Problem 1 One-by-one Feature Selection

* Method:

In this problem, we use Similarity-Based: **Fisher Score** method, which refers to lecture slide: DSLecture 5 Feature Selection p.25.

Hence, we follow the formula of Fisher score on lecture slide and use it as our score function of each feature.

Hence, our implementation is: For each feature, calculate its score function, and sort these features based on their score function.

* Experiment Result:

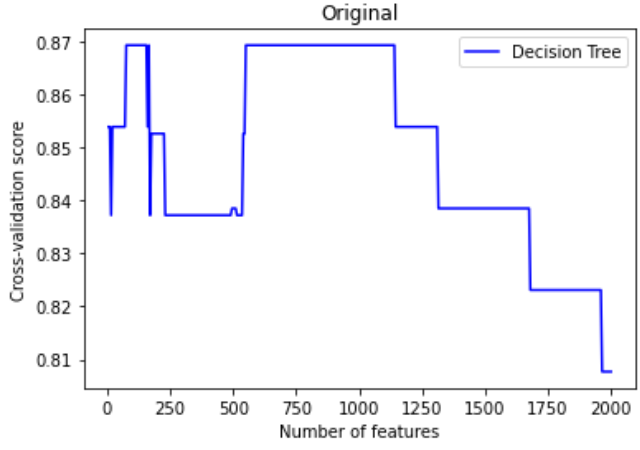
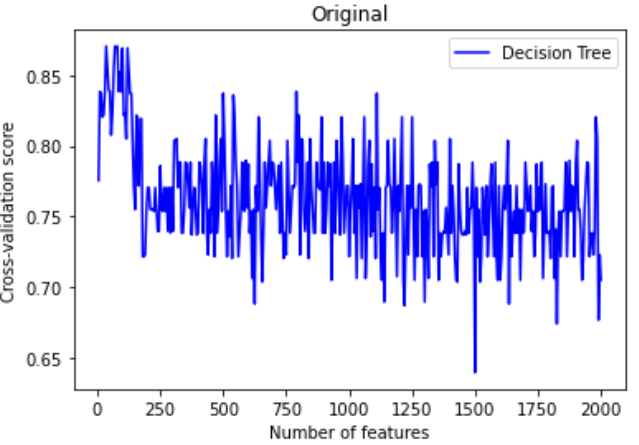
At first, to ensure that our implementation is right, we sort the features in the ranking\_idx **randomly**. And get the better result of score/# of feature = 0.8538461538461538 / 460 than SVC. So we can use it as our baseline.

Then we implement prob.1 and use the sample code in feature evaluation part. We get results:

|  |  |  |
| --- | --- | --- |
|  | DecisionTree | SVC |
| scores | 0.8705128205128204 | 0.8692307692307694 |
| Number of feature | 35 | 75 |

Visualization:

Decision Tree: SVC:



Comment:

As above, we can see that our implementation indeed has improved, whose results are better than that of the baseline.

However, although the **features are cut significantly**, we can see that the **improvement in the score is slight**

Moreover, we can see the visualization of two methods, both of **can’t see a positive trend** of the curve. It means **the order of features in ranking\_idx is not exactly representing the significance of feature evaluation**, else we should observe a positive trend since as we add more “significant”(at the front of ranking\_idx) features into evaluation, we should get a better score.

