic=}white)-3.212(\text{smokenow=}Yes)}} = 0.0167843$$

A potential limitation of this conclusion, and so of the study, lies in the sampling method utilized in section 2, impacting K-NN's performance due to dataset imbalance. However, techniques such as class weighting, batching, or resampling could alleviate this issue, enhancing overall model performance on imbalanced datasets.