Bayesian regularization-Quadratic Discriminant Analysis

Zinan ZHOU , Xinyu WANG

AOS1 - Fall 2019

Object:

We will use the classification method -Quadratic Discriminant Analysis(QDA),then implement the regularized version of the QDA method, and then apply it to the datasets provided which are the following datasets : Optdigits, Pageblocks, Satimage, Segment.The object is to compare the results obtained without and with priors.

The main equations:

- Regulanization:

  The Gaussian-inverse-Wishart prior is

$$\mu_k \sim \mathcal{N}\left(\mu_{kp}, \Sigma_k/\kappa_{kp}\right), \quad \Sigma_k \sim IW\left(\nu_{kp}, \Lambda_{kp}\right) \tag{1}$$

  We can obtain the following equations:

$$\begin{aligned} f_{\mu_k}(\mu) &\propto (\Sigma_k)^{-\frac{1}{2}} \cdot \exp\left(-\frac{k_{kp}}{2}\left(\mu-\mu_{kp}\right)^t \Sigma_k^{-1}\left(\mu-\mu_{kp}\right)\right) \\ f_{\Sigma_k}(\Sigma) &\propto (\Sigma)^{-\left(\frac{V_{kp}+p+1}{2}\right)} \cdot \exp\left(-\frac{1}{2} tr\left(\Sigma^{-1}\Lambda_{k_p}^{-1}\right)\right) \end{aligned} \tag{2}$$

- Parameter estimate for category mean:

$$\hat{\mu}_k = \frac{\sum_i z_{ik} x_i + k_{kp}\mu_{kp}}{\sum_i z_{ik} + k_{kp}} \tag{3}$$

- Parameter estimate for covariance matrix :

$$\hat{\Sigma}_k = \frac{\sum_i z_{ik} B_{ik} + k_{kp}(\mu_k-\mu_{kp})(\mu_k-\mu_{kp})^t + \Lambda_{kp}^{-1}}{\sum_i z_{ik} + V_{kp} + p + 2} \tag{4}$$

- The equation for prediction :

$$\hat{P}\left(W_k|x\right) = \frac{\hat{\pi}_k f_k\left(x_i, \hat{\mu}_k, \hat{\Sigma}_k\right)}{\sum_l \hat{\pi}_k f_l\left(x_i, \hat{\mu}_k, \hat{\Sigma}_k\right)} \tag{5}$$

- The equation for QDA:

$$X \sim N\left(\mu_k, \Sigma_k\right) \tag{6}$$

  We can write the equation as follows:

$$f_k(x) = (2\pi)^{-\frac{p}{2}} \left(\Sigma_k\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left(x-\mu_k\right) \cdot \Sigma_k^{-1}\left(x-\mu_k\right)\right) \tag{7}$$

Define inputs in the training code:

- Xapp: $n \times d$
- zapp: $n \times K$
- mprior: $K \times p$
- Sprior: $p \times p \times K$
- df_exp: should satisfy ≥ 0
- df_cov: should satisfy $\geq d-1$

It should provide for training:

- pi: $1 \times K$
- mu: $K \times p$
- Sig: $p \times p \times K$

Define inputs in the test code:

- Xtst: $n \times p$
- pi: $1 \times K$
- Sig: $p \times p \times K$

It should provide for test:

- prob: $n \times K$
- pred: $n \times 1$

Implementation:

1.Get the four datasets

2.Define the train_test function:we divide the data into train set and test set.We set Zapp and Ztst the 'class' of each data.And the other inputs as Xapp and Xtst. zapp and ztst are the binary of the class.

3.Define the function of scale.Scale the data to provide the influence of extreme data value.