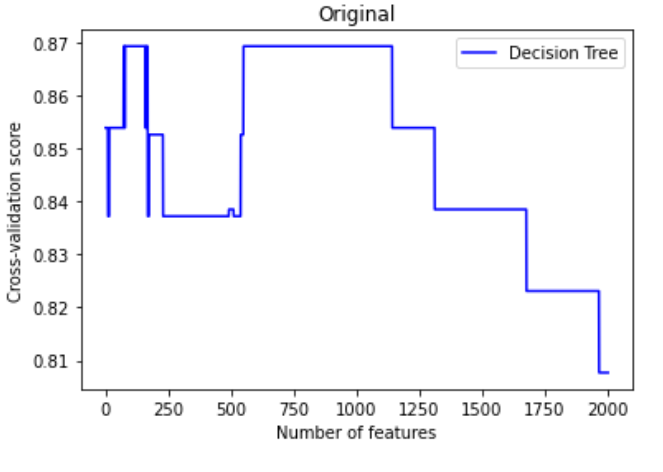
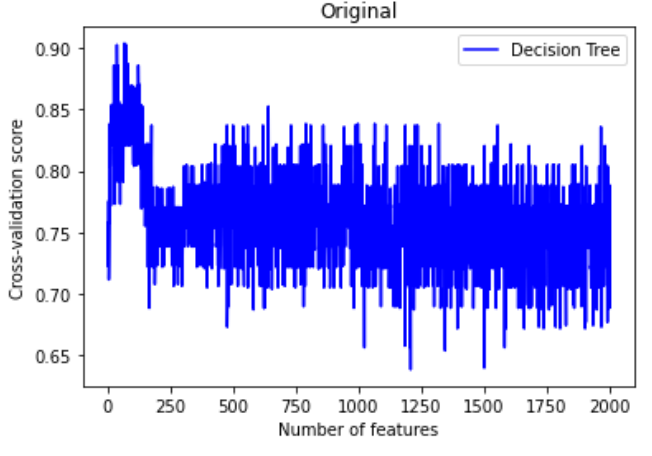
Or discussing more precisely, ideally, ranking\_idx sort features based on the significance of feature evaluation of each feature. We can imagine that the front part of features is good for feature evaluation, the medium part of features can be seen as no help on evaluation, and the end part of features may worsen the evaluation. Hence, the visualization curve should have a positive trend, flat trend, and negative trend when we take more features into consideration. However, we observe the curve differs far away from the ideal, so we can say that one-by-one selection is not a good method.

Next, we observe that the sample code only examines the score that takes the first 5\*n features into evaluation and ignores other possibilities. Hence, we next try to **examine the score that takes the first n features into evaluation**. We get the results:

|  |  |  |
| --- | --- | --- |
|  | DecisionTree | SVC |
| scores | 0.9038461538461539 | 0.8692307692307694 |
| Number of feature | 67 | 72 |

Visulization:

Decision Tree: SVC:



Comment:

As above, we can see that we improve the score to about 0.9 and cut #\_of\_features to 67. However, the problem discussed above still exists since we use the same score function in this experiment.

Problem 2 Subset-Based Feature Selection

The metaheuristic we choose here is **PSO (particle swarm optimization)**.

* Idea of the algorithm:

Following the idea of PSO, we construct a search place with 2000-D(since the dataset has 2000 features) and boundaries [0,1] at each dimension. Each particle in the search space represents a “method” of feature selection. Given a particle x, we have the selection method:

**For i=1~2000, feature fi is chosen if x[i]>0.5, else we don’t choose fi.**

Where x[i] means the position of x along the i-th dimension.

Hence, we run the PSO algorithm to find the best global position, which gives the best feature selection that achieves a minimal cost function during training.

* Algorithm:

The objective function(/cost function) we use is:

**f(x) = a \* (1 - P) + (1-a) \*gamma\*(#\_selected\_features/#\_features).**

Where P is the **feature evaluation score**. We use the method in the sample code and choose the **Decision Tree with 5-fold validation** method.

Since we want to **maximize P and minimize the number of features** we choose, we need to **minimize f(x)**.

And “a” is a tunable parameter between 0-1 that we can tune if we prefer a **better evaluation score** or a **fewer number of features selection**.

And “gamma” is a scaling factor that balances the importance of “P” and “#\_selected\_features”.

As for the implementation of PSO algorithm, we refer to the lecture slide: DSLecture 5 Feature Selection p.50 and implement the PSO algorithm.

However, there are some differences in my algorithm.

At first, we add a tunable parameter **V\_max** so that the velocity won’t be too large to make particles move out of search space.

And we also add security to ensure that after a particle updates its position, it won’t go beyond the search space.

Moreover, we make “**w**”, the weight of the original velocity of particles, linearly decay to gradually slow down the pace during training.

The tunable parameters are as below:

1. a : a tunable parameter in the objective function. a = 0 – 1
2. gamma: a scaling factor in the objective function. Default set gamma=1
3. w: weight of particle’s original velocity
4. phi\_p: weight of particle’s probabilistic velocity towards local best-known position
5. phi\_g: weight of particle’s probabilistic velocity towards global best-known position
6. lr: the proportion at which the velocity affects the movement of the particle. Here, we **don’t restrict lr=0~1** since we have no idea to tune w, phi\_p, and phi\_g parameters.
7. Iteration: the round of PSO algorithm would run
8. S: number of particles in the search space
9. K: Vmax = k\*((b\_up-b\_low)/2) = 0.5\*k , k = 0~1. Default set k=1.
10. w\_decay: Boolean to decide if “w” is a constant or would linearly decay during training.

