Tut 4: Logistic Regression (cont)

Jan 2022

1. (May 2020 Final Q2(a)) The testing dataset of an insurance claim is given in Table 2.1. The variables "gender", "bmi", "age_bracket" and "previous_claim" are the predictors and the "claim" is the response.

Table 2.1: The testing data of an insurance claim (randomly sampled with repeated entry).

gender	bmi	age_bracket	previous_claim	claim
female	under_weight	18-30	0	no_claim
female	under_weight	18-30	0	no_claim
male	over_weight	31-50	0	no_claim
female	under_weight	50+	1	no_claim
male	normal_weight	18-30	0	no_claim
female	under_weight	18-30	1	no_claim
male	over_weight	18-30	1	no_claim
male	over_weight	50+	1	claim
female	normal_weight	18-30	0	no_claim
female	obese	50+	0	claim

The "gender" is binary categorical data, the "bmi" is a four-value categorical data with values under_weight, normal_weight, over_weight and obese, the "age_bracket" is a three-value categorical data with value "18-30", "31-50" and "50+", the "previous_claim" is a binary categorical data with 0 indicating "no previous claim" and 1 indicating "having a previous claim". The "claim" is a binary response with values "no_claim" (negative class, with value 1) and "claim" (positive class, with value 0).

Suppose a logistic regression model is trained and the coefficients are stated in Figure 2.2.

Figure 2.2: The coefficients of the logistic regression based on an insurance claim data.

```
Call:
glm(formula=Purchased~., family=binomial, data=data.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9882  -0.5640  -0.1372   0.5532   2.1820

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.188e+01  2.497e+00  -4.757 1.96e-06 ***
GenderMale       4.221e-01  5.927e-01   0.712 0.476319
Age              2.178e-01  4.751e-02   4.584 4.56e-06 ***
EstimatedSalary  3.868e-05  1.001e-05   3.863 0.000112 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 135.37  on 99  degrees of freedom
Residual deviance:  74.91  on 96  degrees of freedom
```

Write down the **mathematical formula** of the logistic regression model and then use it to **predict** the “claim” of the insurance data in Table 2.1 as well as **evaluating** the performance of the model by calculating the confusion matrix, accuracy, sensitivity, specificity, PPV, NPV of the logistic model. [**Note**: The default cut-off is 0.5] (4 marks)

2. (Jan 2021 Final Q2(b)) The testing dataset of a social network advertisement is given in Table 2.2. The variables "Gender", "Age" and "EstimatedSalary" are the predictors and the variable "Purchased" is the response. The "Gender" is a binary categorical data with levels "Male" and "Female", the "Age" and the "EstimatedSalary" are quantitative data. The "Purchased" is a binary response with values 0 (representing "no purchase", assuming **0 is the positive class**) and 1 (representing "purchase").

Table 2.2: The testing data of a social network advertisement.

Gender	Age	EstimatedSalary	Purchased
Male	29	80000	0
Male	45	26000	1
Female	48	29000	1
Male	45	22000	1
Female	47	49000	1
Male	48	41000	1
Male	46	23000	1
Male	47	20000	1
Male	49	28000	1
Female	47	30000	1

Figure 2.1: The coefficients of the logistic regression based on an insurance claim data.

```
Call:
glm(formula=Purchased~., family=binomial, data=data.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9882  -0.5640  -0.1372   0.5532   2.1820

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.188e+01  2.497e+00  -4.757 1.96e-06 ***
GenderMale       4.221e-01  5.927e-01   0.712 0.476319
Age              2.178e-01  4.751e-02   4.584 4.56e-06 ***
EstimatedSalary  3.868e-05  1.001e-05   3.863 0.000112 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 135.37  on 99  degrees of freedom
Residual deviance:  74.91  on 96  degrees of freedom
```

Suppose a logistic regression model is trained and the coefficients are stated in Figure 2.1. Write down the **mathematical formula** of the logistic regression model and then use it to **predict** the variable "Purchase" of the insurance data in Table 2.2 as well as **evaluating** the performance of the model by calculating the confusion matrix, accuracy, sensitivity, specificity, PPV, NPV of the logistic model (assuming 0 is the positive class). [**Note**: The default cut-off is 0.5] (5 marks)