GITA305 Data Mining
Exam 1, April 23, 2009

Name: **Student No.:**

1. (40 points) Consider the following training dataset for predicting whether a mushroom is edible or not based on its shape, color and odor:

Shape	*Color*	*Odor*	*Edible*
C	B	1	Yes
D	B	1	Yes
D	W	1	Yes
D	W	2	Yes
C	B	2	Yes
D	B	2	No
D	G	2	No
C	U	2	No
C	B	3	No
C	W	3	No
D	W	3	No

(a) Draw the full decision tree that would be learned by the ID3 algorithm for this data. You need to show the entropy calculations **for the top level only**. (10 pts.)

(b) Generate the conditional probability tables for each attribute given the class (i.e. $P(Shape|Edible)$, $P(Color|Edible)$, and $P(Odor|Edible)$) to be used in a naive Bayes classifier. (10 pts.)

(c) What is the training error for each algorithm? (10 pts.)

(d) Suppose we have the following validation set:

Shape	*Color*	*Odor*	*Edible*
C	B	2	No
D	B	2	No
C	W	2	Yes

What is the validation set error for each algorithm? (10 pts.)

2. Bayesian Learning (20 points)
   After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious disease and that the test is 99% accurate (i.e. the probability of testing positive when you do have the disease is 0.99, as is the probability of testing negative when you don't have the disease). The good news is that this is a rare disease, striking only 1 in 10,000 people of your age. Why is it good news that the disease is rare? What are the chances that you actually have the disease?

3. Give a brief but precise answer for each of the following questions ($2 \times 10 = 20$ points):

   (a) What is the measure the ID3 algorithm employs? State the rationale behind it.

   (b) State the objective that the ID3 algorithm is trying to optimize.

   (c) Does ID3 algorithm guarantee to find the global optimum hypothesis (in terms of the objective mentioned above)?

   (d) Explain the relationship between Occam's razor, generalization, and overfitting.

   (e) Explain the relationship between training, test, and cross-validation sets of data.

   (f) Explain the relationship between maximum likelihood and maximum a posteriori hypothesis.

   (g) Give one advantage of ID3 over Bayesian learning.

   (h) Give one advantage of Bayesian learning over ID3.

   (i) Explain the relationship between machine learning and data mining.

   (j) Explain the relationship between naive Bayes and ordinary Bayesian learning.

$$\log_2 \frac{3}{5} = \frac{\log_{10} \frac{3}{5}}{\log_{10} 2}$$

GITA305 Data Mining

Exam 1, April 23, 2009

ch.3 p.33

Name: **Student No.:**

1. (40 points) Consider the following training dataset for predicting whether a mushroom is edible or not based on its shape, color and odor:

class는 2개

Shape(C, D)

Color (B, W, G, U)

Odor (1, 2, 3)

Shape	*Color*	냄새 *Odor*	먹을수 있는 *Edible*
C	B	1	Yes
D	B	1	Yes
D	W	1	Yes
D	W	2	Yes
C	B	2	Yes
D	B	2	No
D	G	2	No
C	U	2	No
C	B	3	No
C	W	3	No
D	W	3	No

5

6

entropy 계산하기

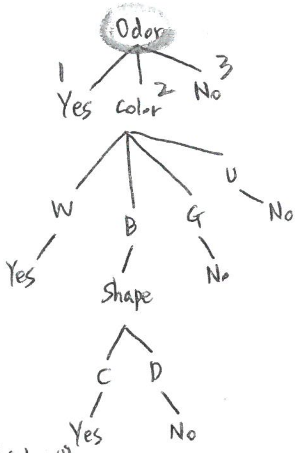
(a) Draw the full decision tree that would be learned by the ID3 algorithm for this data. You need to show the entropy calculations **for the top level only**. (10 pts.)

최상위 level의 entropy를 계산해서 나타내야 한다.

