northumbria UNIVERSITY NEWCASTLE

(Form MOD1): External Examiner Approval of Assessment Tasks

(coursework and examination)

Process:
Assessment tasks/briefs and related information should be carefully checked before[1] being sent to the External Examiner for approval: briefs and papers should, therefore, be marked draft until confirmed by the External Examiner. Further information can be found under Marking Moderation Processes on the Assessment Web Page[2].

Once the checklist below has been completed by the Module Lead, this form should be provided to the external examiner with the assessment information.

Faculty	[ADSS][BL][EE][HLS]	**Department**	**ITM, SeoulTech**
Module Code	**ITM 522**	**Module Title**	**Data Mining**
Module Tutor	**Kyoungok Kim**	**Level**	**Level 5**
Student Year Group / Intake	**2016 / Fall**	**Credit Value**	**10**
Description of Assessment / Component	**EXAM (Mid-term Exam)**	**% of Module Mark**	**25%**
Delivery Location		**SeoulTech**	

Checklist for coursework / module assignments / examination	**YES / NO** *(please specify)*
Alignment to module learning outcomes and assessment task on MD	YES
Free from typing / grammatical errors	YES
Clear information / instructions / rubric provided	YES
Word lengths and penalties included where appropriate (coursework)	YES
Marks allocation for components included (*if relevant*)	YES.
Hand-in date, feedback process and timeframe clearly stated	YES
Marking criteria provided for students (which map to grade descriptors)	YES
Marking guidelines / solutions / answers (*for internal / external moderator*)	YES
Option for alternative assessment considered (*in this case this is required*)	N/A
Any extra documentation (*e.g. for examinations*) clearly labelled	N/A
Adheres to any PSRB requirements	YES
Liaison with placement provider (*where appropriate*)	N/A
Assessment for referrals included or approval date agreed with external examiner	N/A
Consultation with EPWO partners to check for contextual relevance	N/A

Length of EE approval for this assessment *(please circle)*	One delivery only 1 year 2 years 3 years Until end of tenure

Name and contact Details of Moderator	**Date**	**Comments**
Hakyeon Lee hylee@seoultech.ac.kr (signature)	17. 10. 2016.	· Total marks are 105 pts. Is this intended? · Questions are relevant and to the point. · Students may need more time.

[1] For educational partners a Northumbria member of staff should sign to assure NU oversight of assessment task

[2] https://www.northumbria.ac.uk/sd/central/ar/qualitysupport/asspolicies/

E1 Form (to be completed by module tutor for all examination assessments)

Exam Description (Paper 1 etc.)	**Exam size** (Please enter number of students expected to sit exam)	**Does exam need to be same day as Part Time/Distance Learning students?** (Please specify day for exam to be sat if taken by part time students.)	**Is the module taught on a Campus other than Newcastle City Campus?** (If runs as CV, does the exam need to start at 09:30?)	**Campus where module is to be examined.**
5	25	Yes	Yes	SeoulTech

EXAM DETAILS

Boxes 1, 2, 3 and 5 along with the information provided above are essential in order that an exam can be scheduled, if any of these boxes are left blank the exam details may not be entered correctly and errors in scheduling may occur.

1. Duration (the duration must **include reading time, if appropriate**)	**2. Type of Exam**		**3. Accommodation Requirements** Please tick or state as requested.		**4. Are there any constraints for this examination?** **You must also provide a valid reason for this constraint, otherwise it will not be included. Exam constraints should only be applied if essential.**	**5. Does this exam need to run at the same time as another exam?** Please enter module code in the box below.
1.5hrs	☐	Open book	**SD** (single desk)	☐	Constraint requested:	This exam must run at the same time as
	☒	Closed Book	**DD** (double desk)	☒		
	☐	Written	**LB** (locally booked,	☐		(module code)
	☐	Written closed book	If a specific room is required, please state room number**, this must be booked by the Faculty**.		Reason for constraint above:	Do the exams need to take place in the same room?
	☐	Unrestricted				
	☐	Restricted notes			.	Yes / No
	☐	Open Notes				
	☐	Pre-work				
	☐	Lab based exam				
	☐	OMR (optical mark recording)	**Lab** (please state lab number) **this must be booked by the Faculty**			
	☐	Other – please specify				

SEOULTECH
Seoul National University of Science & Technology

Mid-term Exam

Date : 2016.10.31

Code	ITM 522		Title	Data Mining	
Time for Exam	**1.5hr**	**Questions**	**7**	**Weighting**	**25%**

1. (10pts, each 2pts) Write definition (equation) of following terms.

(1) SSE

$\sum_T (x_T - \hat{x}_T)^2$

(2) SSR

$\sum_T (\hat{x}_T - \bar{x})^2$

(3) logit function

$\sigma(t) = \frac{1}{1+e^{-t}}$

(4) chi-square statistic

(5) Mahalanobis distance

$\sqrt{(x_1 - x_2)^T S^{-1} (x_1 - x_2)}$

S is covariance matrix

2. (30pts) Data set with four variables such as api00, acs_k3, meals and full is given and we want to build a regression model on them

[Variable Description]

api00: academic performance of the school

acs_k3: the average class size in kindergarten through 3rd grade

meals: the percentage of students receiving free meals

full: the percentage of teachers who have full teaching credentials

The purpose of regression model is to predict api00 using the other variables.

In data set, there are 313 data points.

[Fitted model]

Variable	DF	Parameter Estimate	Standard Error	t value
Intercept	1	906.73916	28.26505	32.09987
acs_k3	1	-2.68151	1.39399	-1.92362
Meals	1	-3.70242	0.15403	-24.03901
Full	1	0.10861		

SEOULTECH
Seoul National University of Science & Technology

Mid-term Exam

Date : 2016.10.31

Code	ITM 522		Title	Data Mining	
Time for Exam	1.5hr	Questions	7	Weighting	25%

(1) From the fitted model, SST=3,906,597 and SSR=2,634,884. Fill out following table (18pts, each 2pts)

Source	DF	Sum of Squares	Mean Square	F value
Model	[3]	[2,634,884]	[878294.6667]	[211.40747]
Error	[309]	[1271713]	[4115.51605]	
Total	[312]	[3906597]		

(2) If sample variance of variable "full" is 1602.77, what is standard error of estimated coefficient for variable "full"? (5pts)

$\sigma(full)^2 = 1602.77 = \frac{\sum(x_i-\bar{x})^2}{n-1}$ $\sum(x_i-\bar{x})^2 = 500064.24$

$se(full) = \sqrt{\frac{4115.51605}{500064.24}} = 0.09072$

(3) If we set significance value α as 0.1, $t(1-\alpha/2)=2.01$. Find all variables (including intercept) which are not significant to predict api00. (4pts)

Intercept, Meals

(4) Calculate R^2. (4pts)

0.60447

(5) Instead of percentage for variable "meals", unit of "meals" are changed to ratio of students receiving free meals in range [0,1]. In this case, what is estimated parameter for variable "meals"? (4pts)

÷ 100

-390.242

3. (total 15pts) Suppose $X = (X_1, X_2, \ldots, X_n)$ is iid sample from a Poisson distribution with parameter λ

independent...

(1) Write the likelihood function based on Poisson distribution. (5pts)

Poisson distribution $f(X;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$ $L(x;\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$

SEOULTECH
Seoul National University of Science & Technology

Mid-term Exam

Date : 2016.10.31

Code	ITM 522		Title	Data Mining	
Time for Exam	**1.5hr**	**Questions**	**7**	**Weighting**	**25%**

(2) What is log likelihood function? (5pts)

$$\log L(x;\lambda) = \sum_{i}^{n} (x_i \log \lambda) - n\lambda - \sum_{i}^{n} \log x_i!$$

(3) What is the maximum likelihood estimator for λ with given samples $\mathbf{X} = (X_1, X_2, \ldots, X_n)$? (5pts)

$$\frac{\partial \log L}{\partial \lambda} = \frac{\sum_i x_i}{\lambda} - n = 0$$

$$\frac{\sum_i x_i}{n}$$

4. (25pts) We build classification model using decision tree algorithm to predict variable "Play Golf".

