Machine Learning Homework 3

1. **Ideally, how should you choose a sample out of a significantly heterogenous dataset in order to train a model on it? For this assignment, randomly split the data into test and training sets such that only 10% of the records are in the training set. Fit a simple linear regression model to predict the overall score of a player and test your model against the test set. Calculate the R^2 for the predictions you made on the test set. How many features are used in this model?**

   The dataset contains over 18,000 soccer players and we randomly split only 10% of the players in to training data and the rest into testing data.If the training data is much smaller than the test data, it would perform a bad model and it will achieve a low precision on the test data. In an ideal situation where we want to train a model on a significantly heterogeneous data, we would split 80 percent of them into the training data and 20 percent into the test data. This guarantees the model to be trained well and use the test data to validate the model.

   Besides, we would perform stratified random sampling, instead of simple random sampling. For example, if there is high heterogeneity in the overall score for people of different nationalities, we can divide the population by clubs, thereby creating strata, and randomly draw samples from each strata.

   We created dummy variables to analyze categorical variables, such as “international reputation” and “weak foot”. After dropping one level for each categorical variable, we include 85 features in the model. The simple linear regression model we applied has a R^2 value 0.890.

2. **Using the same training and test sets, fit a simple regression model but with 5-fold cross validation and predict the overall scores of players in the test set. Calculate R^2 for the predictions and compare with the R^2 from question 1. Please explain your observation.**

	model 1 - linear regression	model 2.1 - linear regression with CV	model 2.3 - linear regression with RFE and CV
R^2 on test set	0.890330	0.894787	0.889298
MSE on test set	5.219964	5.006234	5.266809
number of features used	85.000000	85.000000	60.000000

   In 5-Fold cross-validation, the data is divided into 5 subsets. Now the holdout method is repeated 5 times, such that each time, one of the 5 subsets is used as the test set and the other 4 subsets are put together to form a training set. This significantly reduces bias as

we are using most of the data for fitting, and also significantly reduces variance as most of the data is also being used in the validation set. By comparing model 1 and model 2.1, we can see that R^2 increases by 0.004 after applying 5-fold cross-validation in the linear regression model.

In addition, we applied 5-fold cross-validation to another simple linear regression model (2.3) with only 60 variables. Model 2.1 has a higher R^2 and a lower MSE than model 2.3, both indicating a better performance, so we decided to keep using model 2.1 with 91 features.

3. **Using the training data from question 1, fit a Lasso regression to predict the overall scores of players in the test set. Use the default value of alpha (alpha is used to tune the penalty. Higher the value of alpha, fewer the number of features) parameter, which is usually 1. How many features are being used by the model? Calculate the R^2 for the predictions you made on the test set and compare with the R^2 from question 1. Please explain your observation.**

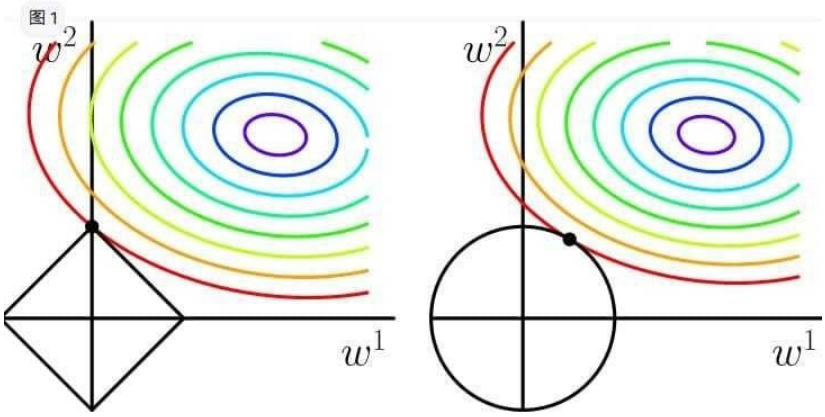
   The Lasso regression model uses 23 features. Compared to the R^2 of linear regression (0.890), the R^2for Lasso is smaller (0.851). This can be explained since the value of R^2naturally increases as more variables are included in the model. However, we also observed that the adjusted R^2for Lasso is also smaller than that of the linear regression. Since the adjusted R^2 takes into account and adjust for the number of predictors, we can say that the linear regression model has more explaining power than the Lasso regression.

	model 1 - linear regression in Q1	model 3 - lasso regression in Q3
R^2 on train set	0.902304	0.858410
Adjusted R^2 on train set	0.897515	0.856597
R^2 on test set	0.890330	0.850822
number of features used	85.000000	23.000000

4. **Do you expect your answer to question 3 to change if you are using ridge- or log- instead of lasso- penalties? Please explain.**

	model 1 - linear regression in Q1	model 3 - lasso regression in Q3	model 4 - ridge regression in Q4
R^2 on train set	0.902304	0.858410	0.902187
Adjusted R^2 on train set	0.897515	0.856597	0.897451
R^2 on test set	0.890330	0.850822	0.890706
number of features used	85.000000	23.000000	84.000000

