CS4375.003

Simon Kim

Sxk190106

ML from Scratch

1. Two paragraphs comparing and contrasting generative classifiers versus discriminative classifiers. Cite any sources you use.

Generative classifier, such as naïve bayes, tries to model class in how a particular class would generate input data [1]. We call this "learning about the environment". Discriminative classifier, like logistic regression learn what features of the model is useful to distinguish each data. Mathematically, discriminative classifier direcly calculates the posterior probability P(y|x) or learn a direct map from input x to label y, and generative classifier tries to learn the joint probability distribution p(x,y) of the inputs x and label y and make their prediction using Bayes rule to calculate the conditional probability, p(y|x). [1]

These two classifiers seems different, but they are similar because both methods use conditional probability to classify. However, the discriminative classifier is more accurate as it tries to directly solve the task, rather than tying to solve a general problem as an intermidiate step as generative models do[1]. This is why accuracy of Logistic regression is better than naïve Bayesian when we implemented them through R.

2. Google this phrase: reproducible research in machine learning. Using 2-3 sources, at least one of which should be academic, write a couple of paragraphs of what this means, why it is important, and how reproducibility can be implemented. Cite your sources using any format.

Reproducibility is a minimal prerequisite for the creation of new knowledge and scientific progress [2]. Through reproducibility in the history of mankind, we have verified existing new technologies and created new technologies. Even in the research of a new era in which machine learning is important or essential, the possibility of verification will be very important.

Reproducibility in machine learning means that you can repeatedly run your algorithm on certain datasets and obtain the same (or similar) results on a particular project. Reproducibility in machine learning means being able to replicate the ML orchestration carried out in a paper, article, or tutorial and getting the same or similar results as the original work. [3]

However, machine learning has a very negative shape for reproducibility.The training of any machine learning models makes use of randomness, and this is especially true for deep learning models, which are trained by a process known as static gradient design [2] It cannot be easily verified by randomness.

As a way to solve this problem, a method of utilizing a shared DevOps platform that guarantees reproducibility, the establishment of a standard neural network is being devised, and it is recommended to leave tracks that others can independently verify.

3. copy/paste runs of your code showing the output

(1) Logistic Regression

```
Opening file titanic_project.csv.
Reading line 1
heading: "","pclass","survived","sex","age"
new length 1046
Closing file titanic_project.csv.
Number of records: 1046
Stats for pclass
Sum: 2309
Mean: 2.20746
Median: 2
Range: 2

Stats for survived
Sum: 427
Mean: 0.408222
Median: 0
Range: 1

Stats for sex
Sum: 658
Mean: 0.629063
Median: 1
Range: 1

Stats for age
Sum: 31231
Mean: 29.8576
Median: 28
Range: 80

Stats for pclass_train
Sum: 1780
Mean: 2.225
Median: 2
Range: 2

Stats for survived_train
Sum: 312
Mean: 0.39
Median: 0
Range: 1

Stats for sex_train
Sum: 510
Mean: 0.6375
Median: 1
Range: 1

Stats for age_train
Sum: 23819
Median: 1
Range: 1

Stats for age_train
Sum: 23819
Mean: 29.7737
Median: 28
Range: 80

Stats for pclass_test

Sum: 523
Mean: 2.14344
Median: 2
Range: 2

Stats for survived_test

Sum: 115
Mean: 0.471311
Median: 0
Range: 1

Stats for sex_test

Sum: 147
Mean: 0.602459
Median: 1
Range: 1

Stats for age_test

Sum: 7360
Mean: 30.1639
Median: 29
Range: 76

Logistic Regression
Coefficients :
0: 159.97
1: -23.7194
2: -149.28
3: -8.71368
Accuracy: 1
Sensitivity: 1
Specificity: 1
Elapsed time in milliseconds: 55887 ms

Program terminated.
C:₩Users₩user₩source₩repos₩Log_Reg₩Debug₩Log
```

