COMP551 Mini Project 1: Machine Learning 101

Matteo Cacciola **Si Yi Li** **Weiqi (Agnes) Liu**

Abstract

Logistic Regression (LR) is believed to perform well under problem settings with large feature space, while Naive Bayes (NB) is more competitive on smaller datasets. Here, we first illustrated the significant impact of learning rate choice on the performances of LR: too high values of the learning rate leads to instability, while too low of it slowers the convergence speed. We then testified the performances of our implementation of LR vs NB over 4 distinct datasets, in which LR has shown an overall higher accuracy than NB. When tested on the same dataset with tuning the size of training set, NB performed better in smaller datasets, whereas LR becomes more advantageous as dataset enlarges. All our code are avaible in our Google Drive.

1 Introduction

Logistic Regression is a discriminative linear classifier with a binomial response variable. A discriminative training process directly maximizes the likelihood function defined on the posterior $p(c|x)$, whereas generative classifiers, such as Naive Bayes, model the joint probability $p(c, x)$ first, then calculate the posterior based on Bayes Rules to make prediction. LR is known for its advantage of the linear dependence of number of parameters over the feature space dimension, thus is usually preferred under large feature-space settings (Ng and Jordan, 2002; Bishop, 2006).

In this project, we implemented models for each of Naive Bayes (NB) and Logistic Regression (LR), then compared their performances from different approaches. We cross compared their accuracy over 4 different datasets, investigated the influences of different learning rates and iterations on LR performance, and compared the changes in accuracy of the models as dataset size increases.

We noticed that both models performed well in all datasets, with LR dominating over NB in terms of accuracy. However, NB outran LR when dataset shrinks. As for influence of LR performance brought by altering the learning rate, our results showed that, if the applied learning rate is too high, the model tends to be suboptimal and cannot guarantee on prediction accuracy. Nevertheless, if the learning rate is overly small, it might severely increase the time complexity of the execution as a result of the suppressed convergence rate.

Apart from comparing the performance between different models, it is also important to know how each model predicts the outputs. Concretely speaking, it is meaningful to investigate important features for LR or NB during prediction. In this project we applies two methods to select the important features of the two models. The first approach is called Drop Column Feature Importance, and it is intuitive as we drop a column of feature at a time, retrain the model again and compare the change accuracy to model trained with the whole dataset. The second approach is called Saliency Map which is used to study computer vision (Simonyan et al., 2013). It uses the knowledge of gradient to study the features, as derivative can be understood as how sensitive the output is with respect to the change a features. In other words, larger derivative indicates a larger impact on the loss function. Following this, the gradient of a feature, during prediction, can be seen as an estimate of its contribution to current prediction.

The 4 datasets we chose to focus on our project are from the UCI machine learning repository, namely the ionosphere dataset, the adult ("Census Income") dataset, the Wine dataset, and the Breast Cancer dataset. The Wine dataset contains results of a chemical analysis on 13 constituents of wines brewed from three different cultivars. The Breast Cancer dataset has clinical information from 10 aspects for patients with this cancer. Prior to training the models, all datasets have been preprocessed to avoid errors caused by invalid datapoints.

2 Datasets

Instances with missing/malformed values were removed from all 4 datasets, and all response variables were binarized. All numerical values x_i were normalized by ($\frac{x_i - \bar{x}}{\sigma}$). In particular, categorical features in the Adult dataset were converted to discrete integers for simplicity, and were performed with One-hot encoding for LR.

2.1 Ionosphere Dataset

This dataset has 34 continuous features, with "Bad"-labeled instances less than $\frac{2}{3}$ of the "Good"s. The features could be divided to "imaginary" and "real" pulses, with means of 0 and 1 respectively. Our plots of label distributions over different features showed that the "Bad" samples are more deviated from the mean, with peaks in every feature centered at ± 1. Amongst the features, "pulse 0 (imaginary)" has all values equal to 0, thus was removed for training.

2.2 Adult Dataset

This dataset of 14 features shows whether the annual income of the sample exceeds $\$50000$. 8 features are multi-class categorical and 6 are numerical.This dataset has a high level of imbalance, with "low salary" ($\leq \$50,000$) instances 3 times more than the "high salary"s ($\geq \$50000$).The 2 classes have varying ratios across features, but the "low" almost always dominates the "high". The only exceptions fell into "self-emp-inc" in the workclass feature and "HS-grad" in the education feature. Notably, the two classes have highly different distributions in "age" and "hour per week", indicating importance of these features in binary prediction.

2.3 Wine Dataset

This dataset has 13 continuous features as chemical analyses used for wine origin identification. We chose a subset of samples with only 2 origins to binarize the response variable. In this dataset, samples in the 2 classes is relatively balanced. However, the 2 classes are distinctively distributed throughout features. For example,"Alcohol" and "Proline" showed highly separated centres for the 2 classes. Such clear distinction might be an indication of the importance of these features. For this dataset, scatter plots were also generated to investigate correlation between features. From the scatter plots, the wine features are relatively independent.