4.Define the coefficient and parameters:

- n = number of the samples;
- d(or p) = the dimension of the data;
- df_exp (κ_{kp}) is a value ≥ 0;
- df_cov(ν_{kp}) is a value $\geq d-1$;
- mprior of expectations μ_{kp} for the Gaussian prior on μ_k:matrix $K \times p$. In code execution, mprior is setted to np.zeros $((K,\, p))$;
- Sprior of covariance matrices $\Lambda_{kp}^{-1}$for the inverse-Wishart prior on Σ_k:matrix $p \times p \times K$. In code execution, Sprior is setted to identity matrix.

The information of the 4 datasets:

- n_opt: 3934 p_opt: 64 K_opt: 10
- n_pag: 3831 p_pag: 10 K_pag: 5
- n_sat: 4501 p_sat: 36 K_sat: 6
- n_seg: 1617 p_seg: 19 K_seg: 7

And then we use the above coefficient to define the following parameters:

- pi =Probability of each class;
- mu :use df_exp and mprior apply the equation of $\hat{\mu_k}$;
- Sig:use df_cov and Sprior apply the equation of $\hat{\Sigma_k}$;

5.Define the prediction function:We use the obtained prior mu,sig,and pi to calculate the probability of each class by applying the equation $f_k(x)$.
the pred is to choose the biggest probability of the values.so that we obtain the prediction of the class of each sample in test.

6.Calculate the accuracy of each dataset with prior.

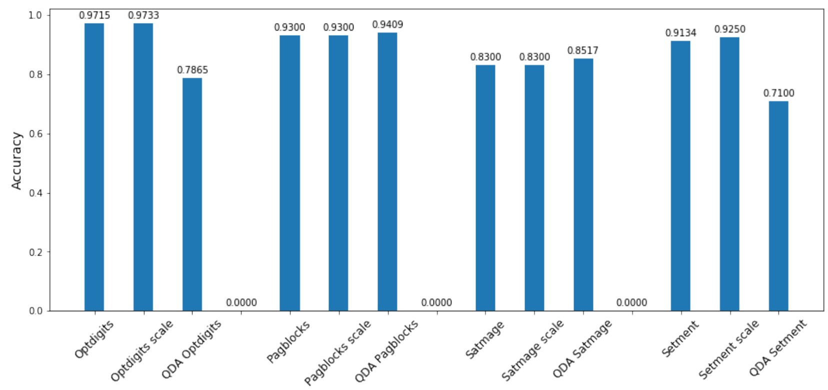

Figure 1: Accuracy of with prior, with proir after scaling and without prior

7.Call the function QDA and calculate the accuracy of each dataset directly without prior.

8.Compare the results of with prior, with proir after scaling and without prior, and draw the conclusion.(figure 1)

9.Change the value of df_exp and df_cov,and compare the different result to find the influence of the parameters.

Result:

We compare the result of four datasets ,each of them has three result:the accuracy of QDA without prior,the accuracy with prior,and the result of data after scale with prior.

From figure 1 we can clearly find that the results of Optdigits and Segment have the obvious difference between accuracy with prior and without prior.The accuracy with prior is more higher than that without prior.From the other two,they are almost the same high.We can conclude that classifier with prior will have higher accuracy in the method quadratic discriminant analysis.
And for the result between the data original and the data after scale ,they are almost the same result.

From Figure 2,we study the influence of the df_exp,which is the κ_{kp},except for the data Satmage ,the accuracy of the other three don't change too much with the vary of the df_exp. We set df_exp = 0.001, 0.01, 0.1,0.2,0.5,1,10 and 100. For the Satmage,the accuracy become lower with the increasing of the df_exp.So from the figure2,we can see when the value of df_exp is ≤ 1,the common accuracy is better than when is ≥ 1.We set the initial value of df_exp = 1.

