Datamining homework 3 by Xiao Liang

Approach 1: Pearson Correlation

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

Approach 2: Mutual Information

$$P(x,y) = \text{joint probability of } (x,y), \quad x \in \{1,\ldots,r\} \text{ and } y \in \{1,\ldots,s\}$$
$$P(x) = \sum_y P(x,y) = \text{marginal probability of } x$$
$$P(y) = \sum_x P(x,y) = \text{marginal probability of } y$$

$$MI(X,Y) = \sum_{xy} P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)}$$

Features 3 and 4 should be used as results of both approaches.

The .dat file was transferred into .csv file with only numerical numbers.

Matlab codes for Pearson Correlation:

```
clear
clc
M = csvread('iris.csv');
Y=zeros(150,1);
Y(1:50)=-1;
Y(51:100)=0;
Y(101:150)=1;
coeff=zeros(4,1);
for i=1:4
X=M(:,i);
C = cov(X,Y);
coeff(i) = C(1,2) / sqrt(C(1,1) * C(2,2));
end
coeff
```

Results:

coeff =

0.7826

-0.4194

0.9490

0.9565

Mutual Information:

```
clc
clear all
M = csvread('iris.csv');
mi=zeros(4,1);
for i=1:4
X=M(:,i);
mu=mean(X);
xs=0.0001;xl=0.0001;
for j=1:50
    if X(j)<=mu
        xs=xs+1;
    else
        xl=xl+1;
    end
end
ps1=xs/150;pl1=xl/150;
xs=0.0001;xl=0.0001;
for j=51:100
    if X(j)<=mu
        xs=xs+1;
    else
        xl=xl+1;
    end
end
ps2=xs/150;pl2=xl/150;
xs=0.0001;xl=0.0001;
for j=101:150
    if X(j)<=mu
        xs=xs+1;
    else
        xl=xl+1;
    end
end
ps3=xs/150;pl3=xl/150;
ps=ps1+ps2+ps3;
pl=pl1+pl2+pl3;
p1=1/3;p2=1/3;p3=1/3;
mi(i)=ps1*log(ps1/ps/p1)/log(2)+ps2*log(ps2/ps/p2)/log(2)+ps3*log(ps3/ps/p3)/
log(2)...
+pl1*log(pl1/pl/p1)/log(2)+pl2*log(pl2/pl/p2)/log(2)+pl3*log(pl3/pl/p3)/log(2
);
end
mi
```

Results:

mi =

0.4874
0.2606
0.7633
0.7303