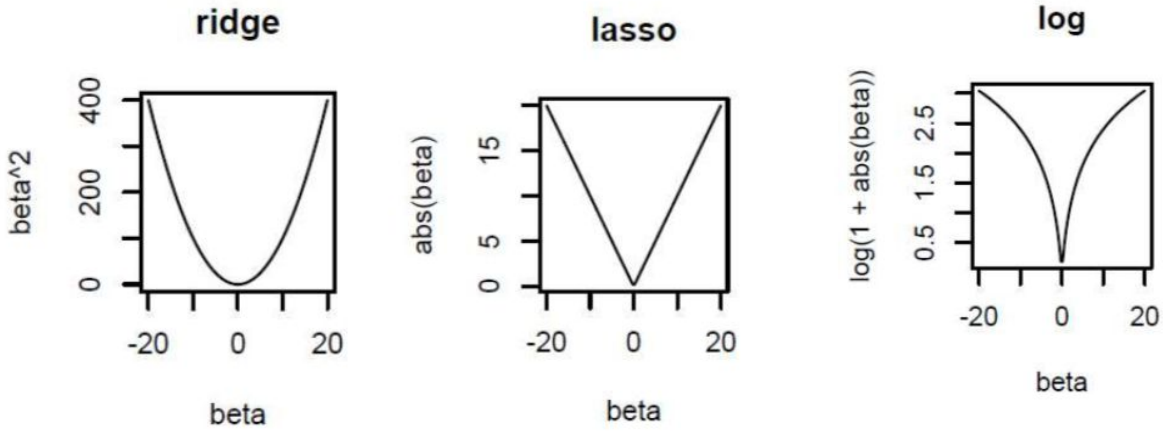
Ridge-penalties use the square of the magnitude of the coefficients as the cost function. Ridge regression shrinks the coefficients and avoids overfitting of data. Comparing to Lasso regression, Ridge regression includes more features in the model, because Lasso not only penalizes high beta values but also converges irrelevant variable coefficients to zero. Please see graph representation (left – Lasso, right – Ridge).



ℓ_1-ball meets quadratic function. ℓ_1-ball has corners. It's very likely that the meet-point is at one of the corners.

ℓ_2-ball meets quadratic function. ℓ_2-ball has no corner. It is very unlikely that the meet-point is on any of axes."

Log-penalties use log(1+|beta|) as the loss function. In this case, there is a clear absolute deviation in the middle, where beta is equal to zero. Therefore, using log-penalties will lead to less number of features being included in the model.



5. **Now try to fit a Lasso regression to predict the overall scores of players with an ideal value for alpha. Your code should try to test different values of alpha and use the ideal one. What, according to your code, is the ideal value of alpha? How many features are being used by the model? Calculate the R^2 for the predictions you made on the test set and compare with the R^2 from question 1. Please explain your observation.**

   Alpha values we tested: 1e-15, 1e-10, 1e-8, 1e-4, 1e-3, 3e-3, 4e-3, 5e-3, 7e-3, 1e-2, 0.05, 0.1, 1, 5, 10, 20

   The ideal value of alpha is 0.004, and 63 features are used in the model. The R^2 for test set predictions is now closer to the R^2 from Q1. Therefore, we can improve the performance of Lasso regression by adjusting the parameter alpha.

	model 1 - linear regression in Q1	model 3 - lasso regression in Q3	model 5 - lasso regression with ideal alpha in Q5
R^2 on train set	0.902304	0.858410	0.901005
Adjusted R^2 on train set	0.897515	0.856597	0.897453
R^2 on test set	0.890330	0.850822	0.890345
number of features used	85.000000	23.000000	63.000000

6. **Calculate AIC and BIC for the models you built in question 1 and question 4. According to each of the measures, which is the better model? Is BIC always greater than AIC? Please explain. Compare the AICs with the corresponding corrected AICs.**

   AIC is an estimate of a constant plus the relative distance between the unknown true likelihood function of the data and the fitted likelihood function of the model. Therefore, a lower AIC means a model is considered to be closer to the truth. The corrected AICc is the corrected form for small sample size. BIC is an estimate of a function of the posterior probability of a model being true, under a certain Bayesian setup. Therefore, a lower BIC means that a model is considered to be more likely to be the true model.

   The value comparison of AIC and BIC is not fixed. When n>e^27.389, n is the number of observations in the test set, AIC is greater than BIC. Otherwise, BIC has a greater value than AIC. In our case, where n=16387, BIC is always larger than AIC.

	model 1 - linear regression in Q1	model 5 - lasso regression with ideal alpha in Q5
R^2 on train set	0.902304	0.901005
Adjusted R^2 on train set	0.897515	0.897453
R^2 on test set	0.890330	0.890345
number of features used	85.000000	63.000000
AIC	27251.362641	27205.120053
BIC	27913.927592	27698.191645
AICc	27252.259519	27205.614080

According to both AIC and BIC, the Lasso regression in Q5 is better than simple linear regression in Q1.

7. ICs are alternatives to CVs. Do you trust them equally? Please explain.

We would consider ICs fully equal to CVs only under certain conditions. Moreover, the components of ICs, namely AIC and BIC, need to be discussed separately.

On one hand, the AIC and leave-one out cross-validation (LOOCV) are asymptotically equivalent (Stone, 1977). However, when we are trying to get asymptotic equivalence, we need to have a very large sample with not that large number of parameters. In this case, the −2log(likelihood) part will be more important than how we penalize the number of parameters. Therefore, with a smaller sample size, AIC will work better; with an infinite-dimensional case, LOOVC will perform optimally. They are often used in prediction and model assessment.

On the other hand, BIC is equivalent to leave-k-out cross-validation (LKOCV) where k=n[1−1/(log(n)−1)], with n= sample size (Shao 1997). BIC is used as an estimate of a function of the posterior probability of a model being true, under a certain Bayesian setup. They are more often used for estimating the test errors when selecting the best models.

Therefore, the applications of the ICs as well as different CVs focus on different goals. In sum. we cannot treat them fully equally.

References

Stone M. (1977) An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. Journal of the Royal Statistical Society Series B. 39, 44–7.

Shao J. (1997) An asymptotic theory for linear model selection. Statistica Sinica 7, 221-242.