* Intuitive parameter setting:

To find an appropriate “a”, under the condition: gamma=1, and other parameters are set “normally”(weights of velocity≒1, iteration=10, S=100, lr=0.8), we have tried a=0.1 and a=0.9, getting results:

|  |  |  |
| --- | --- | --- |
|  | a=0.1 | a=0.9 |
| P | 0.8679487179487179 | 0.9012820512820513 |
| #\_of\_selected\_feature | 876 | 1015 |

As above, we can see that if we choose too small “a”, “P” may be even worse than the baseline. To see if this method can achieve a better P than that in problem1, we should choose a = 0.9.

To find appropriate velocity weights, which contain “w”,”phi\_p”, and “phi\_g”, we can see the formula in the PSO algorithm that says:

**vi,d = w \* vi,d + phi\_p\*rp\*(pi,d-xi,d) + phi\_g\*rg\*(gd - xi,d)**

Hence, since the search base has boundary [0,1], we should want **vi,d has an average value = 0.2**. Moreover, since rp,rg ~U(0,1), which has an average value = 0.5, and we estimate that (pi,d-xi,d),(gd - xi,d) have an average value = 0.5 since both pi,d/gd and xi,d are bound in [0,1]. Finally, we want three velocities to have the same contribution to vi,d, we should **set “w”,”phi\_p”, and “phi\_g” to a ratio about 1:1:1** since in average, vi,d = w\*0.25+phi\_p\*0.5\*0.5+ phi\_g\*0.5\*0.5. (here we assume vi,d has average value=0.25 for convenience)

To find an appropriate “S”, since the search space is 2000-D, we should set S to an order of 2000. However, considering the training time, we choose S=1000.

Hence, we initialize parameters as below:

a = 0.9

gamma=1

w = 0.9 / 3

phi\_p = 0.8 / 3

phi\_g = 0.8 / 3

lr = 1

iteration = 20

S = 100

k = 1

w\_decay = False

* Experiment for parameters tuning:

Exp1: find an appropriate lr

In Exp1, we tune lr to find which lr can achieve better performance. It will determine the “exact” average velocity during training.

Exp2: find appropriate velocity weights ratio

In Exp2, we use the best lr we found in Exp1, and we tune different velocity weights ratios to see which ratio can achieve better performance.

Exp3: find appropriate “a”

In Exp3, we use the best parameters found in Exp2, and we tune different “a” to see which one can achieve better performance.

Exp4: see the effect of w\_decay

In Exp4, we use the best parameters found in Exp3, and we turn w\_decay = True. Then test bigger and smaller “w” to see if it helps on boosting performance.

* Experiment Result:

Exp1:

We set

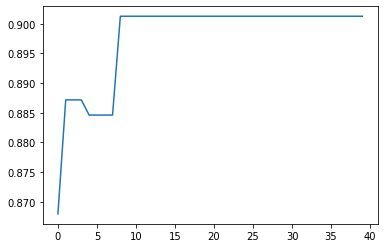
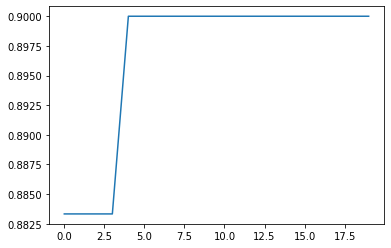
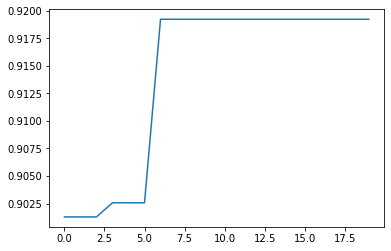
a=0.9 w=0.9/3 phi\_p=0.8/3 phi\_g=0.8/3 iter=20/40 S=1000 k=1

and tune different lr to results:

