COMS 7350 DataMining
Professor Cen Li
Homework 3

Xin Yang

February 5, 2014

1 Pearson Correlation

Here, we can use function CORREL(X,Y) in Excel to calculate Pearson correlation between the attribute and class.

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \tag{1}$$

However, three are three classes which are not numerical values. So we can set attribute "Iris Setosa" to 1, "Iris Veriscolour" to 2, "Iris Virginica" to 3. Then, we can use these numerical values to calculate Pearson correlation. Based on the correlation values, we can choose the best two features.

PearsonCorrelation	$\rho(A1,Y)$ = 0.782561232	$\rho(A2,Y)$ = -0.4194462
PearsonCorrelation	$\rho(A3,Y)$ = 0.949042545	$\rho(A4,Y)$ = 0.956463824

Here, A1,A2,A3,A4 represents the four attributes,and Y is the class labels. From the results, we can see the fourth attribute is the best one, and third attribute is better than the another two attributes.

2 Signal to Noise

We can use the following formula to calculate Signal to Noise, and then choose the best two attributes.

$$S2N = \frac{|(\mu^+ - \mu^-)|}{\sigma^+ + \sigma^-} \tag{2}$$

Because there are 3 classes, so we have to compare every each pair alternatively. Before this, we have to calculate mean and standard deviation for each class.

- Calculate the S2N between class 1 and class 2.

$\mu1$	5.006	3.418	1.464	0.244
$\mu2$	5.936	2.77	4.26	1.326
$\sigma1$	0.352	0.381	0.173	0.107
$\sigma2$	0.516	0.314	0.4699	0.198
$S2N_{12}$	1.237	0.7649	3.4415	2.6918

- Calculate the S2N between class 1 and class 3.

$\mu1$	5.006	3.418	1.464	0.244
$\mu3$	6.588	2.974	5.552	2.026
$\sigma1$	0.352	0.381	0.173	0.107
$\sigma3$	0.636	0.322	0.551	0.275
$S2N_{13}$	0.8743	0.1503	1.3447	1.4757

- Calculate the S2N between class 2 and class 3.

$\mu3$	6.588	2.974	5.552	2.026
$\mu2$	5.936	2.77	4.26	1.326
$\sigma3$	0.636	0.322	0.551	0.275
$\sigma2$	0.516	0.314	0.4699	0.198
$S2N_{23}$	1.237	0.7649	3.4415	2.6918

- Calculate the average of the above three results.

$S2N_{ave}$	1.0791	0.6281	3.7485	3.2321

Based on the results, we can see that attribute 3 and attribute 4 are the best two attributes, which is agree with Pearson Correlation.