Homework 3

CS420, Machine Learning, Shikui Tu, Spring 2020

1 SVM vs. Neural Networks

1.1 Data A

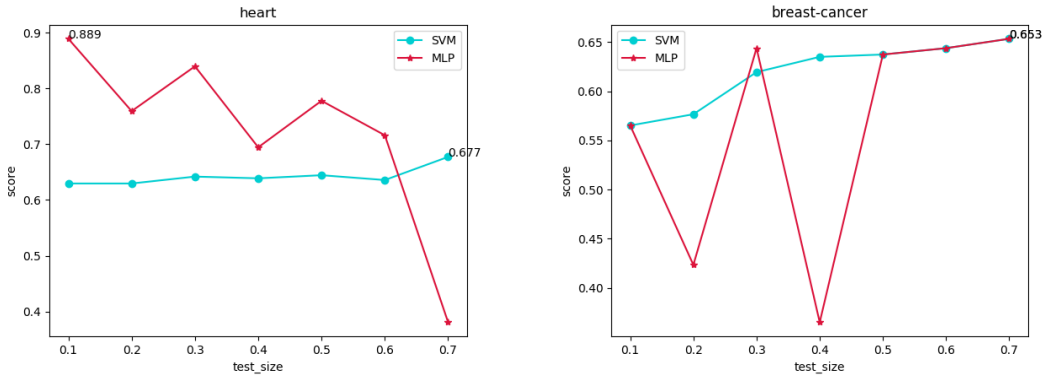
From Data A, I choose datasets named ”heart” and ”diabetes” for experiments.

Table 1: Data Description

name	class	data	feature
heart	2	270	13
breast-cancer	2	683	10

1.1.1 Comparison between SVM and MLP for different training sizes

Because neither the heart and breast-cancer data sets have test samples, I used $train\_test\_spilit$ in sklearn to divide the datasets into a training set and a test set. The reasult of different training sizes for SVM and MLP are shown in Figure 1.It can be seen from the results that the number of training sets has a great influence on the results. It stands to reason that as the training samples increase, the experimental results will change from underfitting to overfitting, that is, the model's score in the test set should first increase and then decrease. However, in this experiment, the score fluctuates with the increase of training samples in MLP, probably because the dataset I selected is too small.

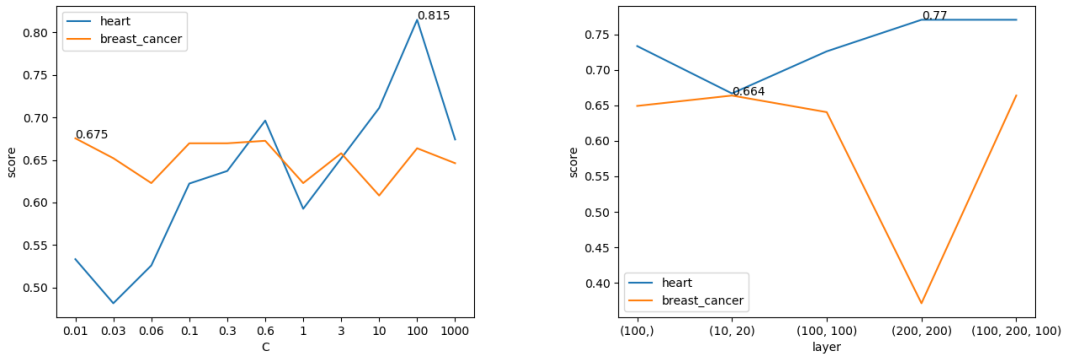


(a) (b)

Figure 1: test_size N of $train\_test\_spilit$ in [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7].

1.1.2 Comparison over different cost value in SVM and layers in MLP

In this part, I tested the effect of different C in SVM and different layers in MLP.The results are shown in fig 2. It can be seen from the results that C and layers have a great influence on the results. C represents the tolerance level of the training set error and the appropriate C can avoid overfitting and underfitting. Besides, for MLP, if the number of hidden layers in the MLP is too small, the network will lack the learning ability, while too much will cause the network structure to be complicated and slow to train. Therefore, it is very important to choose the appropriate C in SVM and the number of hidden layers in MLP.



(a) (b)

Figure 2: different C in SVM and layers in MLP.

1.2 Data B

In this part, I chose the CIFAR-10 dataset. Due to limited computing resources, I performed PCA dimensionality reduction and StandScaler preprocessing on the data. The experimental results are shown in Table 3. From the results, we can see that the performance of SVM becomes worse after dimensionality reduction, which shows that PCA damages some important information of the original picture. In addition, it is obvious that the classification effect of neural networks on large data sets is much better than that of SVM.

Table 2: Data Description

name	class	training samples	testing samples	feature
CIFAR-10	10	50000	10000	3072

Table 3: Data Description

classifier	accuracy
SVM(pca-0.1)	0.2833
SVM(pca-0.3)	0.2737
SVM(pca-0.5)	0.2758
SVM(pca-0.7)	0.2754
SVM(pca-0.9)	0.2754
SVM(pca-1.0)	0.5213
ResNet	0.936
VGG	0.924

Strength:

1. The essence of the convex optimization method is to ensure optimality. The solution is guaranteed to be a global minimum, not a local minimum;

2. SVM can use different kernel functions to improve classification performance;

3. SVM works well on low-dimensional and high-dimensional data spaces. It can effectively work on high-dimensional data sets, because the complexity of the training data set in SVM is usually characterized by the number of support vectors rather than dimensions;

Weakness :

1. The space cost and time cost of training are large;

2. It is difficult to solve multi-classification problems with SVM;

2 Casual Discovery Algorithms

In this part, I studied the relationship between lung cancer, smoking and some related factors, such as the pressure between peers, whether they cough, whether their fingers are yellow, whether they are focused, whether they have a traffic accident, etc. I downloaded the dataset from $www.phil.cmu.edu/tetrad/$.

I use PC algorithm for causal discovery. The PC algorithm has 4 main steps:

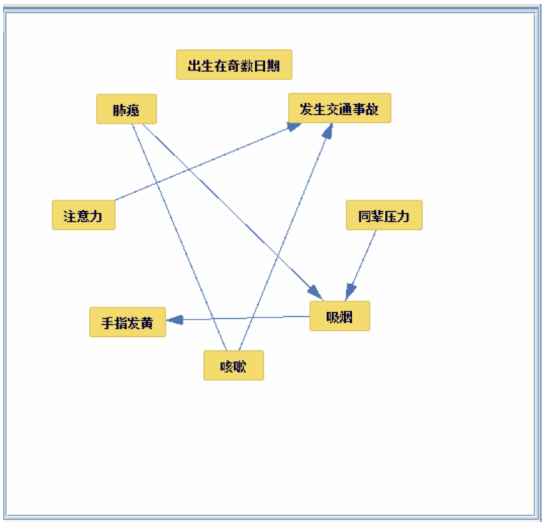
1. Form the complete undirected graph;

2. Remove edges according to n-order conditional independence relations;

3. Orient edges by v-structures;

4. Orient edges

The result is shown in Figure 3. It can be seen from the results that the birth date has no causal relationship with other factors. Smoking can cause yellowing of fingers, and peer pressure can cause smoking. Inattention and coughing can cause traffic accidents. It is worth noting that the result in lung cancer will cause smoking, which is not in line with common sense. I speculate that the reason is that I trimmed the data set and affected the judgment result.



(a)

Figure 3: casual discovery