Rainy: no – 3, yes – 2 (0.68)
OC: no – 0, yes – 4 (1)
Sunny: no – 2, yes – 3 (0.98)

NO → 5
YES → 9

Input				Output
Outlook	Temperature	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

SEOULTECH
Seoul National University of Science & Technology

Mid-term Exam

Date : 2016.10.31

Code	ITM 522		Title	Data Mining	
Time for Exam	1.5hr	Questions	7	Weighting	25%

(1) At initial state, calculate Gini impurity of variable "Play Golf". (5pt)

$1-(\frac{5}{14})^2-(\frac{9}{14})^2 = 0.4591$

(2) At first, root node is split by variable "Outlook" and each children node corresponds to each value. After split, calculate impurity of children nodes. (9pt)

Rainy 0.48
overcast 0
Sunny 0.48

(3) Calculate information gain of this split. (6pt)

$0.4591 - 2 \times \frac{5}{14} \times 0.48 = 0.11632$

(4) Which variable is the best for further split of the children node with "Rainy"? (5pts)

Humidity

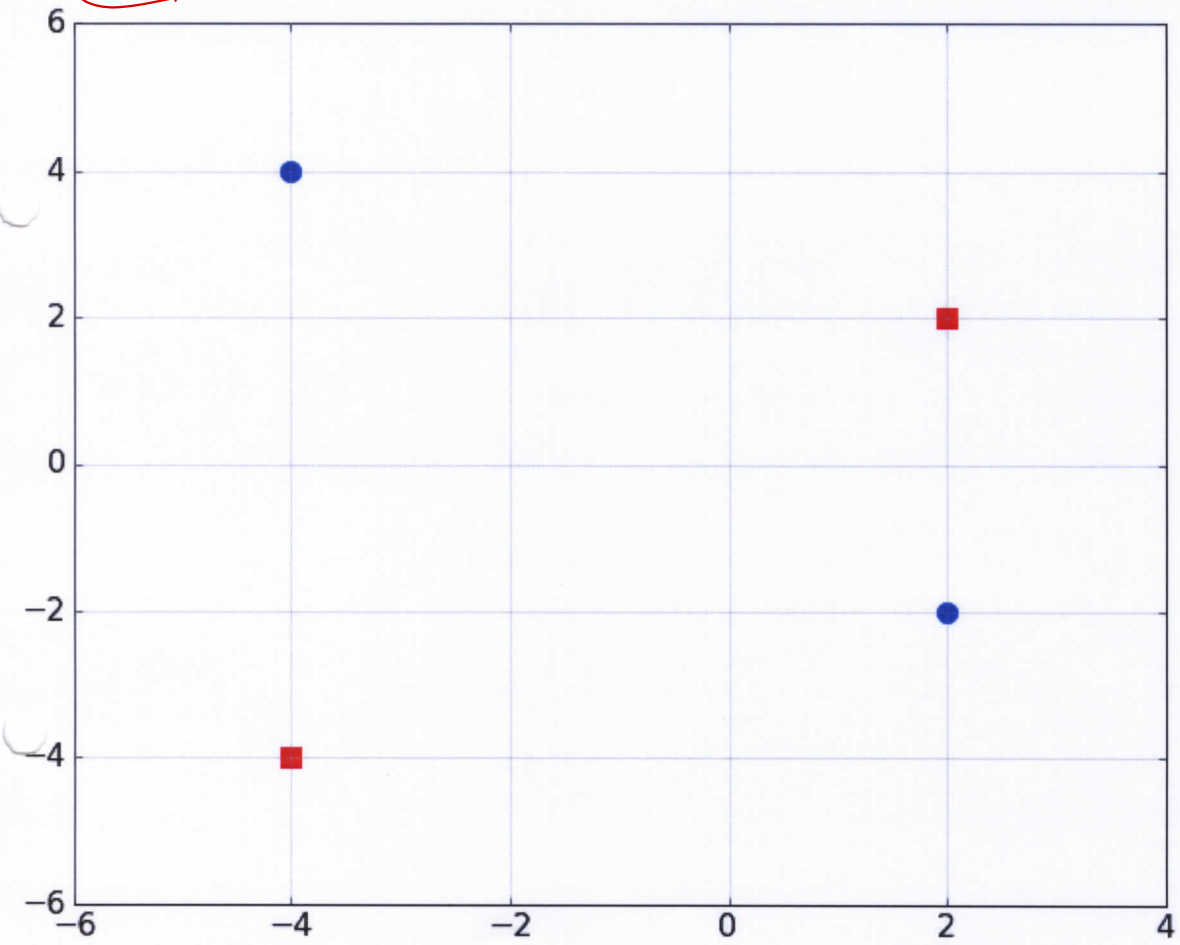
SEOULTECH
Seoul National University of Science & Technology

Mid-term Exam

Date : 2016.10.31

Code	ITM 522		Title	Data Mining	
Time for Exam	1.5hr	Questions	7	Weighting	25%

3. (10pts) Decision boundary is the region of a problem space in which the output label of a classifier is ambiguous. Draw the decision boundaries produced by 1-nearest neighbor classifier on the following dataset. Use Euclidean distance and shade regions where points are classified as red square.



SEOULTECH
Seoul National University of Science & Technology

Mid-term Exam

Date : 2016.10.31

Code	ITM 522		Title	Data Mining	
Time for Exam	**1.5hr**	**Questions**	**7**	**Weighting**	**25%**

6. (10pts) We want to use k-nearest neighbor classifier to predict output class of new data based on following training data set

index	x_1	x_2	Class
1	7.0	3.2	1
2	6.3	3.3	2
3	6.4	3.2	1
4	5.5	2.3	1
5	6.3	2.9	2
6	7.6	3.0	2
7	6.5	2.8	1
8	5.2	2.7	1
9	7.3	2.9	2
10	6.7	2.5	2

(1) Predict output class of new data point, $x_1 = 6.75, x_2 = 2.85$ based on given training samples when $k = 3$ and distance measure is Euclidian distance. (5pts)

(2) If we change distance measure from Euclidian distant to Manhattan distance, list up all nearest neighbors and how is the answer of problem (2) changed? (5pts)

SEOULTECH
Seoul National University of Science & Technology

Mid-term Exam

Date : 2016.10.31

Code	ITM 522		Title	Data Mining	
Time for Exam	**1.5hr**	**Questions**	**7**	**Weighting**	**25%**

7. (5pts) Using binary classifiers, it is possible to build multiclass classifier. Explain what one-versus-one approach and its problems.

SEOULTECH
Seoul National University of Science & Technology

Solutions of Mid-term Exam

Date : 2016.10.31

Code	ITM 522	Title	Data Mining

1. (10pts, each 2pts) Write definition (equation) of following terms.

♦**Difficulty : Easy**
♦**Amount of work: 10 min**
♦**Solution :**

(1) SSE

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

(2) SSR

$$\sum_{i=1}^{n}(y_i - \bar{y})^2$$

(3) logit function

$$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$$

(4) chi-square statistic

$$\chi^2 = \sum_{j=1}^{c}\sum_{k=1}^{r}\frac{(O_{jk} - E_{jk})^2}{E_{jk}}$$

(5) Mahalanobis distance

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)^T S^{-1}(\mathbf{x}_1 - \mathbf{x}_2)}$$

S is sample covariance matrix

2. (30pts) Data set with four variables such as api00, acs_k3, meals and full is given and we want to build a regression model on them

[Variable Description]

api00: academic performance of the school

acs_k3: the average class size in kindergarten through 3rd grade

meals: the percentage of students receiving free meals

full: the percentage of teachers who have full teaching credentials

SEOULTECH
Seoul National University of Science & Technology

Solutions of Mid-term Exam

Date : 2016.10.31

Code	ITM 522	Title	Data Mining

The purpose of regression model is to predict api00 using the other variables.
In data set, there are 313 data points.

