Homework 3

Kiran Kumar Donthula

02/06/14

The two approaches used for feature selection are

1) Signal to Noise ratio method:

In the given iris data set, we have three classes. Iris Setosa was considered as positive and remaining two were considered as negative classes. The signal to noise ratio values were calculated for all the four attributes. Now, the Iris Versicolor was taken as positive with two remaining classes as negative. The same procedure was followed with the Iris Virginica. The average of S2N values for each attribute was calculated.

Case 1: Iris-Setosa is taken positive and other two classes as negative

		sepal len	sepal width	petal len	petal width
		A1	A2	A3	A4
positive					
	mean	5.006	3.418	1.464	0.244
	stdev	0.35249	0.381024398	0.173511	0.1072095
negative	mean	6.262	2.872	4.906	1.676
	stdev	0.662834	0.332751006	0.825578	0.4247685
S2N		1.237043	0.764946504	3.445136	2.6918406

Case 2: Iris- versicolor is taken positive and other classes as negative

		sepal len	sepal width	petal len	petal width
		A1	A2	A3	A4
positive					
	mean	5.936	2.77	4.26	1.326
	stdev	0.516171	0.313798	0.469911	0.197753
negative	mean	5.797	3.196	3.508	1.135
	stdev	0.945319	0.416071	2.094229	0.919198
S2N		0.095108	0.583666	0.293276	0.171001

Case 3: Iris- verginica is taken positive and other classes as negative

		sepal len	sepal width	petal len	petal width
		A1	A2	A3	A4
positive					
	mean	6.588	2.974	5.552	2.026
	stdev	0.63588	0.322497	0.551895	0.27465
negative	mean	5.471	3.094	2.862	0.785
	stdev	0.641698	0.476057	1.448565	0.566288
S2N		0.874311	0.150272	1.344691	1.475733

Average of S2N values

	sepal len	sepal width	petal len	petal width
Case 1	1.23704339	0.764947	3.445136	2.691841
Case 2	0.09510843	0.583666	0.293276	0.171001
Case 3	0.87431065	0.150272	1.344691	1.475733
Average	0.73548749	0.499628	1.694368	1.446192

Result: The signal to noise ratio values are higher for petal length and petal width. So, they can be selected as best predictive attributes.

2) Relief Method:

In this method, one of the object is randomly selected. The near hit was calculated from the same class where random pick belongs. The near miss was calculated from the remaining two classes and then the weight was updated using the formula. The near hit and miss were selected using the distance between two coordinates formula and compared which object is closer to the random pick compared to other objects.

	sepal length	sepal width	petal length	petal width		
	A1	A2	A3	A4		
1	5.4	3.4	1.7	0.2	Iris-setosa	**near hit**
2	5.4	3.4	1.5	0.4	Iris-setosa	**random pick1**
3	5.4	3	4.5	1.5	Iris-versicolor	**near miss**
	3.6	2.4	5.9	2.4	**RANGE**	

$W_1 = 0 - \left(\frac{5.4-5.4}{3.6}\right)^2 + \left(\frac{5.4-5.4}{3.6}\right)^2 = 0$

$W_2 = 0 - \left(\frac{3.4-3.4}{2.4}\right)^2 + \left(\frac{3.4-3}{2.4}\right)^2 = 0.02778$

$W_3 = 0 - \left(\frac{1.5-1.7}{5.9}\right)^2 + \left(\frac{1.5-4.5}{5.9}\right)^2 = 0.2574$

$W_4 = 0 - \left(\frac{0.4-0.2}{2.4}\right)^2 + \left(\frac{0.4-1.5}{2.4}\right)^2 = 0.2031$

The following table obtained after we randomly pick second object and doing same calculations.

	sepal length	sepal width	petal leng	petal width		
	x	y	z	a		
1	5.7	2.8	4.5	1.3	Iris-versicolor	**near hit**
2	5.7	2.6	3.5	1	Iris-versicolor	**random pick1**
3	5.7	2.5	5	2	Iris-virginica	**near miss**
	3.6	2.4	5.9	2.4	RANGE	

	w1	w2	w3	w4
	0	0.022569	0.293307	0.361111111
average	0	0.011285	0.146653	0.180555556

Result : If we select our threshold value to be 0.1, then petal length and petal width can be considered as best predictive attributes.

Final Conclusion: Both the results obtained from two different methods agree each other.