Clemson University
School of Computing
CPSC6300 Section 002 - Applied Data Science
Final Exam on December 12, 2019
Duration: 2.5 hours (7:00PM-9:30PM)

Honor Policy.

- For this exam, you must **work independently**. You may not aid or accept aid from other students in any means.
- During the exam, you may consult two pages of double-sided notes which you have prepared before the exam. However, you must not consult any other paper or online materials.
- You are allowed to use a battery-powered calculator in the exam. However, you are not allowed to use the calculators on a smart phone, a laptop, or any other types of mobile devices that is not exclusively designed for calculator purpose.

Directions.

1. There are 18 pages total (including this page) in this exam. Check if you have all the pages.
2. Answer the questions on the exam pages in the space provided. You may use the back of the pages for scratch work but final answer must be in the provided spaces.
3. Read questions carefully and answer as neatly, clearly and concisely as possible; illegible or incomprehensible answers will not get any point.
4. Should a question be unclear or ambiguous, make a reasonable interpretation and state what you have assumed before answering.
5. Partial credit will be given for clear formulations of how to solve the problems.
6. Mysterious or unsupported answers will not receive full credit.
7. Answer all the questions, including all sub-parts.

Last Name: _____

First Name: _____

Signature: _____

Question:	1	2	3	4	5	6	7	Total
Points:	15	6	20	16	18	10	15	100
Score:								

1. (15 points) **Multiple choices question**. Choose the best answer for each question.

   (a) A hospital is building a new care unit for elderly patients. In order to streamline their services, the hospital administration is trying to place patients with similar medical conditions into rooms on the same floor. Overwhelmed with the amount of medical data that is available for each patient, they ask for your help to identify patient groups. What learning problem do you consider this problem to be?

   A. Supervised learning
   B. Classification
   C. Regression
   D. Clustering
   E. PCA

   (b) A consulting company collects data on the top 500 firms in the US. For each firm they record CEO salary, profit, number of employees, and industry. They ask you to build a data science model that explains CEO salary. What learning problem do you consider this task to be?

   A. Semi-supervised learning
   B. Classification
   C. Regression
   D. Clustering
   E. PCA

   (c) Which of the following metrics is most appropriate for linear model selection?

   A. R^2
   B. Training MSE
   C. t-statistic
   D. Mallow's C_p
   E. z-statistic

   (d) What does ***F*-Statistic** measure?

   A. The variability explained by the model.
   B. The variability left unexplained by the model.
   C. The variability explained by the model normalized by the number of predictors.
   D. The ratio of variability explained by the model to the variability left unexplained by the model.
   E. The statistically significance of a single variable.

   (e) Suppose you add another predictor to an existing linear regression model. What will happen to R^2 and adjusted R^2?

   A. R^2 will decrease and adjusted R^2 will either increase or decrease depending on how well the new predictor explains the response.
   B. R^2 will increase and adjusted R^2 will either increase or decrease depending on how well the new predictor explains the response.
   C. Adjusted R^2 will increase and R^2 will either increase or decrease depending on how well the new predictor explains the response.
   D. Adjusted R^2 will decrease and R^2 will either increase or decrease depending on how well the new predictor explains the response.
   E. The change of both R^2 and adjusted R^2 depends on how well the new predictor explains the response.

(f) Which statistic measure is more proper than others to represent a twitter text when you apply a Logistic Regression classifier to determine whether a twitter is positive, negative, or neutral?

A. Document frequency
B. Inverse document frequency
C. TFIDF
D. Term frequency
E. Term count

(g) In a random forests model, the algorithm selects a random sample of m predictors each time a split in a tree is considered. What is the main reason for this random sampling of predictors?

A. It combines decision trees with feature selection.
B. It reduces the complexity of the optimization problem.
C. It causes the tress estimated using different samples correlated with each other.
D. It is necessary when there are more predictors than observations.
E. It overcomes the problem in the decision tree method that most of the trees will be highly correlated when they all use the strong predictors in the top split.

(h) Suppose you have a data set with 100 observations and 200 features. You are tasked with performing feature selection. Which of the follow approaches is a good choice for this particular problem?