Variable	DF	Parameter Estimate	Standard Error	t value
Intercept	1	906.73916	28.26505	
acs_k3	1	-2.68151	1.39399	
Meals	1	-3.70242	0.15403	
Full	1	0.10861		

♦**Difficulty : Intermediate to hard**
♦**Amount of work: 25min**
♦**Solution :**

(1) From fitted model, SST=3,906,597 and SSR=2,634,884. Fill out following table. (18pts, each 2pts)

Source	DF	Sum of Squares	Mean Square	F value
Model	3	2634884	878295	213.41
Error	309	1271713	4115.57673	
Total	312	3906597		

(2 If sample variance of variable "full" is 1602.77, what is standard error of estimated coefficient for variable "full"? (5pts)

$se(\text{"full"}) = \sqrt{\frac{MSE}{\sum_{i=1}^{n}(x_{full}-\bar{x}_{full})^2}}$ and $\sigma^2_{full} = \frac{\sum_{i=1}^{n}(x_{full}-\bar{x}_{full})^2}{n-1} = 1602.77$

So, $\sum_{i=1}^{n}(x_{full}-\bar{x}_{full})^2 = (n-1)\sigma^2_{full} = 312 \times 1602.77 = 500064.24$

$se(\text{"full"}) = \sqrt{\frac{MSE}{\sum_{i=1}^{n}(x_{full}-\bar{x}_{full})^2}} = \sqrt{\frac{4115.57673}{500064.24}} = 0.09072$

(3) If we set significance value α as 0.1, $t(1-\alpha/2)=2.01$. Find all variables (including intercept) which are not significant to predict api00. (4pts)

To find insignificant variables, we have to calculate t values.

t value=(Parameter Estimate)/(Standard Error).

So,

Variable	t value
Intercept	32.08
acs_k3	-1.92
Meals	-24.04
Full	1.20

Because variables whose absolute t value is less than 2.01 is insignificant, only variable which is not significant is variable "full" and "acs_k3"

SEOULTECH
Seoul National University of Science & Technology

Solutions of Mid-term Exam

Date : 2016.10.31

Code	ITM 522	Title	Data Mining

(4) Calculate R^2(4pts)

R^2=SSR/SST=(1271713)/(3906597)=0.6745

(5) Instead of percentage for variable "meals", unit of "meals" are changed to ratio of students receiving free meals in range [0,1]. In this case, what is estimated parameter for variable "meals"? (4pts)

new "meals"=old "meals"/100, so estimated parameter increases to 100times.

-3.70242*10=-370.242

3. (15pts) Suppose $X = (X_1, X_2, \dots, X_n)$ is iid sample from a Poisson distribution with parameter λ

♦**Difficulty : Intermediate**

♦**Amount of work: 10min**

♦**Solution :**

(1) Write the likelihood function based on Poisson distribution. (5pts)

Poisson distribution $f(X;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$

$$\mathcal{L} = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

(2) What is log likelihood function? (5pts)

$$\log\mathcal{L} = \sum_{i=1}^{n} (x_i \log\lambda - \log x_i!) - n\lambda$$

(3) What is the maximum likelihood estimator for λ with given samples $X = (X_1, X_2, \dots, X_n)$? (5pts)

$$\frac{\partial \log\mathcal{L}}{\partial\lambda} = \frac{\sum_{i=1}^{n} x_i}{\lambda} - n = 0$$

$$\lambda = \frac{\sum_{i=1}^{n} x_i}{n}$$

SEOULTECH
Seoul National University of Science & Technology

Solutions of Mid-term Exam

Date : 2016.10.31

Code	ITM 522	Title	Data Mining

4. (20pts) We build classification model using decision tree algorithm to predict variable "Play Golf".

Input				Output
Outlook	Temperature	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

♦**Difficulty : Intermediate to hard**

♦**Amount of work: 20min**

♦**Solution :**

(1) At initial state, calculate Gini impurity of variable "Play Golf". (5pt)

$$1 - \sum p_i^2 = 1 - \left(\frac{5}{14}\right)^2 - \left(\frac{9}{14}\right)^2 = 0.4592$$

(2) At first, root node is split by variable "Outlook" and each children node corresponds to each value. After split, calculate Gini impurity of children nodes. (9pt)

Sunny: (Yes/No)=(3/2)

$$1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48$$

Overcast: (Yes/No)=(4,0)

$$1 - \left(\frac{4}{4}\right)^2 - \left(\frac{0}{4}\right)^2 = 0$$

Rainy: (Yes/No)=(2,3)

$$1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48$$

(3) Calculate information gain of this split. (6pt)

information gain=$0.4592 - 0.48 \times \frac{5}{14} \times 2 = 0.1163$

SEOULTECH
Seoul National University of Science & Technology

Solutions of Mid-term Exam

Date : 2016.10.31

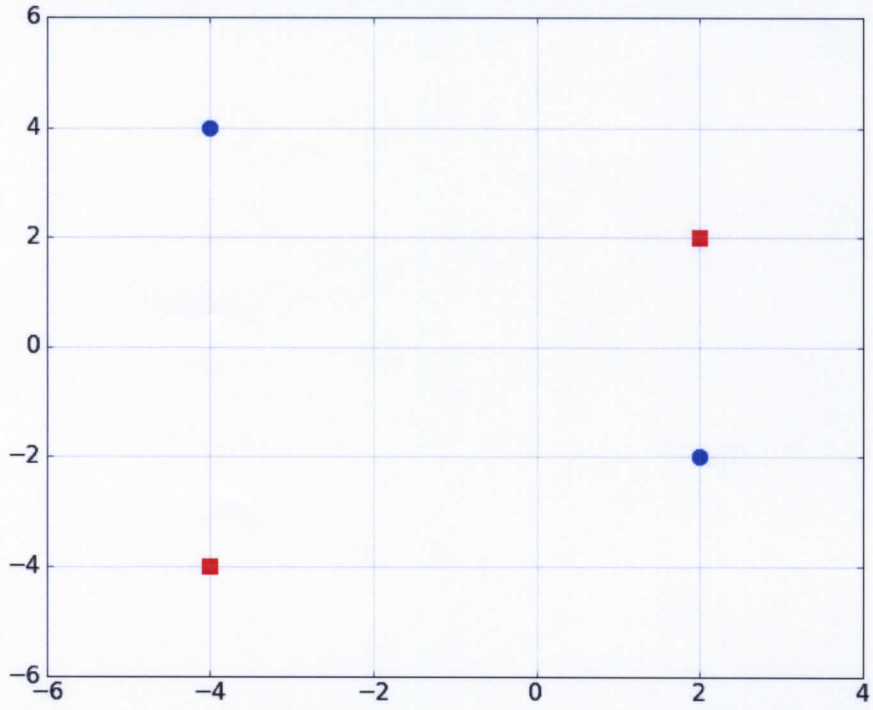
Code	ITM 522	Title	Data Mining

(4) Which variable is the best for further split of the children node with "Rainy"? (5pts)

Input				Output
Outlook	Temperature	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Rainy	Mild	Normal	True	Yes

When humidity is used to split further, the perfect split can be achieved. Therefore, humidity (high or normal) is the best split

5. (10pts) Decision boundary is the region of a problem space in which the output label of a classifier is ambiguous. Draw the decision boundaries produced by 1-nearest neighbor classifier on the following dataset. Use Euclidean distance and shade regions where points are classified as red square.