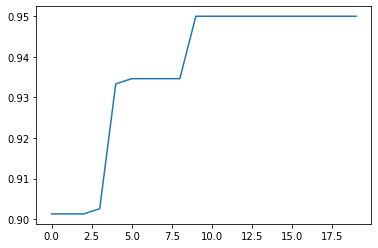
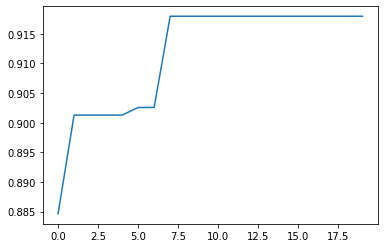
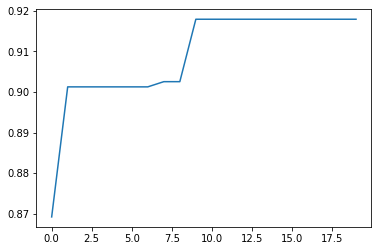
|  |  |  |
| --- | --- | --- |
|  | P | #\_selected\_feature |
| lr=0.2 (iter=40) | 0.9013 | 975 |
| lr=1 (iter=20) | 0.9000 | 981 |
| lr=2.8 (iter=20) | 0.9192 | 977 |
| lr=3 (iter=20) | 0.9500 | 988 |
| lr=3.2 (iter=20) | 0.9179 | 922 |
| lr=5 (iter=20) | 0.9179 | 859 |
| lr=8 (iter=20) | 0.8833 | 1038 |

Visualization: (x-axis: iteration times, y-axis: P(evaluation score))

lr=0.2 lr=1 lr=2.8

lr=3 lr=3.2 lr=5

lr=8



Comment:

As above, we have observations below:

1. At smaller lr(=0.2), it takes more rounds to achieve its best performance.
2. At too big or too small lr, performance gets worse.
3. lr = 3 has the best performance.

Exp2:

Using the result of Exp1, we set

a=0.9 lr = 3 iter=20 S=1000 k=1

and tune the ratio of three velocity weights while keeping the average velocity at 0.2. That is, if w : phi\_p : phi\_g = 3 : 1 : 1, we set

w=0.8\*3/5 phi\_p=0.8\*1/5 phi\_g=0.8\*1/5

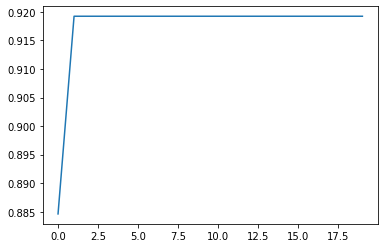
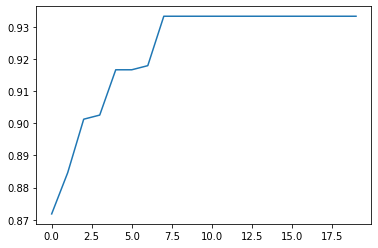
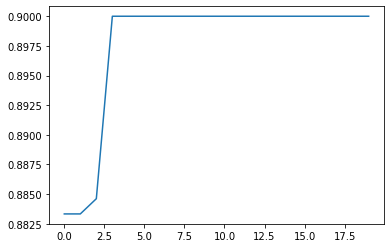
since vi,d = w\*0.25+phi\_p\*0.5\*0.5+ phi\_g\*0.5\*0.5

Results:

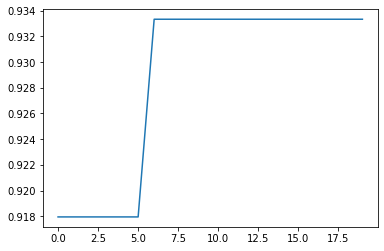
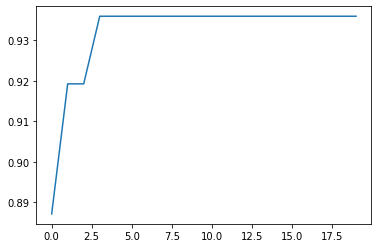
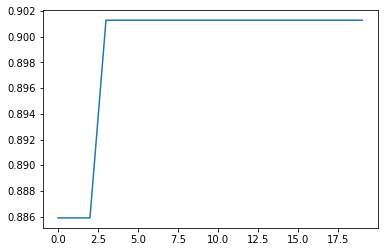
|  |  |  |
| --- | --- | --- |
| w : phi\_p : phi\_g | P | #\_selected\_feature |
| 3 : 1 : 1 (w big) | 0.9192 | 891 |
| 1 : 3 : 3 (w small) | 0.9333 | 957 |
| 1 : 3 : 1 (phi\_p big) | 0.9000 | 909 |
| 3 : 1 : 3 (phi\_p small) | 0.9333 | 764 |
| 1 : 1 : 3 (phi\_g big) | 0.9359 | 948 |
| 3 : 3 : 1 (phi\_g small) | 0.9013 | 825 |

