Stat 427/627 Statistical Machine Learning Quiz 3 Jun Lu. American University

Quiz 3. Classification I: KNN, Intro to Logistic Regression

Name: Lisa Liubovich Attempt (circle one): BEFORE AFTER

1. A classification model yields the following confusion matrix on the entire data set.

|  |  |  |
| --- | --- | --- |
|  | Predicted Positive (1) | Predicted Negative (0) |
| Actual (i.e., Observed) Positive (1) | 28 | 7 |
| Actual (i.e., Observed) Negative (0) | 14 | 21 |

Calculate the following metrics and show your work. For example:

Classification Rate (Accuracy) = (𝑇𝑃 + 𝑇𝑁 )/𝑁 = (28 + 21)/(28 + 21 + 14 + 7) = 49/70 = 0.7

* + 1. True Positive Rate (TPR)

TPR = TP/P = 28/(28+7) = 28/35 = 0.80

* + 1. False Positive Rate (FPR)

FPR = FP/N = 14/(14 + 21) = 14/35 = 0.40

1. An insurance company tries to predict whether a new customer has an accident during the next 3 years. This may depend on the driving experience. They use the following data for 9 randomly chosen customers. (Data are sorted by driving experience)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Experience (years) | 0 | 3 | 3 | 4 | 8 | 10 | 13 | 17 | 22 |
| Accident in 3 years | N | N | Y | Y | N | Y | N | N | N |

Consider a new customer with 10 years of experience.

(a) Use KNN algorithm to predict whether the new customer would have an accident in 3 years. Consider K = 1 and 3, respectively.

* + - 1. K = 1

Distance of each existing customer from new customer based on experience:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Experience (years) | 0 | 3 | 3 | 4 | 8 | 10 | 13 | 17 | 22 |
| Distance | 10-0 = 10 | 10-3 = 7 | 10-3 = 7 | 10-4 = 6 | 10-8 = 2 | 10-10 = 0 | |10-13| = 3 | |10-17| = 7 | |10 -22| = 12 |

Nearest neighbor/smallest distance = customer with 10 years of experience (0) 🡪 the accident status of the nearest neighbor with 10 years of experience is Y

Therefore, for k = 1, the prediction is that the new customer will have an accident.

* + - 1. K = 3

The three smallest distances are 0 (experience = 10), 2 (experience = 8), and 3 (experience = 13). The accident statuses for these experiences are Y (experience = 10), N (experience = 9), and N (experience = 13). 🡪 N is the majority class

Therefore, for k = 3, the prediction is that the new customer will not have an accident.

(b) A logistic regression model is also fit on the data. Use the following output to predict the *probability* that the new customer (10 years of experience) would have an accident. (You may keep your result in the form of 𝑒𝑎.)

Call: glm(formula = accident ~ years, family = binomial(link=logit)) Coefficients:

(Intercept) years

0.261 -0.122

Pr(accident = Y|experience = 10) = (e0.261 + (-0.122)(10))/(1 + e0.261 + (-0.122)(10)) = (e0.261 – 1.22)/(1 + e0.261 -1.22) = (e-0.959)/(1 + e -0.959) = 0.38327597037/1.38327597037 = 0.2770784562

The Probability that a new customer with 10 years of experience would have an accident is 1/e0.959, which is approximately 0.2771 or 27.71%.

Stat 427/627 Statistical Machine Learning Quiz 3 Jun Lu. American University

~~Other questions I removed from this quiz: distance of higher dimensional data. Flexibility of KNN.~~