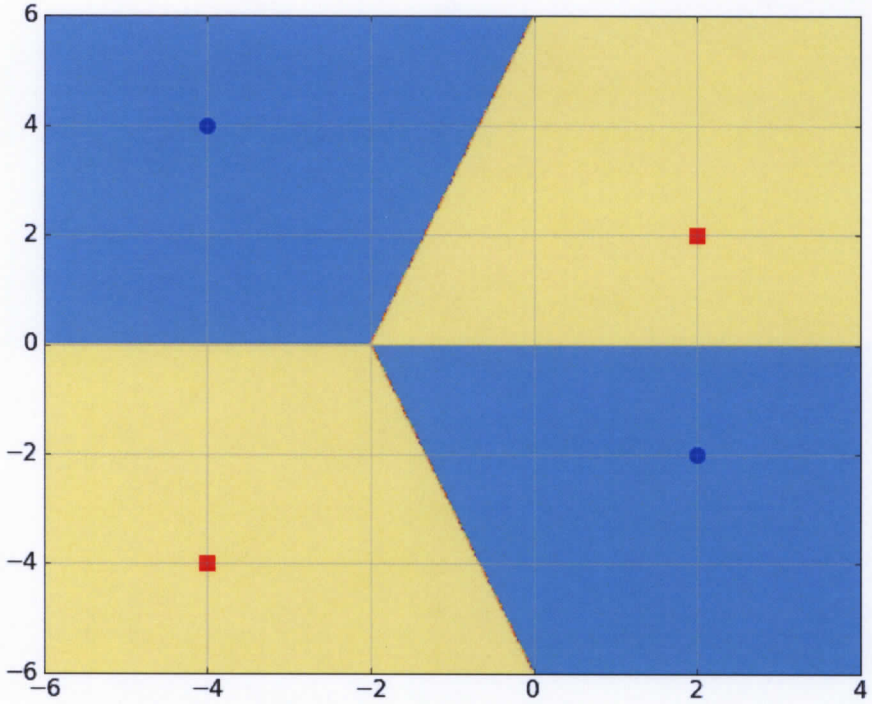
♦**Difficulty : Difficult**
♦**Amount of work: 5min**
♦**Solution :**

SEOULTECH
Seoul National University of Science & Technology

Solutions of Mid-term Exam

Date : 2016.10.31

Code	ITM 522	Title	Data Mining



6. (10pts) We want to use k-nearest neighbor classifier to predict output class of new data based on following training data set

index	x_1	x_2	Class
1	7.0	3.2	1
2	6.3	3.3	2
3	6.4	3.2	1
4	5.5	2.3	1
5	6.3	2.9	2
6	7.6	3.0	2
7	6.5	2.8	1
8	5.2	2.7	1
9	7.3	2.9	2
10	6.7	2.5	2

♦Difficulty : Intermediate

♦Amount of work: 15min

♦Solution :

(1) Predict output class of new data point, $x_1 = 6.75, x_2 = 2.85$ based on given training samples when $k = 3$ and distance measure is Euclidian distance. (5pts)

index	Euclidian distance for 5th point
1	0.4301
2	0.6364
3	0.4950
4	1.3657

SEOULTECH
Seoul National University of Science & Technology

Solutions of Mid-term Exam

Date : 2016.10.31

Code	ITM 522	Title	Data Mining

5	0.4528
6	0.8631
7	0.2550
8	1.5572
9	0.8746
10	0.3536

So, 3-nearest neighbors are 1st point, 7th point, 10th point.
Output class of 1st point and 7th point is 1 and output class of 10th point is 2. Therefore output class of new input data is 1 because majority class is 1.

(2) If we change distance measure from Euclidian distant to Manhattan distance, list up all nearest neighbors and how is the answer of problem (2) changed? (5pts)

index	Manhattan distance for 5th point
1	0.6
2	0.9
3	0.7
4	1.8
5	0.5
6	1.0
7	0.3
8	1.7
9	1.2
10	0.4

So, 3-nearest neighbors are 5th point, 7th point, 10th point.
Output class of 7th point is 1 and output class of 5th point and 10th point is 2. Therefore output class of new input data is 2 because majority class is 2.

(5pts) Using binary classifiers, it is possible to build multiclass classifier. Explain what one-versus-one approach and its problems.

♦**Difficulty : Easy**
♦**Amount of work: 5min**
♦**Solution :**

(3pts) For multiclass classification commonly used approach is to construct K separate binary classifiers. Each model is trained using the data from class C_k as the positive examples and the data from the remaining $K-1$ classes as the negative examples.

- Drawbacks (2pts)

① Because each classifier was trained on different task, there is no guarantee that the real-values quantities $y_k(\mathbf{x})$ will have appropriate scales
② Imbalance of data on training

northumbria UNIVERSITY NEWCASTLE

(Form MOD1): External Examiner Approval of Assessment Tasks

(coursework and examination)

Process:
Assessment tasks/briefs and related information should be carefully checked before[1] being sent to the External Examiner for approval: briefs and papers should, therefore, be marked draft until confirmed by the External Examiner. Further information can be found under Marking Moderation Processes on the Assessment Web Page[2].

Once the checklist below has been completed by the Module Lead, this form should be provided to the external examiner with the assessment information.

Faculty	[ADSS][BL][EE][HLS]	**Department**	**ITM, SeoulTech**
Module Code	**ITM 522**	**Module Title**	**Data Mining**
Module Tutor	**Kyoungok Kim**	**Level**	**Level 5**
Student Year Group / Intake	**2016 / Fall**	**Credit Value**	**10**
Description of Assessment / Component	**EXAM (Final Exam)**	**% of Module Mark**	**50%**
Delivery Location		**SeoulTech**	

Checklist for coursework / module assignments / examination	**YES / NO** *(please specify)*
Alignment to module learning outcomes and assessment task on MD	YES
Free from typing / grammatical errors	YES
Clear information / instructions / rubric provided	YES
Word lengths and penalties included where appropriate (coursework)	YES
Marks allocation for components included (*if relevant*)	YES
Hand-in date, feedback process and timeframe clearly stated	YES
Marking criteria provided for students (which map to grade descriptors)	YES
Marking guidelines / solutions / answers (*for internal / external moderator*)	YES
Option for alternative assessment considered (*in this case this is required*)	N/A
Any extra documentation (*e.g. for examinations*) clearly labelled	N/A.
Adheres to any PSRB requirements	YES
Liaison with placement provider (*where appropriate*)	N/A
Assessment for referrals included or approval date agreed with external examiner	N/A
Consultation with EPWO partners to check for contextual relevance	N/A.

Length of EE approval for this assessment *(please circle)*	One delivery only 1 year 2 years 3 years Until end of tenure	
Name and contact Details of Moderator	**Date**	**Comments**
Hakyeon Lee hylee@seoultech.ac.kr (signature)	17.10.2016	· Total marks are 105 pts. Is this intended? · Students may need more time.

[1] For educational partners a Northumbria member of staff should sign to assure NU oversight of assessment task

[2] https://www.northumbria.ac.uk/sd/central/ar/qualitysupport/asspolicies/

E1 Form (to be completed by module tutor for all examination assessments)

Exam Description (Paper 1 etc.)	**Exam size** (Please enter number of students expected to sit exam)	**Does exam need to be same day as Part Time/Distance Learning students?** (Please specify day for exam to be sat if taken by part time students.)	**Is the module taught on a Campus other than Newcastle City Campus?** (If runs as CV, does the exam need to start at 09:30?)	**Campus where module is to be examined.**
5	25	Yes	Yes	SeoulTech

EXAM DETAILS

Boxes 1, 2, 3 and 5 along with the information provided above are essential in order that an exam can be scheduled, if any of these boxes are left blank the exam details may not be entered correctly and errors in scheduling may occur.

