Feature selection with univariate methods

1) Data

Obj	A_1	A_2	A_3	…	class
1	30	28	.		Pos
2	24	16	.		Pos
3	20	15	.		Neg
4	28	17	.		Pos
5	10	19	.		Neg
6	20	20	.		Neg
7	16	16	.		Neg
8	34	15	.		Pos
.	.	.	.		.
.	.	.	.		.

Which attribute is better for predicting the class label?

A_1 or A_2 ?

- signal to noise ratio
- t-test
- correlation
- chi-square test
- Mutual Information

1. signal to noise ratio method:

First Step : divide up the data by clauses:

Obj	A_1	A_2	Class
1	30	28	positive
2	24	16	
4	28	17	
8	34	15	

$\mu_1 = 29$ $\mu_2 = 19$

$\sigma_1 = 4.16$ $\sigma_2 = 6.05$

Obj	A_1	A_2	Class
3	20	15	negative
5	10	19	
6	20	20	
7	16	16	

$\mu_1 = 16.5$ $\mu_2 = 17.5$

$\sigma_1 = 4.72$ $\sigma_2 = 2.38$

$$\frac{|\mu^+ - \mu^-|}{\sigma^+ + \sigma^-}$$

A_1 : $\frac{|29 - 16.5|}{4.16 + 4.72} = \frac{12.5}{8.88} = 1.4$ ← larger signal to noise ratio

A_2: $\frac{|19 - 17.5|}{6.05 + 2.38} = 0.17$

Attribute A_1 is better

t-test method:

Degree of freedom

$m^+ + m^- - 2 = 4 + 4 - 2 = 6$

p = 0.05

null hypothesis: A_{1i}positive distribution ≡ A_i negative distribution

t-test critical value: 2.447 ← from table

First Step : divide up the data by classes:

Obj	A_1	A_2	Class
1	30	28	positive
2	24	16	
4	28	17	
8	34	15	

Obj	A_1	A_2	Class
3	20	15	negative
5	10	19	
6	20	20	
7	16	16	

Then, compare the distribution of attribute A1 for objects having class "positive" to the distribution of the attribute A1 for the objects having class "negative".
The null hypothesis is that the two distributions are the same.

use excel functions:

For attribute A1:
ttest(A2:A5, B2:B5, 2, 3)

{first paramenter: "positive" class object attribute values are in A2:A5
second parameter: "negative" class object attribute values are in B2:B5
fourth parameter: 2 – two tailed t test
third parameter: 3 - assuming equal variance
}

ttest result for A1 is 0.0076 < critical value 2.447, accept null hypothesis

Repeat the above for attribute A2:
ttest result for A2 is 0.669 < 2.447, accept

In this case, both A1 and A2 are not good predictors for the class. Comparably speaking, 0.669 > 0.0076, A2 is slightly better attribute than A1

3. <u>correlation method:</u>

class
Use positive ← 1
negative ← -1

excel function
correl(A1:A8, B1:B8)
A1:A8 - attribute A1's values, B1:B8 – class values
correlation (A_1, class) = 0.44

correlation (A_2, class) = 0.123

A_1 correlates more with the class, therefore is more predictive, i.e., better

Chi square method:

- not readily applicable here since Attribute values are numeric
- what we can do is to discretize it, i.e., define bins of values and assign the numeric values into the corresponding bins
- have practiced in hw #2 for testing whether to perform decision tree pruning

Mutual Information method:

- N/A due to numeric value
- can be computed if needed based on join probability distribution functions
- have practiced in hw #1 w/ decision tree