Visualization:

w big: w small: phi\_p big

phi\_p small: phi\_g big: phi\_g small:

Comment:

As above, we have observations below:

1. If we make ”w” or “phi\_p” bigger, performance gets worse. Since the global best position is the really best, whose contribution should not be less.
2. However, the experiment results told us that setting the ratio of three velocity weights to 1:1:1 may be the best, which achieves the highest P.
3. There are some explanations for observation 2.. Maybe it is True that the ratio 1:1:1 is the best. Or it is just luck that we see it the best. Or we may have misestimated the formula of average velocity, which may need more rigorous probabilistic provement.

Exp3:

Based on the result of Exp1 and Exp2, we set:

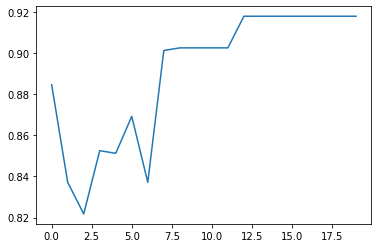
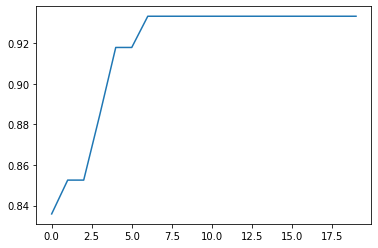
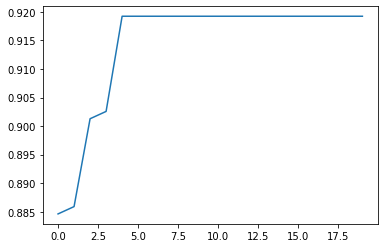
w=0.9/3 phi\_p=0.8/3 phi\_g=0.8/3 lr=3 iter=20 S=1000 k=1

and tune “a” to get results:

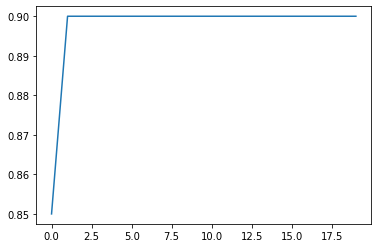
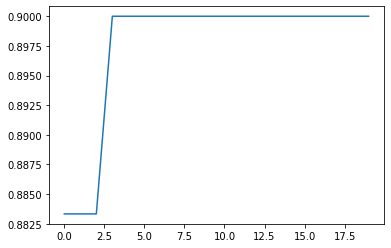
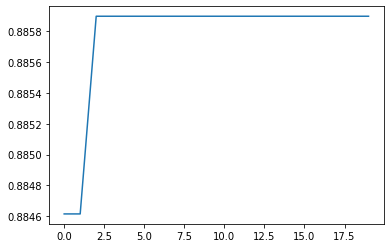
|  |  |  |
| --- | --- | --- |
|  | P | #\_selected\_feature |
| a=0.1 | 0.9179 | 543 |
| a=0.3 | 0.9333 | 684 |
| a=0.5 | 0.9192 | 843 |
| a=0.7 | 0.9000 | 939 |
| a=0.8 | 0.9000 | 1006 |
| a=1 | 0.8859 | 994 |

Visualization:

a=0.1 a=0.3 a=0.5

a=0.7 a=0.8 a=1

Comment:

As above, we have observations below:

1. It is indeed that when we have “a” smaller, we minimize the #\_selected\_feature, while when “a” is bigger, P doesn’t correspondingly get bigger. We guess it’s because the contribution of P and that of #\_selected\_feature to f(x) are not the same. Tune gamma
2. At small “a”, since the plot

Problem 3 ARIMA Forecast

* Analyze Data

At first, we plot some properties of data to know more about the data.

一開始拿純training data去fit，發現MSE都落在10萬上下，最好的不過7萬多多

後來發現預測必從最後一點開始預測，於是撥了一點尾巴的train給test，發現結果有變好。變成落在6,7萬，最好的出現5萬多的MSE了