1. Duration (the duration must **include reading time, if appropriate**)	**2. Type of Exam**		**3. Accommodation Requirements** Please tick or state as requested.		**4. Are there any constraints for this examination?** **You must also provide a valid reason for this constraint, otherwise it will not be included. Exam constraints should only be applied if essential.**	**5. Does this exam need to run at the same time as another exam?** Please enter module code in the box below.
2.5hrs	☐	Open book	**SD** (single desk)	☐	Constraint requested:	This exam must run at the same time as
	☒	Closed Book	**DD** (double desk)	☒		
	☐	Written	**LB** (locally booked,	☐		(module code)
	☐	Written closed book	If a specific room is required, please state room number**, this must be booked by the Faculty**.		Reason for constraint above:	Do the exams need to take place in the same room?
	☐	Unrestricted			.	
	☐	Restricted notes				Yes / No
	☐	Open Notes				
	☐	Pre-work				
	☐	Lab based exam				
	☐	OMR (optical mark recording)	**Lab** (please state lab number) **this must be booked by the Faculty**			
	☐	Other – please specify				

SEOULTECH
Seoul National University of Science & Technology

Final Exam

Date : 2016.12.19

Code	ITM 522		Title	Data Mining	
Time for Exam	**2.5hr**	**Questions**	**11**	**Weighting**	**50%**

1. (10pts, each 2pts) Write definition (equation) of following terms.

(1) false positive error

(2) eigenvector If A is matrix, the following vector X which satisfies the equation is eigenvector.

$$Ax = \lambda x$$

(3) Bayes' theorem

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

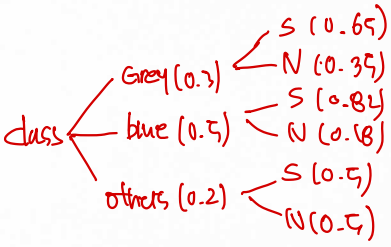
(4) confidence of association rule

$$\frac{n(X \cup Y)}{n(X)}$$

(5) covariance of two random variables X,Y

$$\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

2. (5pts) In a certain day care class, 30% of the children have grey eyes, 50% of them have blue and the other 20%'s eyes are in other colors. One day they play a game together. In the first run, 65% of the grey eye ones, 82% of the blue eyed ones and 50% of the children with other eye color were selected. Now, if a child is selected randomly from the class, and we know that he/she was not in the first game, what is the probability that the child has blue eyes?



개별케이에서 선택X 일때 blue eyes 확률 $= \frac{P(\text{blue}) P(\text{선택X} \mid \text{blue})}{P(\text{선택X})}$

$$\frac{0.5 \times 0.5 \times 0.18}{0.3 \times 0.35 + 0.5 \times 0.18 + 0.2 \times 0.5}$$

$= 0.1525$

3. (16pts) Logistic regression model was trained on the dataset to classify whether a patient has diabetes or not (diabetes=1, otherwise 0). The total number of train samples is 768.

p=5

variable	coefficient
intercept	-7.8141
pregnant	0.1449
glucose	0.0363
blood	-0.0118
insulin	-0.0010
BMI	0.0907

β_0

$$\frac{e^{-7.8141 + 0.1449}}{e^{-7.8141 + 0.0363}}$$

SEOULTECH
Seoul National University of Science & Technology

Final Exam

Date : 2016.12.19

Code	ITM 522		Title	Data Mining	
Time for Exam	**2.5hr**	**Questions**	**11**	**Weighting**	**50%**

(1) What are the degree of freedom of the model and the degree of freedom of error(residuals)? (4pts)

Model = 6개
Error = 5

(2) Calculate odds ratio of variables "pregnant" and "glucose". (4pts) 정확히 어떤건지 모름.

(3) Write the link function of the trained logit model (logit function was used) with trained coefficients. (4pts)

$-8.4 + 0.1231 \times_{preg} + 0.0363 \times_{glucose} - 0.0138 \times_{blood} - 0.001 \times_{insulin} + 0.0901 \times_{BMI}$

(4) Calculate the probability of diabetes for the following sample. (4pts)

	pregnant	glucose	blood	insulin	BMI
value	3	78	50	88	31

0.08207

SEOULTECH
Seoul National University of Science & Technology

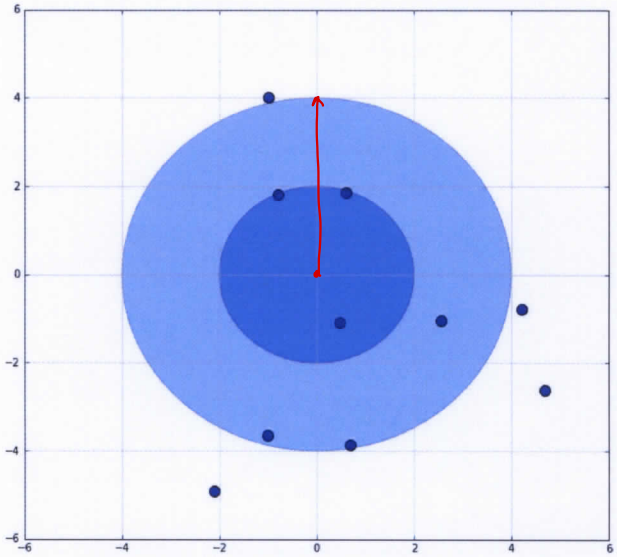
Final Exam

Date : 2016.12.19

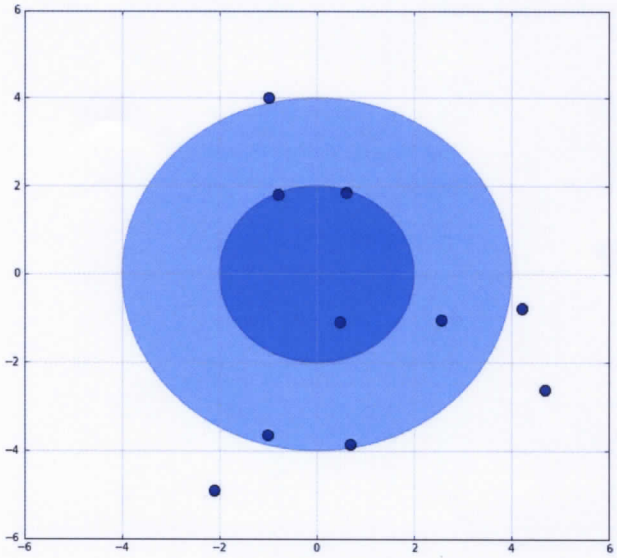
Code	ITM 522		Title	Data Mining	
Time for Exam	**2.5hr**	**Questions**	**11**	**Weighting**	**50%**

4. (10pts) Using fixed radius neighbor regression, regression model is trained based on given train data.

(1) When Euclidean distance is used and $r = 4$, how many points are included into the neighbor set of point (0,0)? Explain. (5pts)



(2) When Manhattan distance is used and $r = 4$, how many points are included into the neighbor set of point (0,0)? Explain. (5pts)



SEOULTECH
Seoul National University of Science & Technology

Final Exam

Date : 2016.12.19

Code	ITM 522	Title	Data Mining		
Time for Exam	**2.5hr**	**Questions**	**11**	**Weighting**	**50%**

5. (10pts) We built Naïve Bayes classifier on the data whose 4 input variables are all binary and output variables is categorical variable with 3 different classes. Trained Naïve Bayes classifier is described in following table.

	Class 1	Class 2	Class 3
prior	0.4	0.3	0.3
$p(x_1 = 1)$	0.7	0.3	0.4
$p(x_2 = 1)$	0.5	0.1	0.6
$p(x_3 = 1)$	0.5	0.9	0.5
$p(x_4 = 1)$	0.2	0.4	0.6

When new input data is $(x_1, x_2, x_3, x_4) = (1,0,1,0)$, determine output class based on the Naïve Bayes classifier.