$$\hat{H}(X) = -\frac{5}{11}\log_2\frac{5}{11} - \frac{6}{11}\log_2\frac{6}{11}$$

$$\hat{H}(X|\text{Shape}=C) = -\underbrace{\frac{2}{5}\log_2\frac{2}{5}}_{-0.53} - \underbrace{\frac{3}{5}\log_2\frac{3}{5}}_{-0.44} = \boxed{0.971}$$

$$\hat{H}(X|\text{shape}=D) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$



$$\hat{H}(X|\text{shape}) = \frac{5}{11}\hat{H}(X|\text{shape}=C) + \frac{6}{11}\hat{H}(X|\text{shape}=D) = 0.986 \text{ bits}$$

$$\hat{H}(X|\text{Color}) = \frac{5}{11}\underbrace{\hat{H}(X|\text{Color}=B)}_{0.97} + \frac{4}{11}\underbrace{\hat{H}(X|\text{color}=W)}_{1} + \frac{1}{11}\underbrace{\hat{H}(X|\text{color}=G)}_{0} + \frac{1}{11}\underbrace{\hat{H}(X|\text{Color}=U)}_{0}$$

$$= 0.805 \text{ bits}$$

$$\hat{H}(X|\text{odor}) = \frac{3}{11}\underbrace{\hat{H}(X|\text{odor}=1)}_{0} + \frac{5}{11}\underbrace{\hat{H}(X|\text{odor}=2)}_{0.97} + \frac{3}{11}\underbrace{\hat{H}(X|\text{odor}=3)}_{0}$$

$$= \frac{5}{11} \times 0.97 = 0.441 \text{ bits}$$

v What is entropy $H(\text{Edible}|\text{Odor}=1 \text{ or Odor}=3?)$

$H(\text{Edible}|\text{Odor}=1 \text{ or Odor}=3) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1$

? (d) Suppose we have the following validation set:

	Shape	*Color*	*Odor*	*Edible*
✗	C	B	2	No
	D	B	2	No
✗	C	W	2	Yes

Yes

What is the validation set error for each algorithm? (10 pts.)

training set error : 0

validation set error : 1

Validation data set 이니까 Validation error 만 측정 가능.

Validation dataset 으로 결과를 도출하면 Validation의 결과지 training의 결과는 아님.

간단하지만 정확한 대답

3. Give a brief but precise answer for each of the following questions ($2 \times 10 = 20$ points):

그 뒤에 이론적 근거를 명시하세요.

(a) What is the measure the ID3 algorithm employs? State the rationale behind it.

측정?

(b) State the objective that the ID3 algorithm is trying to optimize.

ID3 알고리즘은 최적화하려고 하는 목적을 쓰세요(?)

글로벌 최적의 가설을 찾기 위해 ID3가 보장(?) 되어 있는가? 상기 언급된 목적의 관점에서

(c) Does ID3 algorithm guarantee to find the global optimum hypothesis (in terms of the objective mentioned above)?

V (d) Explain the relationship between Occam's razor, generalization, and overfitting.

V (e) Explain the relationship between training, test, and cross-validation sets of data.

최대 우도 최대 사후확률

(f) Explain the relationship between maximum likelihood and maximum a posteriori hypothesis.

(g) Give one advantage of ID3 over Bayesian learning.

(h) Give one advantage of Bayesian learning over ID3.

V (i) Explain the relationship between machine learning and data mining.

V (j) Explain the relationship between naive Bayes and ordinary Bayesian learning.

GITA305 Data Mining
Jihoon Yang
Exam 2, June 18, 2009

Name: **Student No.:**

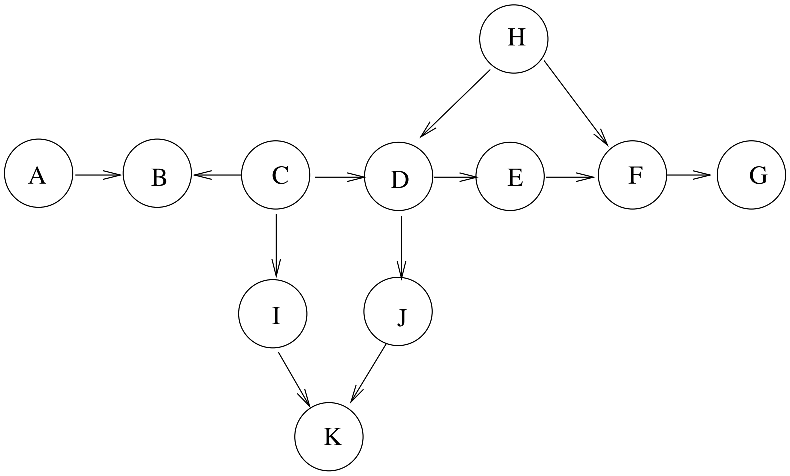
1. Perceptron Learning (20 points)
   Suppose we have the following 5 1-dimensional data points:

Input	Class
1	1
2	1
4	-1
5	-1
6	1

   Initialize all the weights (including the bias) to the zero-vector. Then give their values after each of the **first two cycles**. Assume learning rate of 0.5 and **incremental (or per-sample)** updates. Show the decision boundary with the final weights.