2.4 Breast Cancer Dataset

The last dataset has 10 continuous clinical features for prediction of binary class "benign", "malignant". One of the features was the ID number of the patient with no impact on diagnosis, thus was removed during cleaning. We calculated mean, standard error and the largest registered value of the 10 different attributes to make them in total of 30 features for the patients. The plot of the distribution of the classes showed significantly more "benign" samples than "malignant". The plots showed highly conservative distributions of the 2 classes across many features, making this dataset a good candidate for performing feature selection. Unsurprisingly, in our scatter plots, some feature pairs did reflect high correlations, further confirming on its suitability of feature selection.

3 Method

Before assessing the performance of the models, we evaluated the correctness of our implementations by creating a toy dataset with known distributions of features and known values of class labels. We compared the posterior probabilities predicted by our NB model to the real posteriors, and calculated the accuracy of our LR-assigned class labels. The successful results from these trials validated the correctness of our implementations. In addition, we compared our results with reference solutions from Sci-Kit learn (Pedregosa et al., 2011) and papers describing experimental trials with the same classifiers over the same datasets.

To better illustrate the advantages special to each of the learning algorithms, we executed experiments to evaluate their performances. All experiments were performed with cross-validation (CV in the following context), with a default of 5-fold. In every run of CV, we selected each of the test sets with balanced case-control ratio to avoid class imbalance problem. We first ran CV of our models on all 4 datasets, and averaged CV results for each dataset to compare their model-dataset-specific general accuracy. Then, we further investigated our CV-averaged LR and NB performance as functions of dataset size. For LR, CV was used to acquire average accuracy for different learning rates, then learning rates associated with best accuracy were selected to plot the accuracy as a function of the number of iterations.

To study feature importance, we performed Drop Column Feature Importance approach to both models. More specifically, we first do CV on the whole dataset, then we dropped 1 column at a time and retrained both models. We also performed Saliency Map on LR only. For this method we did CV on the breast cancer dataset, calculated the gradient of each feature on the validation set, then reported the mean gradient.

In order to reproduce our result, we fixed the random seed to be 2020.

4 Results

4.1 Correctness over Toy Dataset

To evaluate the NB Implementation, a toy dataset composed of 6000 samples of class labels $= 1$, 4000 samples with class labels $= 0$, summing to a total of 10000 samples, was created. The values of feature 1 for samples with class label $= 1$ are randomly generated from the normal distribution $\sim \mathcal{N}(0, 1)$, and their values for feature 2 all equal to 0. For samples with $c = 0$, the feature 1 values were sampled from $\mathcal{N}(2, 1)$, with feature 2 values all equal to 1. With these given, we can calculate the real log posteriors using the prior probability and likelihood:

$$\begin{aligned} \log[p_u(c)] &= \log(\frac{Ni}{N}) \\ \log[p(x|c)]^1 &= \log(\frac{1}{\sqrt{2\pi\sigma^2}}) - \frac{x^2}{2\sigma^2} \\ \log[p(c|x)]^2 &= \log[p_u(c)] + \log[p(x|c)] \end{aligned} \tag{1}$$

We then trained NB model[3] to compute the predicted posteriors.

For the evaluation of the LR model, we used the same sample points, but defined real class labels with the logistic function:

$$y = \frac{1}{1 + e^{-wTx}} \tag{2}$$

The experiments using toy dataset were performed 100 times. Over the 100 trials, we received a max absolute error of 0.0665 for the NB model, and min accuracy of 0.9187 for LR, confirming the correctness of our implementations.

4.2 Performance w.r.t. Reference Solutions

The UCI datasets have been widely used in justifying various learning models and optimization algorithms such as feature selection. We compared the learning accuracy of our implementations on the 4 datasets to that of the Sci-Kit Learn implementation, as well as those reported by published papers. Overall, the accuracy of our models deviated insignificantly from reference performances.

We used the Adult dataset when comparing our implementation performances to Sci-Kit learn. Both Sci-Kit learn and our implementation received an accuracy for categorical NB of 0.7912, and for for gaussian NB, 0.7892. For LR, Sci-kit learn obtained accuracy of 0.8498, whereas our result was 0.8443. In summary, our implementation for NB was able to achieve the same level of accuracy as Sci-kit learn, with a slightly less accurate logistic (difference level at 10^{-3}).