① $0.4 \times 0.7 \times 0.5 \times 0.5 \times 0.8 = 0.056$

② $0.3 \times 0.3 \times 0.9 \times 0.9 \times 0.6 = 0.0486$

③ $0.3 \times 0.4 \times 0.4 \times 0.5 \times 0.4$

class 1

6. (15pts) We want to build Gaussian naïve Bayes model. $= \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

(1) Write the likelihood function and log likelihood function of Gaussian naïve Bayes model. (5pts)

$$L = \prod_k p(C_k) \prod_i \frac{1}{\sqrt{2\pi\sigma_{ki}^2}} e^{\frac{-(x_{ki}-\mu_{ki})^2}{2\sigma_{ki}^2}}$$

$$\log L = \sum_k \log p(C_k) + \sum_k \sum_i \frac{1}{2} \log(2\pi\sigma_{ki}^2) - \frac{(x_{ki}-\mu_{ki})^2}{2\sigma_{ki}^2}$$

(2) Calculate maximum likelihood estimates of $\mu_{k,i}$ and $\sigma_{k,i}$ for Gaussian naïve Bayes model(k represents different class, i represent feature) (10pts)

7. (5pts) Write primal form of optimization problem of support vector machine for binary classification problem.

SEOULTECH
Seoul National University of Science & Technology

Final Exam

Date : 2016.12.19

Code	ITM 522	Title	Data Mining		
Time for Exam	**2.5hr**	**Questions**	**11**	**Weighting**	**50%**

8. (10pts) Using Gaussian kernel with $\gamma = 1$, support vector machine binary classification was trained. After training, support vectors and their coefficient were obtained. Intercept is 0.044.

	x_1	x_2	y	a
1	0.63	0.48	-1	1
2	-0.05	0.99	-1	1
3	0.88	0.12	-1	1
4	-1.01	0.64	-1	1
5	-0.59	0.65	-1	0.27
6	0.09	-0.59	1	1
7	1.10	0.42	1	1
8	2.10	-0.51	1	1
9	0.11	-0.43	1	1
10	1.82	-0.43	1	0.27

[Hint: SVM]
Decision function for linear case
$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} - b$

Guassian kernel
$k(\mathbf{x},\mathbf{x}') = \exp(-\gamma\|\mathbf{x} - \mathbf{x}'\|^2)$

(1) Write decision function. (3pts)

(2) Calculate decision function value for data point (0,0) (5pts)

(3) Determine output class. (2pts)

9. (5pts) Using k-means clustering, we want to group data points into two classes. Following table describes a certain iteration of k-means clustering.

SEOULTECH
Seoul National University of Science & Technology

Final Exam

Date : 2016.12.19

Code	ITM 522	Title	Data Mining
Time for Exam	2.5hr	Questions	11	Weighting	50%

	1	2	3	4	5
x_1	1	2	3	3	4
x_2	1	1	2	4	2
Group	1	1	1	2	2

(1) Calculate centroids of two groups. (2pts)

1: $(x_1 = 2,\ x_2 = \frac{4}{3})$

2: $(x_1 = 3.5,\ x_2 = 3)$

(2) Using new centroids, update group based on Manhattan distance. (3pts)

1 1 2 2 2.

10. (10pts) After training the linear regression model, adaptedness should be tested.

(1) Write test statistics of Jarque-Bera test and purpose of this test. (5pts)

Use following notations.

n: the number of train samples

p: the number of independent features

S: sample skewness

ample kurtosis

$$\frac{n-p}{6}\left(S^2 + \frac{(c-3)^2}{4}\right)$$

(2) The number of train samples is 300 and the number of independent features is 10. Sample skewness is 0.297 and sample kurtosis is 2.813. In this case, determine whether this linear regression is appropriate method for fitting ($\alpha = 0.1$). (5pts)

$$\frac{290}{6} \times \left(0.297^2 + \frac{(2.813-3)^2}{4}\right) \qquad chi^2(\alpha = 0.1)$$

11. (9pts) Based on following transaction data, we want to create association rules using Apriori algorithm. Set *s_min*=0.6

TID	Items
1	M, O, N, K, E, Y
2	D, O, N, K, E, Y
3	M, A, K, E

SEOULTECH
Seoul National University of Science & Technology

Final Exam

Date : 2016.12.19

Code	ITM 522	Title	Data Mining		
Time for Exam	**2.5hr**	**Questions**	**11**	**Weighting**	**50%**

4	M, U, C, K, Y
5	C, O, O, K, I, E

(1) Generate C_1 and L_1 (3pts)

$C_1 = \{\{M\}, \{O\}, \{N\}, \{K\}, \{E\}, \{Y\}, \{O\}, \{A\}, \{U\}, \{C\}, \{I\}\}$

$L_1 = \{\{M\}, \{O\}, \{N\}, \{K\}, \{E\}, \{Y\}\}$

(2) Generate C_2 and L_2 (3pts)

$C_2 = \{\{M,O\}, \{M,N\}, \{M,K\}, \{M,E\}, \{M,Y\}, \{O,N\}, \{O,K\}, \{O,E\}, \{O,Y\}, \{N,K\}, \{N,E\}, \{N,Y\}, \{K,E\}, \{K,Y\}, \{E,Y\}\}$

$L_2 = \{\{M,K\}, \{O,K\}, \{O,E\}, \{K,E\}, \{K,Y\}\}$

(3) What is support and confident of rule "If K, then E" (3pts)

$$Supp(K \to E) = \frac{n(K \cup E)}{n(all)} = \frac{4}{6} = 0.8$$

$$Conf(K \to E) = \frac{n(K \cup E)}{n(K)} = \frac{4}{6} = 0.8$$

SEOULTECH
Seoul National University of Science & Technology

Solutions of Final Exam

Date : 2016.12.19

Code	ITM 522	Title	Data Mining

1. (10pts, each 2pts) Write definition (equation) of following terms.

♦**Difficulty : Easy**
♦**Amount of work: 10 min**
♦**Solution :**

(1) false positive error

The incorrect rejection of a true null hypothesis or detecting an effect that is not present

(2) eigenvector

For $n \times n$ matrix A, eigenvector $\mathbf{v}$ of A is the vector that satisfies $A\mathbf{v} = \lambda\mathbf{v}$

(3) Bayes' theorem

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

(4) confidence of association rule

$$Confidence = P(result|condition) = \frac{P(condition \cap result)}{P(condition)}$$
$$= \frac{\#\ of\ transactions\ that\ include\ both\ condition\ and\ result}{\#\ of\ transactions\ that\ include\ condition}$$

(5) covariance of two random variables X,Y

$$Cov(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])^2]\mathbb{E}[(Y - \mathbb{E}[Y])^2]$$

or

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

SEOULTECH
Seoul National University of Science & Technology

Solutions of Final Exam

Date : 2016.12.19

Code	ITM 522	Title	Data Mining

2. (5pts) In a certain day care class, 30% of the children have grey eyes, 50% of them have blue and the other 20%'s eyes are in other colors. One day they play a game together. In the first run, 65% of the grey eye ones, 82% of the blue eyed ones and 50% of the children with other eye color were selected. Now, if a child is selected randomly from the class, and we know that he/she was not in the first game, what is the probability that the child has blue eyes?

