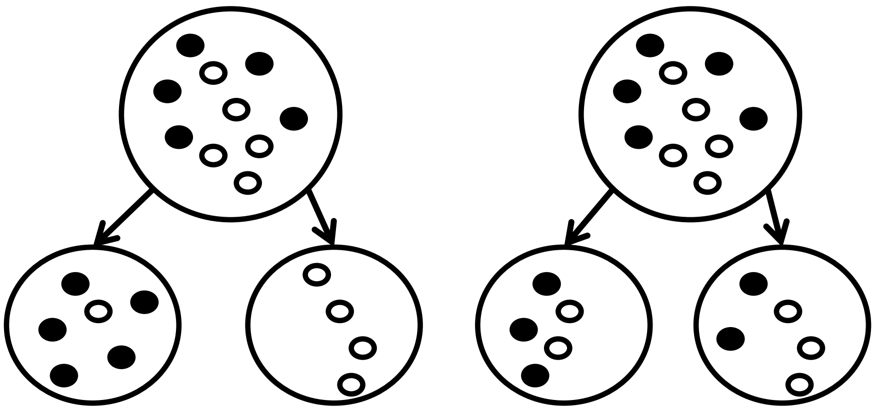
CSCI 485. Introduction to data mining. Spring 2012.

Midterm exam (Solutions)

1. Consider two possible splits of training records during the decision tree induction.



A B

Figure 1. Possible splits of training records according to attributes A and B. Black and white colors represent class labels.

a. What is the entropy of each split? Which split produces more pure nodes? (2 points):

Entropy(A)=6/10*(-5/6*log(5/6,2)-1/6*log(1/6,2))+4/10*(-4/4*log(4/4,2))=0.39

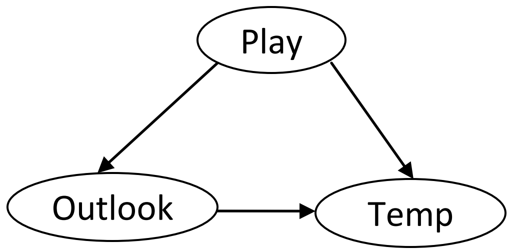
Entropy(B)=5/10*(-3/5*log(3/5,2)-2/5*log(2/5,2))+5/10*(-2/5*log(2/5,2)-3/5*log(3/5,2))=0.97

Split A produces more pure nodes.

b. What is the information gain of each split? (1 point)
Entropy(All)=-5/10*log(5/10,2)-5/10*log(5/10,2)=1

Info gain (A)=1-0.39=0.61
Info gain (B)=1-0.97=0.03

2. The following figure is a Bayesian belief network for the weather data summarized in Table 1. (Assume that all attributes are binary).



Outlook	**Temp**	**Play**
Sunny	Hot	Yes
Rainy	Cold	No
Sunny	Hot	Yes
Rainy	Cold	No
Rainy	Hot	No
Rainy	Cold	No
Rainy	Cold	No
Rainy	Hot	Yes

Table 1. Dataset for Question 2

a. Draw the probability table for each node in the network (3 points).
b. Apply Laplace correction (to all values) (1 point)
c. What is the probability of Play=Yes on a hot day? (2 points)

$P(y|h,S)=P(y)*P(S|y)*P(h|y,S)*\alpha$

$=P(y)*P(s|y)*P(h|y,s)+P(y)*P(\neg s|y)*P(h|y,\neg s) *\alpha$

$= 4/10*3/5*3/4+4/10*2/5*2/3*\alpha = 0.29*\alpha$

$P(\neg y|h,S)=P(\neg y)*P(S|\neg y)*P(h|\neg y,S)*\alpha$

$=P(\neg y)*P(s|\neg y)*P(h|\neg y,s)+P(\neg y)*P(\neg s|\neg y)*P(h|\neg y,\neg s) *\alpha$

$= 6/10*1/7*1/2+6/10*6/7*2/7*\alpha = 0.19*\alpha$

$0.19 \alpha +0.29 \alpha =1.0$

$\alpha =1.0/0.48=2.08$

$P(y|E)=2.08*0.29=60.4\%$

$P(\neg y|E)=2.08*0.19=39.6\%$

Probability of play on a hot day is 60.4%

Conditional probability tables:

Play	
y	3/8
¬y	5/8

Play	
y	4/10
¬y	6/10

	Outlook	
Play	s	¬s
y	2/3	1/3
¬y	0/5	5/5

	Outlook	
Play	s	¬s
y	3/5	2/5
¬y	1/7	6/7

		Temp	
Play	Outlook	h	¬h
y	s	2/2	0/2
y	¬s	1/1	0/1
¬y	s	0/0	0/0
¬y	¬s	1/5	4/5

		Temp	
Play	Outlook	h	¬h
y	s	3/4	1/4
y	¬s	2/3	1/3
¬y	s	1/2	1/2
¬y	¬s	2/7	6/7

3. The fraction of undergraduate students who smoke is 15% and the fraction of graduate students who smoke is 25%. One-fifth of the University students are graduate students and the rest are undergraduates.

a. What is the probability that a student who smokes is a graduate student? (2 points)

$P(g|s)= \alpha*P(s|g)*P(g)= \alpha*0.25*0.20=0.05\ \alpha$

$P(\neg g|s)= \alpha*P(s|\neg g)*P(\neg g)= \alpha*0.15*0.80=0.12\ \alpha$

$0.05\ \alpha+0.12\ \alpha=1.00$

$\alpha=5.88$

$P(g|s)=5.88*0.05=29.4\%$

b. Does this probability change if we manage to reduce the number of smokers to 3% and 5% in undergrads and grads respectively? What is the value of this new probability (that a student who smokes is a graduate student)? (2 points)

The value is the same – 29.4%

4. We have compared two classifiers through cross-validation on 5 different datasets (folds).

The success rates are:

Dataset	Classifier A	Classifier B	Difference
1	89.4	89.8	-0.4
2	90.2	90.6	-0.4
3	88.7	88.2	0.5
4	90.3	90.9	-0.6
5	91.2	91.7	-0.5

Which classifier is significantly better at significance level 10%? (2 points)

m_d=(-0.4-0.4+0.5-0.6-0.5)/5=-0.28

s^2=(2*(-0.4+0.28)^2+(-0.5+0.28)^2+(0.5+0.28)^2+(-0.6+0.28)^2)/4 =0.197

σ=sqrt(0.197)=0.44

H0: no difference 0 ± t * σ /sqrt(N)

0.44/sqrt(5)*2.353=0.46

The difference (-0.28) is not significant at significance level 10%

T-table (pre-computed values of Student's distribution)

One Sided	**90%**	**95%**	**97.5%**	**99%**	**99.5%**	**99.75%**	**99.9%**	**99.95%**
Two Sided	**80%**	**90%**	**95%**	**98%**	**99%**	**99.5%**	**99.8%**	**99.9%**
1	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869