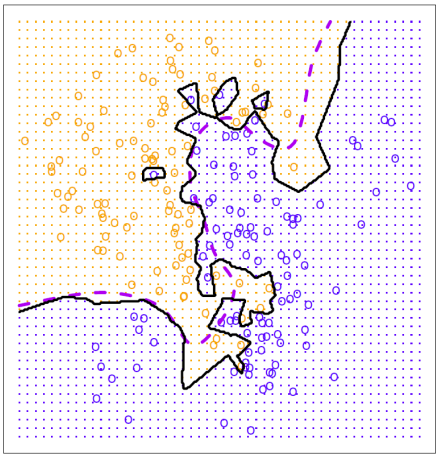
A. best subset selection.
B. forward stepwise selection.
C. backward stepwise selection.
D. ridge regression.
E. the lasso.

(i) Suppose you have trained an SVM classifier with linear decision boundary. After training the SVM, you have correctly inferred that your SVM model is under fitting. Which of the following option would you more likely to consider iterating SVM next time?

A. You want to increase your data points.
B. You want to decrease your data points.
C. You will try to calculate more variables.
D. You will try to reduce the features.
E. You will try to use a polynomial kernel instead of a linear kernel.

(j) Which of the following classifiers could have generated the decision boundary shown in the figure below?

A. Linear SVM.
B. KNN: K=1.
C. Logistic Regression.
D. LDA.
E. None of the above.



(k) Which of the following is **NOT** an assumption of linear regression model?

A. There exists a linear relationship between the predictors and the response variable.
B. The variance of error term is a constant.
C. The error term is uncorrelated across the observation.
D. The expected means of the error term is zero.
E. The least squares method is an unbiased estimation method for linear regression.

(l) Which of the following statements is **NOT** true?

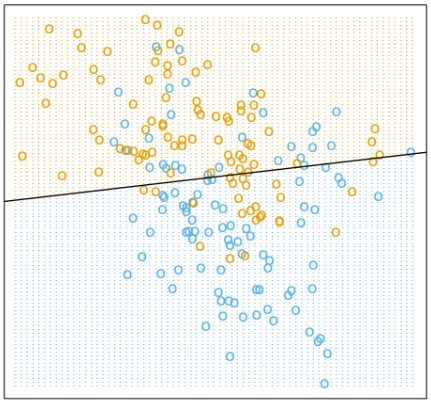
A. In a data science project, you should always use a more flexible model than a simple model because a flexible model has a lower bias and thus make more accurate prediction.
B. For KNN regression models, the flexibility of the model decreases as the model uses more nearest neighbors in its prediction.
C. In SVM based classifier, a linear kernel is preferred over a nonlinear kernel if the data set if linearly separable.
D. A natural spline model generally has a low variance than a higher order of polynomial regression model.
E. A shrinkage method like Lasso and ridge regression is biased towards shrinking the estimated coefficients towards zero.

(m) Which of the following problems can **NOT** be solved by a k-fold cross-validation?

A. Estimate the tuning parameter λ in the lasso method.
B. Determine the degree of polynomials in logistic regression.
C. Determine the number of principal components to be used in noise reduction.
D. Determine the turning parameter C in the optimization problem for SVM classifiers.
E. Estimate the test MSE in a subsect selection medthod.

(n) Which of the following classifiers could have generated the decision boundary shown in the figure below?

A. SVM with a radial kernel.
B. KNN.
C. LDA.
D. QDA.
E. Decision Tree.



(o) Which of the follow statement is **FALSE**?

A. Given a point $\mathbf{x}$ in the p-dimension space and a vector $\mathbf{u}$ that represents the direction of a principal component, the length of the project of $\mathbf{x}$ along $\mathbf{u}$ is $\mathbf{x} \cdot \mathbf{u}/\|\mathbf{u}\|^2$.
B. The principal components for a given data set aim to preserve the variance during a linear transformation. It also explains the variance of the data set.
C. PCA is susceptible to local optima.
D. PCA is an effective dimension reduction technique in regression analysis to address the curse of dimensionality problem.
E. PCA can be used for noise reduction in image preprocessing.