♦**Difficulty : Intermediate**
♦**Amount of work: 10min**
♦**Solution :**

Let's say B= blue, G= grey and O= "Other color" and NR= "not selected for the first run"

$P(B|NR)=P(NR|B)P(B)P(G)P(NR|G)+P(B)P(NR|B)+P(O)P(NR|O)$

On substituting values

$P(B|NR)=0.5\cdot(1-0.82)(0.3\cdot(1-0.65))+(0.5\cdot(1-0.82))+(0.2\cdot(1-0.5))$

$P(B|NR)=0.305$

3. (16pts) Logistic regression model was trained on the dataset to classify whether a patient has diabetes or not (diabetes=1, otherwise 0). The total number of train samples is 768.

variable	coefficient
intercept	-7.8141
pregnant	0.1449
glucose	0.0363
blood	-0.0118
insulin	-0.0010
BMI	0.0907

♦**Difficulty : Intermediate**
♦**Amount of work: 15min**
♦**Solution :**

(1) What are the degree of freedom of the model and the degree of freedom of error(residuals)? (4pts)

df model = 5

df residuals = 762

(2) Calculate odds ratio of variables "pregnant" and "glucose". (4pts)

odds ratio = e^{β_i}

odds ratio of pregnant = $e^{0.1449} = 1.1559$

SEOULTECH
Seoul National University of Science & Technology

Solutions of Final Exam

Date : 2016.12.19

Code	ITM 522	Title	Data Mining

odds ratio of glucose = $e^{0.0363} = 1.0370$

(3) Write the link function of the trained logit model (logit function was used) with trained coefficients. (4pts)

$$g(P) = \ln\frac{P}{1-P} = -7.8141 + 0.1449x_{pregnant} + 0.0363x_{glucose} - 0.0118x_{blood} - 0.001x_{insulin} + 0.0907x_{BMI}$$

(4) Calculate the probability of diabetes for the following sample. (4pts)

	pregnant	glucose	blood	insulin	BMI
value	3	78	50	88	31

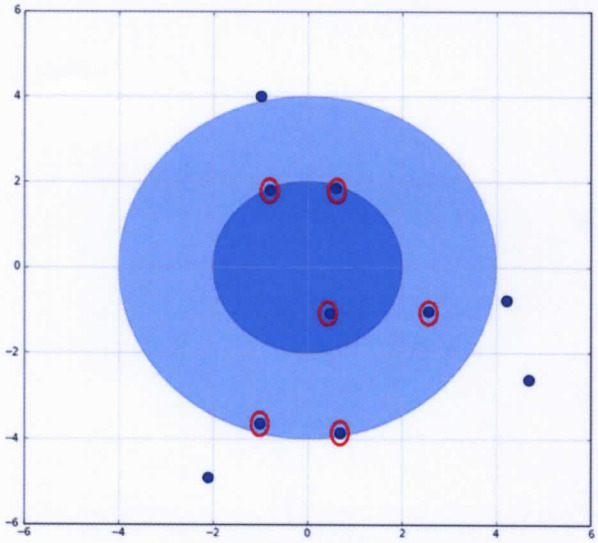
$$t = -7.8141 + 0.1449x_{pregnant} + 0.0363x_{glucose} - 0.0118x_{blood} - 0.001x_{insulin} + 0.0907x_{BMI} = -2.4144$$

$$p(Y=1) = \frac{1}{1+e^{-t}} = \frac{1}{1+e^{2.4144}} = \frac{1}{1+11.1827} = 0.082$$

4. (10pts) Using fixed radius neighbor regression, regression model is trained based on given train data.

♦**Difficulty : Intermediate**
♦**Amount of work: 10min**
♦**Solution :**

(1) When Euclidean distance is used and $r = 4$, how many points are included into the neighbor set of point (0,0)? Explain. (5pts)



Data points inside the circle with radius 4 are neighbors of point (0,0). → the number of elements in neighbor set = 6

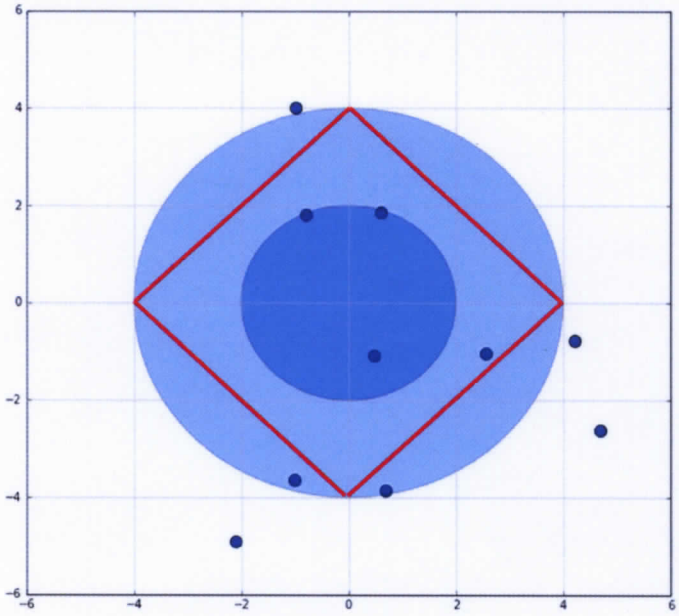
SEOULTECH
Seoul National University of Science & Technology

Solutions of Final Exam

Date : 2016.12.19

Code	ITM 522	Title	Data Mining

(2) When Manhattan distance is used and $r = 4$, how many points are included into the neighbor set of point (0,0)? Explain. (5pts)



The red line represents equidistance line from (0,0) of length 4 based on Manhattan distance. → the number of elements in neighbor set = 4

5. (10pts) We built Naïve Bayes classifier on the data whose 4 input variables are all binary and output variables is categorical variable with 3 different classes. Trained Naïve Bayes classifier is described in following table.

	Class 1	Class 2	Class 3
prior	0.4	0.3	0.3
$p(x_1 = 1)$	0.7	0.3	0.4
$p(x_2 = 1)$	0.5	0.1	0.6
$p(x_3 = 1)$	0.5	0.9	0.5
$p(x_4 = 1)$	0.2	0.4	0.6

When new input data is $(x_1, x_2, x_3, x_4) = (1,0,1,0)$, determine output class based on the Naïve Bayes classifier.

♦**Difficulty : Intermediate**

♦**Amount of work: 15min**

♦**Solution :**

Given input point, we have to calculate posterior probability for each class.

$$p(C_k|\mathbf{x}) \propto p(C_k)p(\mathbf{x}|C_k)$$

For class 1,

$p(C_1)p(\mathbf{x}|C_1) = 0.4 \times 0.7 \times 0.5 \times 0.5 \times 0.8 = 0.056$

SEOULTECH
Seoul National University of Science & Technology

Solutions of Final Exam

Date : 2016.12.19

Code	ITM 522	Title	Data Mining

For class 2,

$p(C_2)p(\mathbf{x}|C_2) = 0.3 \times 0.3 \times 0.9 \times 0.9 \times 0.4 = 0.029$

For class 3,

$p(C_3)p(\mathbf{x}|C_3) = 0.3 \times 0.4 \times 0.4 \times 0.5 \times 0.4 = 0.010$

Because decision function is $\hat{y} = \underset{k\in\{1,2,3\}}{\operatorname{argmax}} p(C_k) \prod_{i=1}^{4} p(x_i|C_k)$, final class of new input is class 1.

6. (15pts) We want to build Gaussian naïve Bayes model.

♦**Difficulty : Difficult**

♦**Amount of work: 25min**

♦**Solution :**

(1) Write the likelihood function of Gaussian naïve Bayes model. (5pts)

$$p(x = v|C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{\frac{(v-\mu_k)^2}{2\sigma_k^2}}$$

$$p(C_k)\prod_{i=1}^{p} p(x_i|C_k) = p(C_k)\prod_{i=1}^{p} \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{\frac{(v_i-\mu_{k,i})^2}{2\sigma_k^2}}$$

$$\mathcal{L} = \prod_{j=1}^{n} p\left(C_{y_j}\right)\prod_{i=1}^{p} \frac{1}{\sqrt{2\pi\sigma_{y_j,i}^2}} e^{\frac{\left(v_{ji}-\mu_{y_j,i}\right)^2}{2\sigma_{y_j,i}^2}}$$

$$\log\mathcal{L} = \sum_{j=1}^{n} \log p\left(C_{y_j}\right) - \sum_{j=1}^{n}\sum_{i=1}^{p} \frac{1}{2}\log(2\pi\sigma_{y_j,i}^2) - \sum_{j=1}^{n}\sum_{i=1}^{p} \frac{\left(v_{ji}-\mu_{y_j,i}\right)^2}{2\sigma_{y_j,i}^2}$$