In comparison with an article citing the UCI datasets (Tan et al., 2016), we found their LR accuracy on the Adult and Wine datasets are 0.8468 and 0.9503 respectively, and the equivalence of NB are 0.8073 and 0.9790. These accuracy for the adult dataset are similar to our results in both the previous test run and test results reflected in later experiments. In addition, their reported accuracy over the Wine dataset was close to our values in experiment 4.4.1 as well, confirming this high accuracy was not caused by randomness, but universal for all test runs. Furthermore, our LR and NB accuracy on the ionosphere dataset are quite closed to the accuracy reported from (Wing et al., 2003) using single layer neural net which are 0.945 and 0.920 for training and testing. We also found LR and NB prediction accuracy on the Breast Cancer dataset, reflected by the figures, are around 0.982 and 0.957 (Patgiri et al., 2018).

4.3 LR Specific Experiments

4.3.1 Learning Rate-Associated Performance

We investigated the impact of altering learning rate in the logistic model. We run a 5-CV on the all four dataset to choose the learning rate with best accuracy. The iteration limit was set to 1000 and

[1] μ was omitted since the toy dataset values were sampled from Normal distributions with $\mu = 0$.

[2] Marginals were further ignored in log-based probabilities as they serve as the same subtrahends for all datapoints.

[3] Model probability computation follows the same logic as how we compute the real log probabilities.

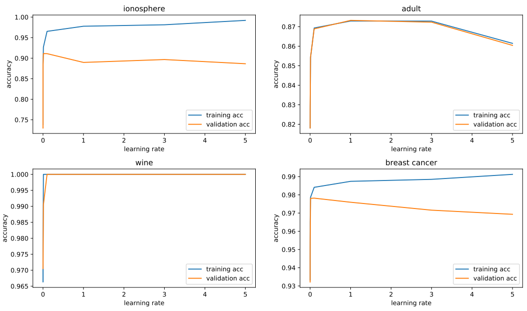

Figure 1: Results with different learning rates

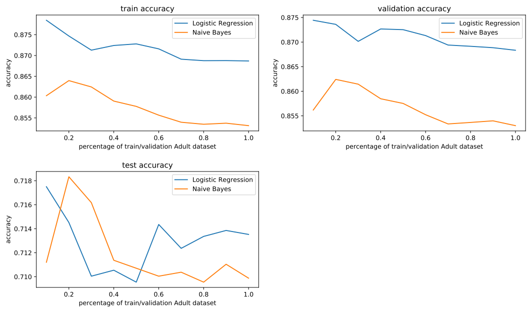

Figure 3: Results with different dataset size

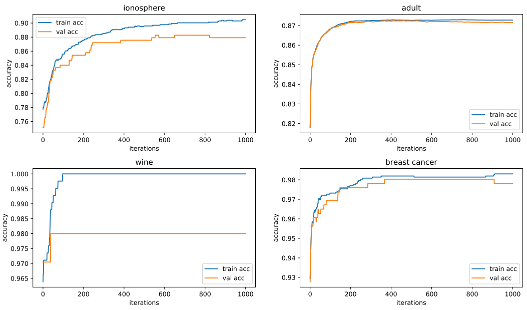

Figure 2: Accuracy

the learning rates were chosen among the following values: 1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1, 3, 5. The results were shown on figure 1.

At the beginning, in all datasets, the accuracy increased as learning rate increased. All curves are sharp at the beginning and became steady after 5*1e-5. Furthermore, as learing rate value exceeds 1, some dataset started to see a decreasing trend in accuracy.

4.3.2 Iteration-Associated Performance

Figure 2 shows the accuracy on training and validation set as a function of iterations (the results are the means of 5-CV):

We observed no sign of over fitting since all validation curves are increasing. The gap between the validation and training curves is the greatest in the wine dataset, and the smallest in adult dataset.

4.4 Performance Assessment for LR and NB

4.4.1 General Accuracy Comparison

We next compared the performances of the 2 models across the 4 datasets. The logistic classification learning rate was set to the best-performing value from previous results, and maximum iteration $= 1000$. Cost function threshold $= 1 \times 10^{-5}$. The result are summarized in table 1.

Amongst learning results from the 4 datasets, LR had 3 performances outstanded the equivalence of NB, when both models reflected nearly perfect accuracy in learning the wine dataset, presumably due to the neatness of this dataset. Noticably, the gap between the two models is tight in adult and wine dataset ($\leq$ 2%), but wide in ionosphere and breast cancer (9% and 6% respectively). Overall, LR performed better, which might be caused by the pre-runned hyper-parameters tuning CV on LR for learing rate selection. Since the learning rate selection process is computationally costly, we can conclude that NB is a better choice with a constraint on runtime complexity, whereas LR has better guarantee in prediction accuracy.

4.4.2 Accuracy as a Function of Dataset Size

For in-depth knowledge of the advantages of the 2 models, we plotted the accuracy obtained on the adult dataset (the dataset biggest in size) with modification in size of training set. Figure 3 summarizes the results obtained, with training set size = 10%, 20%, 30%, 40%, 50, 60%, 70%, 80%, 90%, 100% of its original value.

