Homework 2 Report Problem

陳俊翰
EE5184 - Machine Learning

November 1, 2018

Problem 1. (1%) 請簡單描述你實作之logistic regression 以及generative model 於此task 的表現，並試著討論可能原因。

Answer 1. From the result, we can see the performance of logistic regression is better than generative model, which is Naive Bayes in this experiment.
The reasons are as below.
Take column "Education" for example.(1 = graduate school; 2 = university; 3 = high school; 4 = others) First, compare logistic regression and Naive Bayes:[1]
1. When the training size reaches infinity, the generative model: logistic regression performs better than the generative model Naive Bayes.
2. The generative model (Naive Bayes) reaches the asymptotic solution for fewer training sets than the generative model (Logistic Regression)
3. Naive Bayes also assumes that the features are conditionally independent.
4. Naive Bayes has a higher bias but lower variance compared to logistic regression.

In my opinion, it's because we have 20,000 training data which is enough for logistic regression to win prediction.On the other hand, Naive Bayes has a higher bias and our data doesn't follow the bias.

Algorithm\Accuracy	Training set	Public leaderboard
Logistic Regression	0.82	0.82120
Generative Model	0.79565	0.808

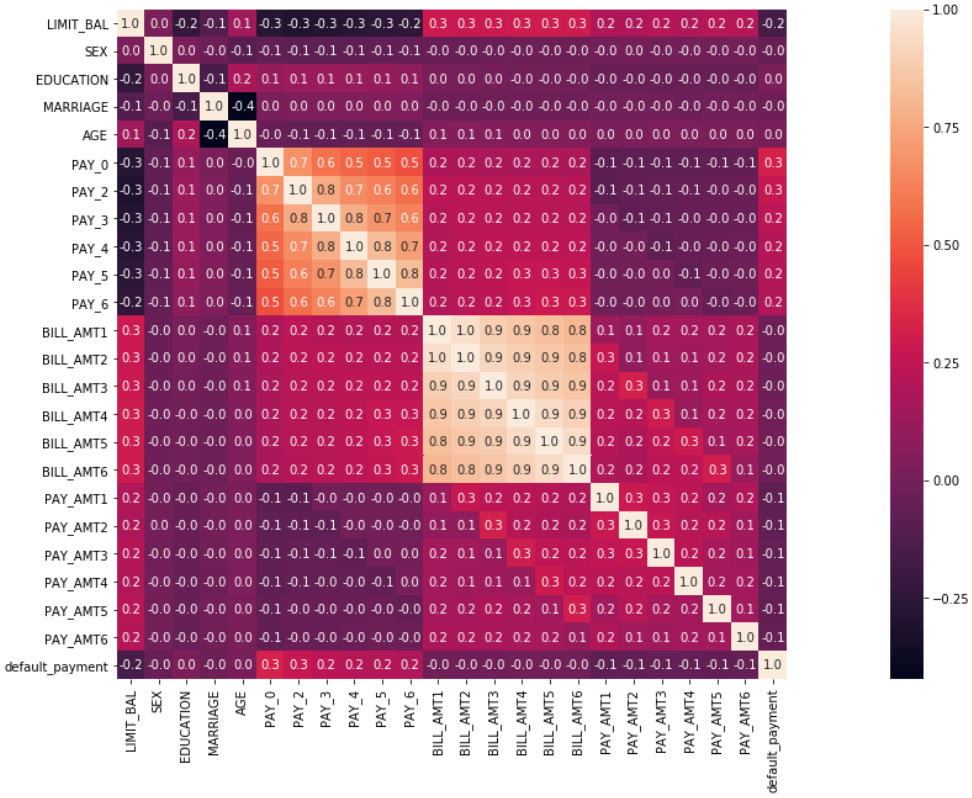
Problem 2. (1%) 請試著將input feature 中的gender, education, martial status 等改為one-hot encoding 進行training process，比較其模型準確率及其可能影響原因。

Answer 2. From the result, we can see that One-Hot Encoding will perform better.Take the column "Education" for example. (1 = graduate school; 2 = university; 3 = high school; 4 = others)We can see the "Education" data is discrete. Without encoding, the model will think 1 is more relative to 2 than 3 because the distance is shorter. However, 1,2,3,4 are categories and it means that they should have the same correlation and distance. Therefore,

we use One-Hot Encoding to make all values of the categorical features are equally away from each other.

Algorithm\Accuracy	Training set	Public leaderboard
Logistic Regression with OneHotEncoding	0.82	0.82120
Logistic Regression w/o OneHotEncoding	0.81185	0.81160

Problem 3. (1%) 請試著討論哪些input features 的影響較大（實驗方法沒有特別限制，但 請簡單闡述實驗方法）。



Answer 3. [2]I use the python package "seaborn" to visualize the data and draw the correlation matrix and heatmap.
From above, we can find that the PAY_0, PAY_X are the strongest predictors of default, followed by the LIMIT_BAL variable.