(2) Calculate maximum likelihood estimates of $\mu_{k,i}$ and $\sigma_{k,i}$ for Gaussian naïve Bayes model(k represents different class, i represent feature) (10pts)

$$\frac{\partial \log\mathcal{L}}{\partial \mu_{k,i}} = \sum_{j\in\{m:y_m=k\}} \frac{(v_{ji}-\mu_{k,i})}{2\sigma_{k,i}^2} = 0$$

$$\mu_{k,i} = \frac{\sum_{j\in\{m:y_m=k\}} v_{ji}}{n_k}$$

where n_k is the number of train sample in class k

SEOULTECH
Seoul National University of Science & Technology

Solutions of Final Exam

Date : 2016.12.19

Code	ITM 522	Title	Data Mining

$$\frac{\partial \log \mathcal{L}}{\partial \sigma_{k,i}^2} = -\sum_{j \in \{m:y_m=k\}} \frac{1}{2\sigma_{k,i}^2} + \sum_{j \in \{m:y_m=k\}} \frac{(v_{ji}-\mu_{k,i})^2}{2(\sigma_{k,i}^2)^2} = 0$$

$$\sigma_{k,i}^2 = \frac{(v_{ji}-\mu_{k,i})^2}{n_k}$$

7. (5pts) Write primal form of optimization problem of support vector machine for binary classification problem.

♦**Difficulty : Easy**
♦**Amount of work: 5min**
♦**Solution :**

$$\underset{(\mathbf{w},b)}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{w}\|^2$$

$$s.t. \quad y_i(\mathbf{w}\cdot\mathbf{x}-b) \geq 1$$

8. (10pts) Using Gaussian kernel with $\gamma = 1$, support vector machine binary classification was trained. After training, support vectors and their coefficient were obtained. Intercept is 0.044.

	x_1	x_2	y	α
1	0.63	0.48	-1	1
2	-0.05	0.99	-1	1
3	0.88	0.12	-1	1
4	-1.01	0.64	-1	1
5	-0.59	0.65	-1	0.27
6	0.09	-0.59	1	1
7	1.10	0.42	1	1
8	2.10	-0.51	1	1
9	0.11	-0.43	1	1
10	1.82	-0.43	1	0.27

♦**Difficulty : Intermediate**
♦**Amount of work: 20min**
♦**Solution :**
(1) Write decision function. (3pts)
Decision function of SVM with kernel

$$f(\phi(\mathbf{x})) = \mathbf{w}^T\phi(\mathbf{x}) + b = \sum_{i=1}^{n} \alpha_i y_i \phi(\mathbf{x}_i)^T\phi(\mathbf{x}) + b = \sum_{i=1}^{n} \alpha_i y_i k(\mathbf{x}_i,\mathbf{x}) + b$$

SEOULTECH
Seoul National University of Science & Technology

Solutions of Final Exam

Date : 2016.12.19

Code	ITM 522	Title	Data Mining

(2) Calculate decision function value for data point (0,0) (5pts)

So, to calculate decision function value, first calculate kernel function value.

	1	2	3	4	5	6	7	8	9	10
$k(\mathbf{x}_i, \mathbf{x})$	0.53	0.37	0.45	0.24	0.46	0.70	0.25	0.01	0.82	0.03

$$f(\phi(\mathbf{x})) = -0.53 - 0.37 - 0.45 - 0.24 - 0.46 \times 0.27 + 0.70 + 0.25 + 0.01 + 0.82 + 0.03 \times 0.27 + 0.044 = 0.12$$

(3) Determine output class. (2pts)

$f(\phi(\mathbf{x})) = 0.12 > 0$, so final output class is 1.

9. (5pts) Using k-means clustering, we want to group data points into two classes. Following table describes a certain iteration of k-means clustering.

	1	2	3	4	5
x_1	1	2	3	3	4
x_2	1	1	2	4	2
Group	1	1	1	2	2

♦**Difficulty : Easy to intermediate**

♦**Amount of work: 15min**

♦**Solution :**

(1) Calculate centroids of two groups. (2pts)

(1pt) Group1: $C_1 = \left(\frac{1+2+3}{3}, \frac{1+1+4}{3}\right) = \left(2, \frac{4}{3}\right)$

(1pt) Group2: $C_3 = \left(\frac{3+4}{2}, \frac{4+2}{2}\right) = (3.5, 3)$

(2) Using new centroids, update group. (3pts)

	1	2	3	4	5
x_1	1	2	3	3	4
x_2	1	1	3	4	4
Distance from C_1	1.33	0.33	1.67	3.67	2.67
Distance from C_2	4.5	3.5	1.5	1.5	1.5
Group	1	1	2	2	2

SEOULTECH
Seoul National University of Science & Technology

Solutions of Final Exam

Date : 2016.12.19

Code	ITM 522	Title	Data Mining

10. (10pts) After training the linear regression model, adaptedness should be tested.

♦**Difficulty : Intermediate**
♦**Amount of work: 10min**
♦**Solution :**

(1) Write test statistics of Jarque-Bera test and purpose of this test. (5pts)

Use following notations.

n: the number of train samples

p: the number of independent features

S: sample skewness

C: sample kurtosis

$$JB = \frac{n-p}{6}\left(S^2 + \frac{1}{4}(C-3)^2\right)$$

Jarque–Bera test is a goodness-of-fit test of whether sample data have the skewness and kurtosis matching a normal distribution. It is used to test normality of residuals.

(2) The number of train samples is 300 and the number of independent features is 10. Sample skewness is 0.297 and sample kurtosis is 2.813. In this case, determine whether this linear regression is appropriate method for fitting ($\alpha = 0.1$). (5pts)

$$JB = \frac{300-10}{6}\left(0.297^2 + \frac{1}{4}(2.813-3)^2\right) = 4.68598$$

If the data comes from a normal distribution, JB statistic asymptotically has a chi-squared distribution with two degrees of freedom.

When $\alpha = 0.1$, $\chi_{0.1;2} = 4.605$ and JB is greater than 4.605. It meant that residuals do not follow the normal distribution. Therefore linear regression is not appropriate model for the data.

11. (9pts) Based on following transaction data, we want to create association rules using Apriori algorithm. Set *s_min*=0.6

TID	Items
1	M, O, N, K, E, Y
2	D, O, N, K, E, Y
3	M, A, K, E
4	M, U, C, K, Y
5	C, O, O, K, I, E

♦**Difficulty : Intermediate**

SEOULTECH
Seoul National University of Science & Technology

Solutions of Final Exam

Date : 2016.12.19

Code	ITM 522	Title	Data Mining

♦ **Amount of work: 15min**

♦ **Solution :**

(1) Generate C_1 and L_1 (3pts)

Item	Number of transactions
M	3
O	3
N	2
K	5
E	4
Y	3
D	1
A	1
U	1
C	2
I	1

Because s_min is 0.6, we should eliminate item that occurs less than 3 transactions.

$$L_1 = \{\{M\},\{O\},\{K\},\{E\},\{Y\}\}$$

(2) Generate C_2 and L_2 (3pts)

$$C_2 = \{\{M,O\},\{M,N\},\{M,K\},\{M,Y\},\{O,K\},\{O,E\},\{O,Y\},\{K,E\},\{K,Y\},\{E,Y\}\}$$

Item Pairs	Number of transactions
MO	1
MK	3
ME	2
MY	2
OK	3
OE	3
OY	2
KE	4
KY	3
EY	2

Because s_min is 0.6, we should eliminate item that occurs less than 3 transactions.

$$L_2 = \{\{M,K\},\{O,K\},\{O,E\},\{K,E\},\{K,Y\}\}$$

(3) What is support and confident of rule "If K, then E" (3pts)

Support({K})=$\frac{5}{5} = 1$

Support({K,E})=$\frac{4}{5} = 0.8$

SEOULTECH
Seoul National University of Science & Technology

Solutions of Final Exam

Date : 2016.12.19

Code	ITM 522	Title	Data Mining

Confidence=$\frac{0.8}{1}$ = 0.8