When running with smaller dataset, NB performed better than LR. As dataset dimension increases, LR outran NB. Notice in this experiment setting we did not use the learning rate with the best performance for the LR. In fact, when the best was applied, NB was always beaten by LR, regardless of the size of the dataset.

4.5 Features Importance

From figure 4, we can clearly see that NB (blue) and LR (orange) put their focus on different features, under Drop Column Feature Importance approach. For example LR think that radius (mean) is not an important feature since the accuracy increased by drop this column, while NB thinks that it is important. Even though Saliency Map tells the size of impact of each feature on the

	LR train Acc	LR val Acc	LR test Acc	NB train Acc	NB val Acc	NB test Acc
Ionosphere	0.9261	0.9113	0.8143	0.8656	0.8612	0.7286
Adult	0.8726	0.8724	0.7323	0.8531	0.8530	0.7099
Wine	1.0	1.0	1.0	0.9832	0.9710	1.0
Breast cancer	0.9830	0.9782	0.9732	0.9475	0.9366	0.9196

Table 1: Table to test captions and labels

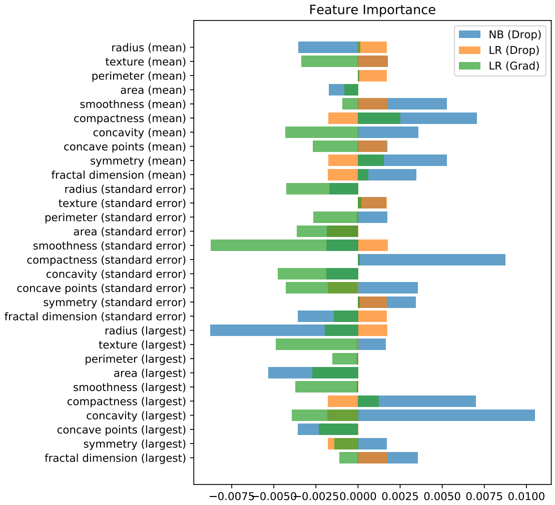

Figure 4: The length of the blue and orange bar indicate the change of accuracy repesct to the full dataset, while the length of the green bar indicate the scale of the gradient of each feature.

outputs, it does not say anything if the impact is good or bad on the accuracy. However, all three method shows that area (largest) and concave points (largest) have a strong impact on prediction. Finally, we can see that the accuracy doesn't change a lot by dropping just one feature, the greatest gap is around 0.01 and so we deduce that, for both model, many features are probably redundant. In fact, based on our data preprocessing results, we observed that many features are correlated.

5 Discussion and Conclusion

We implemented a generative (Naive Bayes) and a discriminative (Logistic Regression) model, and assessed their performances on 4 different datasets based on 5-fold cross validation.

The results showed a generally worse performance of NB comparing to LR. However, its advantage stands out when dataset size is small. Moreover, a LR model with best learning rate, though computationally costly to identify, is guaranteed to have better performance than NB. Therefore, when there is no constraint on runtime of the models, the LR model should be preferred over NB.

Through our experiments we can conclude that the choice of learning rate has a big impact on logistic model performances, since it being too small lowers the convergence speed (takes too much iteration) while it being too large leads to prediction instability (overly greedy algorithm leads to suboptimal solutions).

We also noticed how the size of dataset has a strong impact on the generalization capacity of our model. the accuracy/iterations plots clearly shown that when we have small dataset the training and validation accuracy can be very different.

Lastly, we showed that each feature of a dataset can have a different importance in LR or NB by doing Drop Column Feature Importance and Saliency Map. However, there are two particular features from the breast cancer dataset, the area (largest) and and concave points (largest) that are showed by all three approaches, which are believed to have a significant impact on prediction.

6 Statement of Contributions

Si Yi Li: data processing, model implementation and experiments
Matteo Cacciola: experiments, plotting and report write up
Agnes Liu: plotting and report write up

References

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

Andrew Y. Ng and Michael I. Jordan. 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 841–848. MIT Press.

Ripon Patgiri, Sabuzima Nayak, Tanya Akutota, and

Bishal Paul. 2018. Machine learning: A dark side of cancer computing.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *preprint*.

Yi Tan, Prakash P. Shenoy, Moses W. Chan, and Paul M. Romberg. 2016. On construction of hybrid logistic regression-naïve Bayes model for classification. In *Proceedings of the Eighth International Conference on Probabilistic Graphical Models*, pages 523–534.

S. Wing, Raymond Greenwald, C.-I Meng, V. Sigillito, and L. Hutton. 2003. Neural networks for automated classification of ionospheric irregularities in hf radar backscattered signals. *Radio Science - RADIO SCI*, 38:2–1.