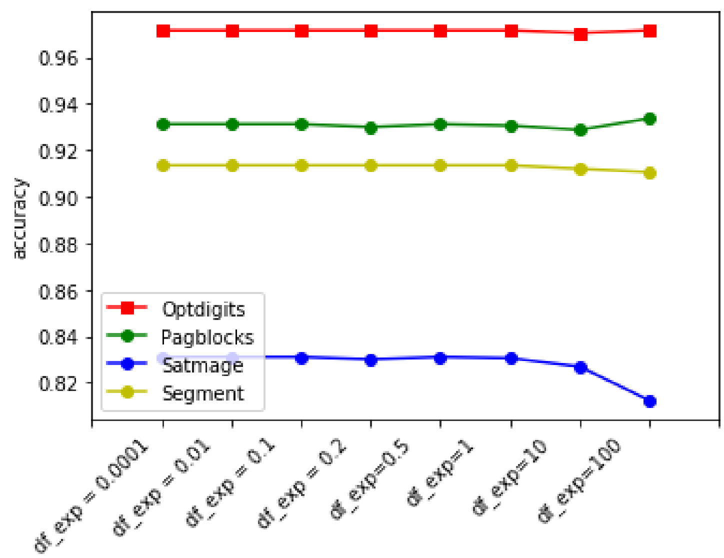

Figure 2: vary the df_exp

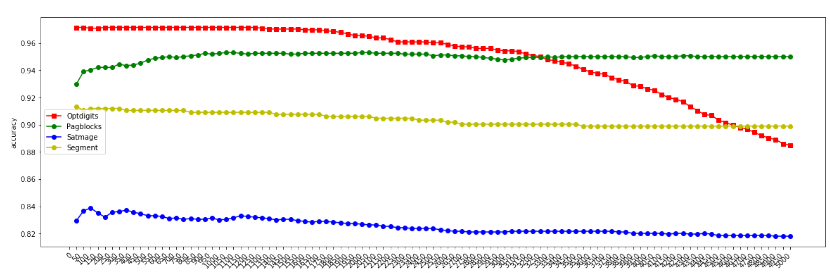

Figure 3: vary the df_exp

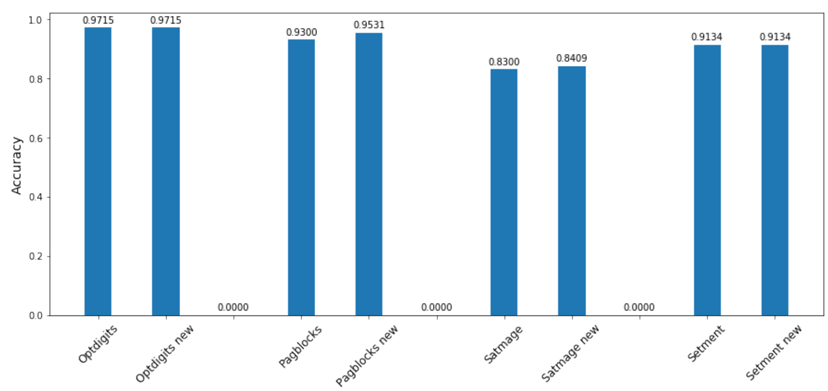

Figure 4: Compare with the new result

The initial value of degrees of freedom is $df\_cov = d + 1$, this figure is 4 data sets from d-1 to d + 5000 divided by 50. From Figure 3, Optdigits, Satmages, and Segments are global decreasing, Satmage increases a little at the begin and then decrease. The most obvious decline in data Optdigits. For the data set Pagblocks, it gradually rises until about d_pag+1000 and stabilize. So at the begin of the degree of freedom increase, it may be good for accuracy. We redo the train and test process with the new df_exp and the new df_cov which we find according to the best performance of accuracy.The result is Figure 4 .It shows that the accuracy of Pageblocks and Satimage becomes higher.We can conclude that the change of df_exp and df_cov can achieve optimization of the result.In practical,we set df_exp of each dataset is 0.1,and the df_cov is d_opt+10,d_pag+1052,d_sat+85,d_seg-1.

Conclusion:

We can conclude that applying Bayesian regularization to Quadratic Discriminant can increase the accuracy of classification.We can optimize the model by changing the parameters of the prior(κ_{kp}, ν_{kp}).We can find the best parameters according to the performance of accuracy.