Machine learning homework01

Name: 張天佑

ID:M10715812

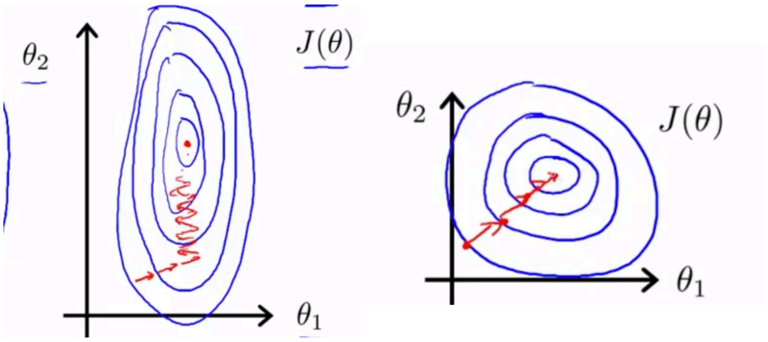
1.1

a.

If you use the above and below diagrams, it will cause the convergence to be too slow during training.

When the difference between the two feature intervals is very large, the contour line that is easy to form is elliptical. It is very likely that the iterative time will take the “zigzag” route (vertical long axis), which leads to the need to iterate many times to converge.

When the two feature intervals are not large, the corresponding contours will become rounded, and the gradient will be faster when the gradient is solved.



Reference:https://blog.csdn.net/index20001/article/details/78044971

b.

The model built according to the bottom one is too complicated and easy to cause overfitting. The separation ability of different models is not the same according to VC dimension, The assumed space is too large, the dimension is too high, and the middle one is preferred when it has the same effect according to the Occam’s razor.

1.2

Advantages:

1. This method is simple and effective.

2. There is no need to generate additional data to describe the rules. Its rule is to train the data (samples) itself, not to ask for consistency of the data, that is, there can be any noise.
3. Although the KNN method relies on the limit theorem in principle, it is only related to a very small number of adjacent samples in class decision making. Therefore, this method can better avoid the imbalance of the number of samples.
4. The KNN method makes the most direct use of the relationship between samples, which reduces the adverse effects of inappropriate selection of category features on the classification results, and can minimize the error in the classification process.

Disadvantages:

1. KNN is a lazy learning method based on instance learning. It stores all training samples first. When it is to be classified, it is temporarily calculated. For high-dimensional samples or large sample sets, the time and space complexity is high, and the time cost is O(mn).

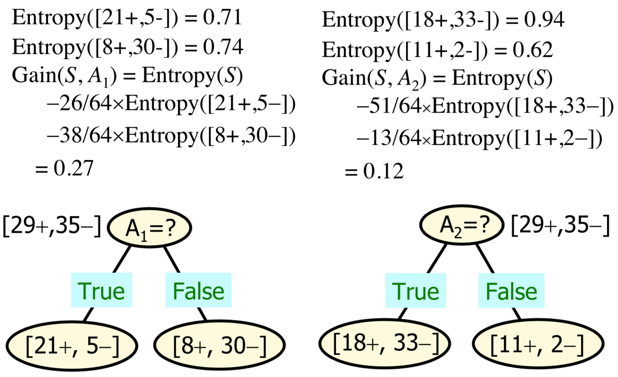
2. The strong capacity dependence of the sample library has a great limitation on the practical application of the KNN algorithm: if the uniform feature space conditions required by the KNN algorithm cannot be satisfied, the error of recognition is large.

3.KNN often assigns the same weight to different features, and the features of the different weights will mislead the classification process.

Reference：http://kesmlcv.blogspot.com/2013/08/knn.html

1.3

a.



Entropy([21+,5-]) = 0.71

Entropy([8+,30-]) = 0.74

$\text{Gain}(S, A_1) = \text{Entropy}(S)$
$-26/64\times\text{Entropy}([21+,5-])$
$-38/64\times\text{Entropy}([8+,30-])$
$= 0.27$

Entropy([18+,33-]) = 0.94

Entropy([11+,2-]) = 0.62

$\text{Gain}(S, A_2) = \text{Entropy}(S)$
$-51/64\times\text{Entropy}([18+,33-])$
$-13/64\times\text{Entropy}([11+,2-])$
$= 0.12$

Considering the content of the class,if we try to use the average entropy ,we may get the following results.

Gain(S,A1) = Average(S) = 0.725

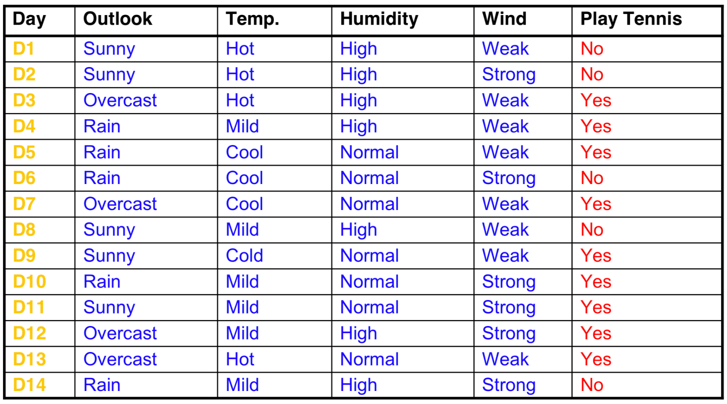
Gain(S,A2) = Average(S) = 0.78

It is obviously not what we want. Cause different categories should be given different weights.

b.

When the features to be considered are very special, the categories we have received are particularly large, but the number of samples for each category is particularly small.

For example:



Day	Outlook	Temp.	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Let's take a look at an extreme situation：If we treat the ID as a attribute that can be used reasonably.We may get 14 categories and there is only one sample in each category.We will get the same gain as the higher entropy.But the ID actually is meaningless.