2. MLP with BP Algorithm (20 points)
   Consider a two-layer feedforward ANN with two inputs a and b, one hidden unit c, and one output unit d. This network has five weights $(w_{ca}, w_{cb}, w_{c0}, w_{dc}, w_{d0})$, where w_{x0} represents the threshold weight for unit x. The activation function of c and d are sigmoid and linear, respectively. Draw the neural network, and derive the weight update equations for all the weights. Then initialize these weights to the values (.1,.1,.1,.1,.1), and give their values after each of the **first two** training iterations of the error backpropagation algorithm. Assume learning rate $\eta = .3$, momentum $\alpha = .9$, **incremental** weight updates, and the following training examples: $[a, b, d] = [101]$ and $[010]$.

3. d-separation in Bayesian Networks (20 points)
Using the Bayes network given below, for each of the following statements indicate whether it is true or false.



(a) $I(B, D, J)$

(b) $I(C, G, H)$

(c) $I(I, J, H)$

(d) $I(A, \{B, E\}, G)$

(e) $I(C, \{\}, H)$

4. Short Questions (20 points)

(a) (2 pts.) True or False: Perceptron algorithm is guaranteed to find the separating decision boundaries for some linearly nonseparable problems.

(b) (2 pts.) True or False: Perceptron algorithm is guaranteed to find the best decision boundaries for linearly nonseparable problems.

(c) (2 pts.) True or False: Generalized delta rule in a multilayer network is guaranteed to find weights that correspond to the global minimum of the error curve.

(d) (2 pts.) True or False: Bayesian network learning is preferred to naive Bayes learning when all the input features are independent given the class.

(e) (2 pts.) True or False: K-means clustering algorithm is guaranteed to find the best clustering.

(f) (2 pts.) True or False: HAC algorithm is guaranteed to find the best clustering.

(g) (2 pts.) True or False: Lazy learning is preferred to eager learning when the generalization speed is critical.

(h) (3 pts.) Briefly describe the difference between a maximum likelihood hypothesis and a maximum a posteriori hypothesis.

(i) (3 pts.) Explain what learning means in ANNs, and give common algorithms/methods.

GITA305 Data Mining

Exam 1, April 28, 2011

Name: **Student No.:**

1. (20 points) An experiment consists of tossing a single die and observing the number of dots that show on the upper face. Events A, B, and C are defined as follows:

A: Observe a number less than 4 $A < 4 \rightarrow 1, 2, 3$
B: Observe a number less than or equal to 2 $B \leq 2 \rightarrow 1, 2$
C: Observe a number greater than 3 $C > 3 \rightarrow 4, 5, 6$

(a) Find the probabilities associated with these compound events:
$A, B, C, A|B, B|A, A \cap B \cap C, A \cap B, A \cup C$

$A = \frac{1}{2}$. $B = \frac{1}{3}$ $C = \frac{1}{2}$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/3}{1/3} = 1$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{1/3}{1/2} = \frac{2}{3}$$

$A \cap B \cap C = 0$

$A \cap B = \frac{1}{3}$

$A \cup C = 1$

(b) Are events A and B mutually exclusive? If $P(B) = 2/3$, are events A and B independent?

A와 B는 상호배반 X $\rightarrow (1, 2)$

$P(B) = \frac{2}{3}$ 이면, $P(A) = \frac{3}{3}$ 이고 $P(A \wedge B) = \frac{2}{3}$ 이므로

$P(A) \cdot P(B) = 1 \times \frac{2}{3} = \frac{2}{3} = P(A \wedge B)$.

$\therefore$ A와 B는 independent

(Data shown again)

Shape	*Color*	*Odor*	*Edible*
C	B	1	Yes
D	B	1	Yes
D	W	1	Yes
D	W	2	Yes
C	B	2	Yes
D	B	2	No
D	G	2	No
C	U	2	No
C	B	3	No
C	W	3	No
D	W	3	No

(b) Generate the conditional probability tables for each attribute given the class (i.e. $P(Shape|Edible)$, $P(Color|Edible)$, and $P(Odor|Edible)$) to be used in a naive Bayes classifier. (10 pts.)

(c) What is the training error for each algorithm? (10 pts.)

