Chair of Decision Sciences and Systems
TUM School of Computation, Information and Technology
Technical University of Munich

Business Analytics & Machine Learning
Homework sheet 3: Logistic Regression

Prof. Dr. Martin Bichler, Prof. Dr. Jalal Etesami
Julius Durmann, Markus Ewert, Johannes Knörr, Yutong Chao

Exercise H3.1 *Master's course applications*

In this exercise, you will fit a logistic regression model for predicting admittance to Master's courses. You will assess its performance based on various metrics.
Note: Use the provided notebook as a template. It implements some portions of the code so you don't need to start from scratch. Be aware that the code is based on statsmodels.

The data (*admit-train.csv* and *admit-test.csv*) consists of values for the variables *admit*, *gre*, *gpa* and *rank*. The attribute *admit* indicates whether a student has been admitted to a Master's course. The attributes *gre* and *gpa* contain the results of certain exams. The attribute *rank* represents the reputation rank of the student's current university. The smaller the rank, the higher is the university's reputation.

Load the data and visualize it by using the following code (already in the template):

```
import pandas as pd
# Read data into dataframe
df_train = pd.read_csv("admit-train.csv")
# Show data
print(df_train.head())
# Pairplot
pd.plotting.scatter_matrix(df_train)
```

a) Briefly describe the data set:
   1) Name the dependent variable and the independent variables.
   2) Which scales of measurement do the variables belong to (e.g., nominal, ordinal, interval or ratio)?

Due to the fact that the dependent attribute *admit* is binary, we use a logistic regression model.
Use the code snippet below (already in the template) to create a logit-model from the training data and to obtain the results.

```
import statsmodels.formula.api as smf
# Mark "rank" as categorical
df_train["rank"] = df_train["rank"].astype("category")
# Define and fit a model
logreg = smf.logit("admit ~ gre + gpa + rank", data=df_train)
result = logreg.fit()
print(result.summary())
```

b) Which attributes are statistically significant regarding a significance level of 5%?

c) Interpret the coefficients.

d) The *rank* attribute is split into multiple sub-variables (`rank[T.2]`, `rank[T.3]`, `rank[T.4]`) by our model. They indicate whether the rank has a certain value:

$$rank[T.i] = \mathbb{I}[rank = i], \qquad i \in \{2, 3, 4\}$$

For $rank = 1$, we don't need another variable since this case can be inferred if all other rank-variables are zero.
To test the significance of the attribute *rank*, you thus need to test if the null-hypothesis

$$H_0 : \quad \beta_{rank=2} = 0 \quad \wedge \quad \beta_{rank=3} = 0 \quad \wedge \quad \beta_{rank=4} = 0$$

can be rejected.
You can do that by a Wald test which allows to test combined hypotheses as the one above:

```
wald_test_result = result.wald_test(
    "(rank[T.2] = 0, rank[T.3] = 0, rank[T.4] = 0)",
    scalar=True
)
```

Is the attribute *rank* statistically significant w.r.t. a level of $\alpha = 5\%$?

e) In order to gain a better understanding of the model, have a look at the predicted probabilities of some observations. Adjust only one parameter and keep the others constant. For example keep *gre* and *gpa* constant (using their mean/average) and vary *rank*. Can you draw any conclusions?

f) Find the McFadden ratio and interpret the results.

g) Load the test data from *admit-test.csv* and predict the probability of its entries.
Construct the confusion matrix for the test data.

h) Compute the accuracy

$$\text{acc} = \frac{TP + TN}{TP + TN + FP + FN}$$

of your trained model on the test set.

Exercise H3.2 *Car ownership by age*

You want to examine the relationship between age and owning a car. Owning a car is modeled as a binary variable, taking on the value of one when true and zero if not. Therefore, you employed a logistic regression and obtained the following results:

Variable	Est. coefficient	Standard error
Age	0.135	0.036
Constant	-3.89	1.73

a) According to the model above, what effect (qualitative and quantitative) does a change in age (+1) have on the dependent variable?

b) Find the age for which the model would be indifferent between owning a car and not owning one. You can either use an exact formula or rely on an approximate graphical solution.