2. (6 points) For each of the following scenarios where machine learning might be applied, indicate what machine learning algorithm you would apply and why.

   - You may limit the choices to those algorithms we have discussed during the class.
   - If there are multiple proper algorithms, you can just pick up one and explain why that one is a proper one (not necessarily to the best one).
   - Points distribution: a proper algorithm (1 point) and explaining why it is proper (2 points).

   (a) (3 points) **Document Classification**. Your company has a large number of documents that need to be sorted into one of three categories: Research & Development, Finance, or Marketing. A staff at the CIO office has been able to identify a number of phrases which are commonly used in these documents. These phrases may help categorize the documents. However, there are a large number (thousands) of these phrases and each one only appears in a small number of documents. The staff has also labeled a few hundred of the documents. Now you are asked to develop a machine learning method to automatically label the rest.

   (b) (3 points) **Neighborhoods Identification**. Your friend wants to open an Asian restaurant in Atlanta and ask your help to identify a best location to open the restaurant. You and your friend both know that the best areas to run a restaurant are those locations which have the highest Asian population and in which the top common venues visited by the public include restaurants. Assume you have already created a Atlanta city data database. You database contains a directory of all neighborhoods (name, demographical distribution of the population) in Atlanta and a directory of all business venues (name, category, neighborhood, etc.) in Atlanta. Now, you want to build machine learning software that categorizes the neighborhoods into similar groups.

3. (20 points) Table 1 shows the first two rows of a preprocessed credit data set. This question is related to applying regression analysis to this data set.

Table 1: The credit card data set.

	Income	Limit	Rating	Cards	Age	Education	Gender	Student	Married	Balance
1	14.891	3606	283	2	34	11	1	0	1	333
2	106.025	6645	483	3	82	15	0	1	1	903

Table 2: Model Coefficients from a linear regression analysis.

	coef	std err	t	$P > \|t\|$	[0.025	0.975]
Intercept	-490.7701	39.562	-12.405	0.000	-568.636	-412.905
Income	-7.5851	0.272	-27.902	0.000	-8.120	-7.050
Limit	0.1558	0.038	4.049	0.000	0.080	0.232
Rating	1.5754	0.577	2.728	0.007	0.439	2.712
Cards	17.1923	5.207	3.302	0.001	6.945	27.440
Age	-0.5275	0.344	-1.534	0.126	-1.204	0.149
Education	-0.1654	1.840	-0.090	0.928	-3.788	3.457
Gender	9.4699	11.324	0.836	0.404	-12.818	31.758
Student	423.2841	19.934	21.234	0.000	384.050	462.518
Married	-9.3162	11.751	-0.793	0.429	-32.445	13.813

(a) (3 points) When apply multiple linear regression to the credit set to the training set and test set for multiple times, you observe that the mean training R^2 is 0.955, the mean test R^2 is 0.951, and the standard variances of both measures are very small. How do you explain this result (consistently high prediction accuracy measures which are statistically equivalent for both the training set and test set?

(b) (3 points) Based on the model coefficients in Table 2, **which predictor(s)** can be excluded from the linear regression model with little or no effect on the prediction accuracy and **why**?

Much like the best subset selection, the lasso performs variable selection. In you analysis, you have tested the following four lasso models:

```
models = {}
models['lasso-default'] = Lasso(max_iter=5000)
models['lasso-normalized'] = Lasso(max_iter=5000, normalize=True)
models['lassoCV-default'] = LassoCV(max_iter=5000)
models['lassoCV-normalized'] = LassoCV(max_iter=5000, normalize=True)
```