(2) Naïve Bayes

```
Opening file titanic_project.csv.
Reading line 1
heading: "","pclass","survived","sex","age"
new length 1046
Closing file titanic_project.csv.
Number of records: 1046
Stats for pclass
Sum: 2309
Mean: 2.20746
Median: 2
Range: 2

Stats for survived
Sum: 427
Mean: 0.408222
Median: 0
Range: 1

Stats for sex
Sum: 658
Mean: 0.629063
Median: 1
Range: 1

Stats for age
Sum: 31231
Mean: 29.8576
Median: 28
Range: 80

Split data into train-test
Stats for pclass_train
Sum: 1780
Mean: 2.225
Median: 2
Range: 2

Stats for survived_train
Sum: 312
Mean: 0.39
Median: 0
Range: 1

Stats for sex_train
Sum: 510
Mean: 0.6375
Median: 1
Range: 1

Stats for age_train
```

```
Stats for age_train
Sum: 23819
Mean: 29.7737
Median: 28
Range: 80

Stats for pclass_test
Sum: 523
Mean: 2.14344
Median: 2
Range: 2

Stats for survived_test
Sum: 115
Mean: 0.471311
Median: 0
Range: 1

Stats for sex_test
Sum: 147
Mean: 0.602459
Median: 1
Range: 1

Stats for age_test
Sum: 7360
Mean: 30.1639
Median: 29
Range: 76

Naive-Bayesian
Prior probability, survived = no, survived = yes.
0.610000 0.390000

Likelihood for p(pclass|survived):
sex | class (1:2:3)
0     0.172131 0.225410 0.602459
1     0.416667 0.262821 0.320513

Likelihood for p(sex|survived):
sex | survived:not
0     0.159836 0.840164
1     0.679487 0.320513

Accuracy: 0.000000
Sensitivity: 0.000000
Specificity: 1.000000

Applied to the first 5 test observations:
```

```
Naive-Bayesian
Prior probability, survived = no, survived = yes.
0.610000 0.390000

Likelihood for p(pclass|survived):
sex | class (1:2:3)
0    0.172131 0.225410 0.602459
1    0.416667 0.262821 0.320513

Likelihood for p(sex|survived):
sex | survived:not
0    0.159836 0.840164
1    0.679487 0.320513

Accuracy: 0.000000
Sensitivity: 0.000000
Specificity: 1.000000

Applied to the first 5 test observations:
0.701591 0.298409
0.540116 0.459884
-0.000000 1.000000
-0.000000 1.000000
-0.000000 1.000000
Elapsed time in milliseconds: 11 ms

Program terminated.
C:₩Users₩user₩source₩repos₩nay_bay₩Debug₩nay_bay.exe(
```

4. analyze the results of your algorithms on the Titanic data

Naïve Bayesian program was faster than Logical regression program. This is because the calculation for Naïve Bayesian is easier than Logical regression. Considering that I used a little bit old machine to run both program, 11 ms is remarkable result, comparing to 55sec of Logistic regression algorithm.

I think there was a mistake on cal_acc function on Naïve Bayes code. Because, when I applied raw probability on test data, I could see reasonable result, but the accuracy turned out to 0, while Logistic regression showed 1. Also, the accuracy of 1 seems nice, but since train-predict cannot actually hit every try, accuracy is hard to be 1. Therefore, my code to calculate accuracy and sensitivity had a problem to analyze full data.

But, I am confident to say that algorithm worked fine because likely hood value survived, and logreg code calculated coefficient value well.

From the learning, we can say: if the passenger is female, she might be able to survive better than male passenger. Also, for same male passenger, if he has better class ticket, he might not survive from the accident.

Works cited

[1] https://www.linkedin.com/pulse/generative-classifiers-vs-discriminative-akanksha-malhotra

[2] Beam, Andrew L et al. "Challenges to the Reproducibility of Machine Learning Models in Health Care." JAMA vol. 323,4 (2020): 305-306. doi:10.1001/jama.2019.20866

[3] https://neptune.ai/blog/how-to-solve-reproducibility-in-ml