GITA305 Data Mining
Exam 2, June 16, 2011

Name: **Student No.:**

1. Perceptron (20 points)
   Suppose we have the following 5 1-dimensional data points:

Input	Class
1	1
2	1
4	-1
5	-1
6	1

   Initialize all the weights (including the bias) to the zero-vector. Then give their values after each of the **first two cycles**. Assume learning rate of 0.5 and **incremental (or per-sample)** updates. Show the decision boundary with the final weights.

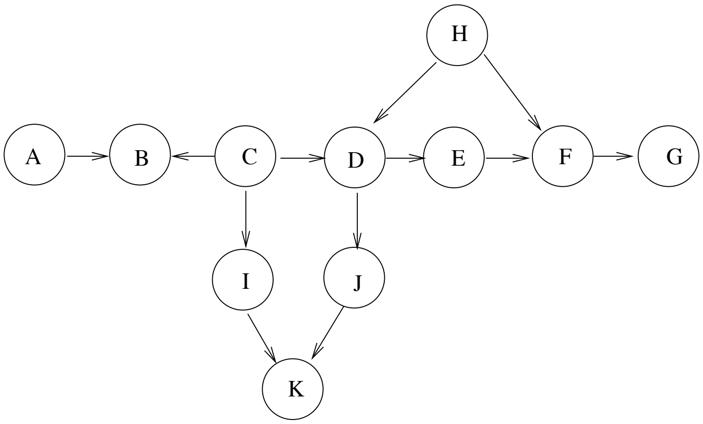
2. MLP with BP Algorithm (20 points)
Consider a two-layer feedforward ANN with two inputs a and b, one hidden unit c, and one output unit d. This network has five weights $(w_{ca}, w_{cb}, w_{c0}, w_{dc}, w_{d0})$, where w_{x0} represents the threshold weight for unit x. The activation function of c and d are sigmoid and linear, respectively. Draw the neural network, and derive the weight update equations for all the weights.

3. Nearest Neighbor Algorithm (15 points)
Suppose we have the following patterns each of which consists of two inputs and an output of an unknown function:

$$([1, 1]; 5), ([2, 2]; 4), ([3, 3]; 3), ([4, 4]; 2), ([5, 5]; 1)$$

Using a 2-nearest neighbor algorithm (and the Euclidean distance metric), compute the output of a test pattern which has [3.1, 3.1] as inputs.

4. d-separation ($5 \times 5 = 25$ points)
Using the given Bayes network, for each of the following statements indicate whether it is true or false.



(a) $dsep(B, D, J)$

(b) $dsep(C, G, H)$

(c) $dsep(I, J, H)$

(d) $dsep(A, \{B, E\}, G)$

(e) $dsep(C, \{\}, H)$

GITA305 Data Mining
Exam 2, December 17, 2014

Name: **Student No.:**

1. Perceptrons (10 points)
   Suppose we have the following 5 1-dimensional data points:

Input	1	2	4	5	6
Class	1	1	-1	-1	1

   Initialize all the weights (including the bias) to the zero-vector. Then give their values after considering each data point for the first cycle. Assume learning rate of 0.5 and incremental updates. Show the decision boundary with the final weights.

2. MLP with BP Algorithm (10 points)
   Consider a two-layer feedforward ANN with two inputs a and b, one hidden unit c, and one output unit d. This network has five weights $(w_{ca}, w_{cb}, w_{c0}, w_{dc}, w_{d0})$, where w_{x0} represents the threshold weight for unit x. The activation function of c and d are sigmoid and linear, respectively. Draw the neural network, and derive the weight update equations for all the weights.

3. Nearest Neighbor Algorithm (10 points)
Suppose we have the following patterns each of which consists of two inputs and an output of an unknown function:

$$([1, 1]; 5), ([2, 2]; 4), ([3, 3]; 3), ([4, 4]; 2), ([5, 5]; 1)$$

Using a 2-nearest neighbor algorithm (and the Euclidean distance metric), compute the output of a test pattern which has [2.1, 2.1] as inputs.

4. SOM (10 points)
Suppose three output neurons (O_1, O_2, O_3) are used in a SOM with two input neurons for the following 8 data points (with (x, y) representing the location):

$$A_1(2, 10), A_2(2, 5), A_3(8, 4), B_1(5, 8), B_2(7, 5), B_3(6, 4), C_1(1, 2), C_2(4, 9)$$

The (static) neighborhood is defined as: $neighborhood(O_i) = \{O_{i-1}, O_i, O_{i+1}\}$ (where invalid indices are ignored). Assuming the weights of O_1, O_2, O_3 are initialized with the values of A_1, B_1, C_1 respectively, show how they are changed in the first iteration with A_2. Assume the learning rate is 0.5.

