In the application project, I implemented logistic regression, decision tree, logistic regression with PCA and Gaussian

Logistic regression is a very common algorithm for linear classification. It uses a sigmoid activation function and then use gradient descent to minimize the loss function. In my project, it has an average accuracy about 95%.

In [probability theory](https://en.wikipedia.org/wiki/Probability_theory) and [statistics](https://en.wikipedia.org/wiki/Statistics), a Gaussian process is a [stochastic process](https://en.wikipedia.org/wiki/Stochastic_process) (a collection of random variables indexed by time or space), such that every finite collection of those random variables has a [multivariate normal distribution](https://en.wikipedia.org/wiki/Multivariate_normal_distribution), i.e. every finite [linear combination](https://en.wikipedia.org/wiki/Linear_combination) of them is normally distributed. The distribution of a Gaussian process is the [joint distribution](https://en.wikipedia.org/wiki/Joint_distribution) of all those (infinitely many) random variables, and as such, it is a distribution over functions with a continuous domain, e.g. time or space. In the Gaussian process classifier, I use a rbf kernel with r = 1 to implement GPC. I got about 74% accuracy.

The last one I use is logistic regression after PCA. Principle component analysis (PCA) is one such method.

• This is a powerful, popular and perhaps the simplest dimensionality reduction method that has various applications. It can also be viewed as a feature selection method, but not in a conventional way.

• Two main ideas are behind PCA: – Adjust (linearly) the view point so that the direction of the major variance becomes one of the coordinate. – Or, PCs are the coordinates along which the data vary the most. – Compress the data by taking fewer coordinates. Pros: less data to store or transmit. Cons: los of information. – The second idea can also be considered as feature selection in the transformed space

I left 80% of the information which decrease the dimension from 31 to 4. After that. I use logistic regression to do the classification. I get 96.8% accuracy which is the best in all implementation project.

A decision tree is a [decision support](https://en.wikipedia.org/wiki/Decision_support_system) tool that uses a tree-like [graph](https://en.wikipedia.org/wiki/Diagram) or [model](https://en.wikipedia.org/wiki/Causal_model) of decisions and their possible consequences, including [chance](https://en.wikipedia.org/wiki/Probability) event outcomes, resource costs, and [utility](https://en.wikipedia.org/wiki/Utility). It is one way to display an [algorithm](https://en.wikipedia.org/wiki/Algorithm) that only contains conditional control statements.

Decision trees are commonly used in [operations research](https://en.wikipedia.org/wiki/Operations_research), specifically in [decision analysis](https://en.wikipedia.org/wiki/Decision_analysis), to help identify a strategy most likely to reach a [goal](https://en.wikipedia.org/wiki/Goal), but are also a popular tool in [machine learning](https://en.wikipedia.org/wiki/Decision_tree_learning). I get a 91.2% accuracy for this method.

To compare all the algorithm, I use wilcoxon’s signed-rank test to choose which is the best. I run the three algorithms on the same partition using 10 fold cross validation. I will list the 10 fold cross validation accuracy as follows.

Logistic regression:

[0.8771929824561403,

0.9473684210526315,

0.9473684210526315,

0.9649122807017544,

0.9649122807017544,

0.9473684210526315,

0.9824561403508771,

0.9649122807017544,

0.9473684210526315,

0.9636363636363636]

Gaussian Process:

[0.45614035087719296,

0.7894736842105263,

0.7192982456140351,

0.6491228070175439,

0.6666666666666666,

0.8771929824561403,

0.7543859649122807,

0.8421052631578947,

0.7894736842105263,

0.8545454545454545]

Logistic regression after PCA:

[0.9473684210526315,

0.9473684210526315,

0.9473684210526315,

0.9824561403508771,

0.9473684210526315,

0.9824561403508771,

1.0,

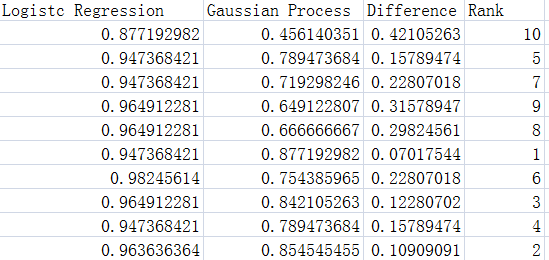
0.9824561403508771,

0.9824561403508771,

0.9636363636363636]

Test 1

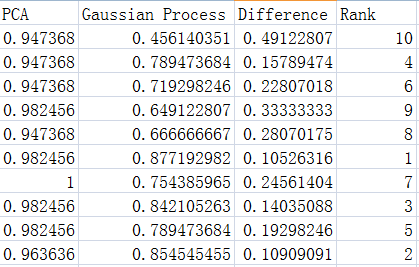
H0: logistic regression H1: Gaussian Process



We can see that Twilcox = 55 > V . I will choose logistic regression.

Test 2

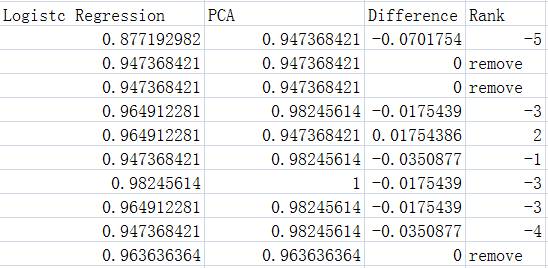
H0: logistic regression with PCA H1: Gaussian Process



We can see that Twilcox = 55 > V. I will choose logistic regression After PCA.

Test 3

H0: Logistic regression H1: logistic regression with PCA

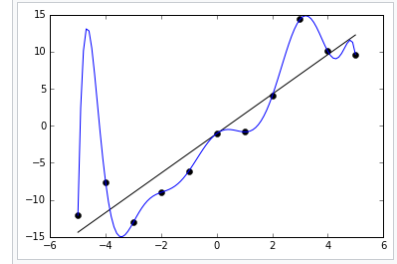


W12 = 2, W21 = 18. Twilcox = 2 < 8. We reject logistic regression and choose PCA

So the order for the performance of 4 algorithms is logistic regression with PCA, logistic regression, decision tree and Gaussian tree.

The best method is logistic regression with PCA. I reduce the dimension from 31 attributes to 4 attributes which has 80% information left. There is a problem that why logistic regression with PCA has better performance with less information. The reason I think is because of noise. For example, overfitting is a common problem in machine learning. It means that if the algorithm fit the training example very well which means use most of the information of the training data, the algorithm may perform very bad in the test set. Figure below shows a good example of overfitting.

In this figure, blue line fits the training data very well which is similar to logistic regression without PCA. All the information are used in blue line while it encounters overfitting problem. PCA in this project is like a noise removal method, it decrease the complexity while keeping most of information. So logistic regression performs very well even if it has less information.



In this project, linear classifier always performs better than the other methods since the data is linearly separable. Gaussian process performs worst because it is not good at linear separable data. Decision tree also gives a reasonable performance, because decision tree is suitable for nearly all type of data.