Table 3: Coefficient estimates and model performance

	lasso-default	lasso-normalized	lassoCV-default	lassoCV-normalized
alpha	1.0	1.0	872.77	0.059
Intercept	-491.80	-427.35	-351.44	-488.94
coef. Age	-0.52	-0.0	-0.0	-0.47
coef. Cards	16.35	0.0	0.0	16.32
coef. Education	-0.0	0.0	0.0	-0.0
coef. Gender	5.24	0.0	-0.0	7.19
coef. Income	-7.58	-5.18	-5.23	-7.44
coef. Limit	0.15	0.08	0.23	0.15
coef. Married	-5.89	-0.0	-0.0	-7.86
coef. Rating	1.62	2.15	0.0	1.57
coef. Student	410.60	351.91	0.0	418.96
train R^2	0.95	0.93	0.86	0.95
test R^2	0.96	0.92	0.84	0.95
train MSE	9177.22	12854.62	26889.63	9171.23
test MSE	11080.05	19275.55	39044.841	11238.40

(c) (2 points) From the coefficient estimates from the lasso-normalized model, we can write the following regression model for balance prediction:

$$Balance = -427.35 - 5.18 \times Income + 0.08 \times Limit + 2.15 \times Rating + 351.91 \times Student + \epsilon$$

Which of the following interpretations of the model is (are) correct?

A. Every \$1,000 increase in Income is associated with an average \$5.18 decrease in balance.
B. The average balance for students is \$351.91.
C. On average, the balance of a student is \$351.91 more than that of a non-student.
D. There is no relationship between balance and limit because the coefficient for the limit is so small.

(d) (2 points) Based on Table 3, does marriage status have any effect on the balance? If yes, on average, does marriage lead to balance increase or decrease?

(e) (3 points) Write down the constrained optimization problem for the lasso and explain how the lasso is connected to the best subsection selection.

(f) (5 points) As a general trend, the training MSE decreases monotonically as the model flexibility increases, and there is a U-shape in the test MSE. The U-shape observed in the test MSE curves turns out to the be result of two competing properties of statistical leaning methods: variance and bias. Using the lasso as example, discuss how bias, variance, and MSE vary with the tuning parameter λ.

(g) (2 points) List one regression model that can output the feature importance as a by-product of its regression analysis.

4. (16 points) Consider the Default data in which the first 4 samples are shown in the table below. We are interested in creating a model to predict whether an individual will default (Y) on his or her credit card payment, on the basis of monthly credit card balance (X_1), annual income (X_2) and student status (X_3).

Table 4: The Default data set

	balance	income	student[Yes]	default
1	729.53	44361.63	0	No
2	817.18	12106.13	1	No
3	1073.55	35704.49	0	No
4	2529.25	38463.50	0	Yes

(a) (2 points) Explain why a simple linear regression model is not appropriate for the task of predicting the default using the predictors X_1, X_2 and X_3.

(b) (2 points) Write the logistic regression model for predicting the default using the three predictors X_1, X_2 and X_3. You can use β_i to represent the model coefficient for the feature X_i.

(c) (2 points) The textbook mentioned that the coefficient of the logistic regression model can be estimated using a method called **maximum likelihood**, which essentially finds the set of coefficients that maximize a likelihood function. Write down the likelihood function for the logistic regression model you chose in (b).

For the question described above, you have estimated the logistic regression model using the following Python code and the results are shown in Table 5.

```
# Python code to estimate the logistic regression model
import statsmodels.api as sm

X = default_df[['balance', 'income', 'student[Yes]']]
y = default_df['default'].factorize()[0]
X_train = sm.add_constant(X.values)
model = sm.Logit(y, X_train).fit()
model.summary2().tables[1]
```

Table 5: The estimated coefficient of the logistic regression model for the Default data set.

	Coef.	Std.Error.	z	P-value
const	-10.869045	0.492273	-22.08	
X1	0.005737	0.000232	24.74	
X2	0.000003	0.000008	0.37	
X3	-0.646776	0.236257	-2.74	

For the following two question, if you don't have a calculator, you can write down the formula. You will receive the partial credit if your formula is correct.

(d) (2 points) Using the above estimated model to predict the probability of default for a **non-student** with a credit card balance of \$2,000 and an income of \$38,000.

Table 6: A confusion matrix comparing the LDA predictions to the true default status.

		True Default Status		
		No	Yes	Total
Predicted Default Status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

(e) (2 points) Using the above confusion matrix, calculate the accuracy and sensitivity of the LDA classifier?