1. (100 points)

A database has five transactions. Let *min sup* = 60% and *min conf* = 75%.

TID	*Items-bought*
T100	M, O, N, N, K, E, Y, Y
T200	D, D, O, N, K, E, Y
T300	M, M, A, K, E, E
T400	M, U, C, C, Y, C, E, O
T500	C, O, O, K, I, I, E

(a) Find all frequent itemsets using Apriori method.

Database is scanned once to generate frequent 1-itemsets. To do this, I use absolute support, where duplicate values are counted only once per TID. The total number of TID is 5, so minimum support of 60% is equivalent to 3/5. Thus itemsets with 1 or 2 support counts are eliminated.

Table 1a. 1-itemset results, raw

Itemset	*support*	*support percentage*
{A}	1	20%
{C}	2	40%
{D}	1	20%
{E}	5	100%
{I}	1	20%
{K}	4	80%
{M}	3	60%
{N}	2	40%
{O}	4	80%
{U}	1	20%
{Y}	3	60%

Table 1b. 1-itemset results, consolidated

Itemset	*support*	*support percentage*
{E}	5	100%
{K}	4	80%
{M}	3	60%
{O}	4	80%
{Y}	3	60%

Now, database is scanned second time to generate frequent 2-itemsets. The possible combinations are 5!/(3!2!) = 10. Using absolute support, each combination is counted per TID, and combinations that are below support value of 3 are eliminated.

Table 2a. 2-itemset results, raw

Itemset	*support*	*support percentage*
{E, K}	4	80%
{E, M}	3	60%
{E, O}	4	80%
{E, Y}	3	60%
{K, M}	2	40%
{K, O}	3	60%
{K Y}	2	40%
{M, O}	2	40%
{M, Y}	2	40%
{O, Y}	3	60%

Table 2a. 2-itemset results, consolidated

Itemset	*support*	*support percentage*
{E, K}	4	80%
{E, M}	3	60%
{E, O}	4	80%
{E, Y}	3	60%
{K, O}	3	60%
{O, Y}	3	60%

I proceed to scan the database again to generate frequent 3-itemsets. Sets {E, K}, {K, O}, {E, O} make {E, K, O} possible. Likewise, {E, O}, {E, Y}, {O, Y} make {E, O, Y}.

Table 3a. 3-itemset results

Itemset	*support*	*support percentage*
{E, K, O}	3	60%
{E, O, Y}	3	60%

Frequent 4-itemsets cannot be generated, because sets {K, O, Y} and {E, K, Y} are missing. So, all frequent itemsets have been found.

(b) List all of the *strong* association rules (with support *s*=60% and confidence *c*=75%) matching the following metarule, where *X* is a variable representing customers, and *item$_i$* denotes variables representing items (e.g., *"A"*, *"B"*, etc.):

buys(*X; item*1) *and buys*(*X; item*2) *) => buys*(*X; item*3) [*s; c*]

The highest itemsets are {E, K, O} and {E, O, Y}. Thus, there can be 2(3!/(1!2!)) = 6 total possible association rules following the metarule of selecting 2 inputs for testing association with 1 output.

Association rules from {E, K, O}:

R1. E ∩ K -> O

confidence = #{E, K, O} / #{E, K} = 3 / 4 = 75%

Therefore, R1 is a strong association rule.

R2. E ∩ O -> K

confidence = #{E, K, O} / #{E, O} = 3 / 4 = 75%

Therefore, R2 is a strong association rule.

R3. K ∩ O -> E

confidence = #{E, K, O} / #{K, O} = 3 / 3 = 100%

Therefore, R3 is a strong association rule.

Association rules from {E, O, Y}:

R4. E ∩ O -> Y

confidence = #{E, O, Y} / #{E, O} = 3 / 4 = 75%

Therefore, R4 is a strong association rule.

R5. E ∩ Y -> O

confidence = #{E, O, Y} / #{E, Y} = 3 / 3 = 100%

Therefore, R5 is a strong association rule.

R6. O ∩ Y -> E

confidence = #{E, O, Y} / #{O, Y} = 3 / 3 = 100%

Therefore, R6 is a strong association rule.

In this case, all 6 association rules are strong, meaning that customers who purchase any of the two products among E, K, O are likely to purchase the remaining one, and customers who purchase two items among E, O, Y are likely to purchase the remaining one.