Problem 4. (1%) 請實作特徵標準化(feature normalization)，並討論其對於模型準確率的影響與可能原因。

Answer 4. From the result, we can see that with feature scaling, the accuracy is much higher and the time is much shorter. [3]Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, the majority of classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.

Another reason why feature scaling is applied is that gradient descent converges much faster with feature scaling than without it.

Algorithm (epoch=1000, learning rate=0.01)	Training set (Accuracy)	Time(s)
Logistic Regression with Feature Scaling	0.8039	3.73
Logistic Regression w/o Feature Scaling	0.7802	34.62

Problem 5. (1%) The Normal (or Gaussian) Distribution is a very common continuous probability distribution. Given the PDF of such distribution

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty$$

please show that such integral over $(-\infty, \infty)$ is equal to 1.

Answer 5.

$$I = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

$$
\begin{aligned}
I^2 &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma^2} e^{-\frac{(x-\mu)^2(y-\mu)^2}{2\sigma^2}} d(x-\mu) d(y-\mu) \\
&= \int_{0}^{2\pi} \int_{0}^{\infty} \frac{1}{2\pi\sigma^2} e^{-\frac{r^2}{2\sigma^2}} r dr d\theta \\
&= \int_{0}^{2\pi} d\theta \frac{1}{2\pi\sigma^2} (-\sigma^2 e^{-\frac{r^2}{2\sigma^2}})|_{r=0}^{\infty} \\
&= \frac{-2\pi\sigma^2}{2\pi\sigma^2} (e^{-\infty} - e^{0}) \\
&= 1
\end{aligned}
$$

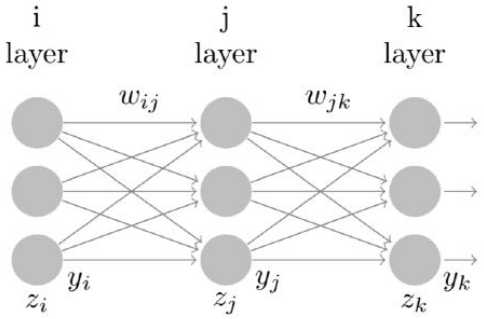
Thus,

$$I = 1$$

, where r and θ are polar coordinate and the origin is (μ, μ), and integral over $r = 0 \to \infty, \theta = 0 \to 2\pi$.

QED

Problem 6. (1%) Given a three layers neural network, each layer labeled by its respective index variable. I.e. the letter of the index indicates which layer the symbol corresponds to. For convenience, we may consider only one training example and ignore the bias



term.Forward propagation of the input z_i is done as follows.Where g(z) is some differentiable function (e.g. the logistic function).

$$y_i = g(z_i)$$

$$z_j = \Sigma_i w_{ij} y_i$$

$$y_j = g(z_j)$$

$$z_k = \sum_j w_{jk} y_j$$

$$y_k = g(z_k)$$

Derive the general expressions for the following partial derivatives of an error function E, also sime differentiable function, in the feed-forward neural network depicted. In other words, you should derive these partial derivatives into "computable derivative"

$$(a)\frac{\partial E}{\partial z_k}(b)\frac{\partial E}{\partial z_j}(c)\frac{\partial E}{\partial w_{ij}}$$

Answer 6. (a)

$$\frac{\partial E}{\partial z_k} = \frac{\partial y_k}{\partial z_k}\frac{\partial E}{\partial y_k}$$

(b)

$$\begin{aligned}\frac{\partial E}{\partial z_j} &= \frac{\partial E}{\partial z_k}\frac{\partial z_k}{\partial z_j} \\ &= \frac{\partial E}{\partial z_k}\frac{\partial z_k}{\partial y_j}\frac{\partial y_j}{\partial z_j} \\ &= \frac{\partial E}{\partial z_k}(\sum_j w_{jk})g^{'}(z_j) \\ &= g^{'}(z_j)\frac{\partial y_k}{\partial z_k}\frac{\partial E}{\partial y_k}\sum_j w_{jk}\end{aligned}$$

(c)

$$\begin{aligned}\frac{\partial E}{\partial w_{ij}} &= \frac{\partial E}{\partial z_j}\frac{\partial z_j}{\partial w_{ij}} \\ &= g^{'}(z_j)\frac{\partial y_k}{\partial z_k}\frac{\partial E}{\partial y_k}(\sum_j w_{jk})\frac{\partial z_j}{\partial w_{ij}}\end{aligned}$$

References

[1] Ng, Andrew Y., and Michael I. Jordan. "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes." Advances in neural information processing systems. 2002.

[2] https://www.kaggle.com/ainslie/credit-card-default-prediction-analysis/notebook

[3] https://en.wikipedia.org/wiki/Feature_scaling