(f) (3 points) Consider the above confusion matrix, explain why LDA does such a poor job of classifying the customers who default? How do you improve the prediction performance by making a small change on how LDA makes the classification decision?

(g) (3 points) Unlike logistic regression, LDA is linked to Bayes's Theorem and derives a discriminant function from the posterior probability that an observation belongs to a certain class based upon two major assumptions. What are the two assumptions of LDA?

5. (18 points) **Email Spam Prediction**. We have the following email spam prediction data set. The data set comprises of three features–suspicious words (X_1), unknown senders (X_2), contains images (X_3)–and a class label (Y) (spam or ham). This question tests Naive Bayes and decision tree classifiers for the email spam prediction problem.

Table 7: An email spam prediction data set

ID	Suspicious Words	Unknown Senders	Contains Images	Class
376	true	false	true	spam
489	true	true	false	spam
541	true	true	false	ham
693	false	true	false	ham
782	false	false	false	ham
976	false	false	false	ham

(a) (2 points) Apply the Bayes's theorem to the email spam problem, we have the following equation

$$P(Y = l|X_1, X_2, X_3) = \frac{P(X_1, X_2, X_3|Y = l) \times P(Y = l)}{P(X_1, X_2, X_3)}$$

In this equation, which term is the likelihood, prior probability, and posterior probability?

Term	Write the Appropriate Term Name in this Column
$P(Y = l\|X_1, X_2, X_3)$	
$P(X_1, X_2, X_3\|Y = l)$	
$P(Y = l)$	
$P(X_1, X_2, X_3)$	

(b) (2 points) A probability based statistical learning methods relies on estimations of several probabilities appeared in the Bayes's theorem. What are the two fundamental principles (i.e. methods) for probability estimation?

(c) (4 points) Naives Bayes uses the production rule

$$P(X_1, X_2, \ldots, X_p|Y = l) = \prod_{i=1}^{p} P(X_i|Y = l)$$

to compute the conditional joint probability $P(X_1, X_2, \ldots, X_p|Y = l)$. Prove that the production rule holds under the conditional independence assumption on the predictors.

(d) (3 points) Using the production rule and the Bayes' Theorem to compute the probability

$P($`Y=spam|Suspicious Words=true, Unknown Senders=false, Contains Image=true`$)$

In your computation, you can assume

$$P(\texttt{Suspicious Words=true, Unknown Senders=false, Contains Image=true}) = \frac{1}{6}$$

(e) (5 points) A decision tree classifier recursively splits the data set into sub data sets using the features that gives the maximum information gain. **Compute the information gain** obtained by splitting the data based on the **Suspicious Words** feature. In this question, assume the information measure is Shannon's entropy. For your reference, below are the equations to compute the entropy of a data set and the remainder after split using a feature.

$$H(y, D) = -\sum_{l \in Classes(y)} P(y = l) log_2(P(y = l))$$

.

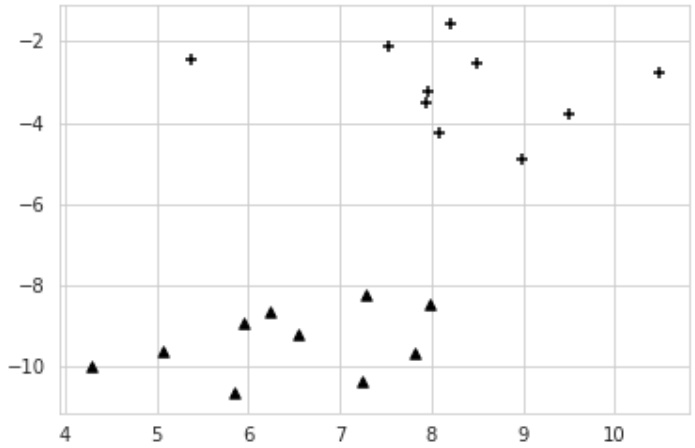
$$Rem(d, D) = -\sum_{l \in Classes(d)} \frac{\|D_{d=l}\|}{\|D\|} \times H(y, D_{d=l})$$

.

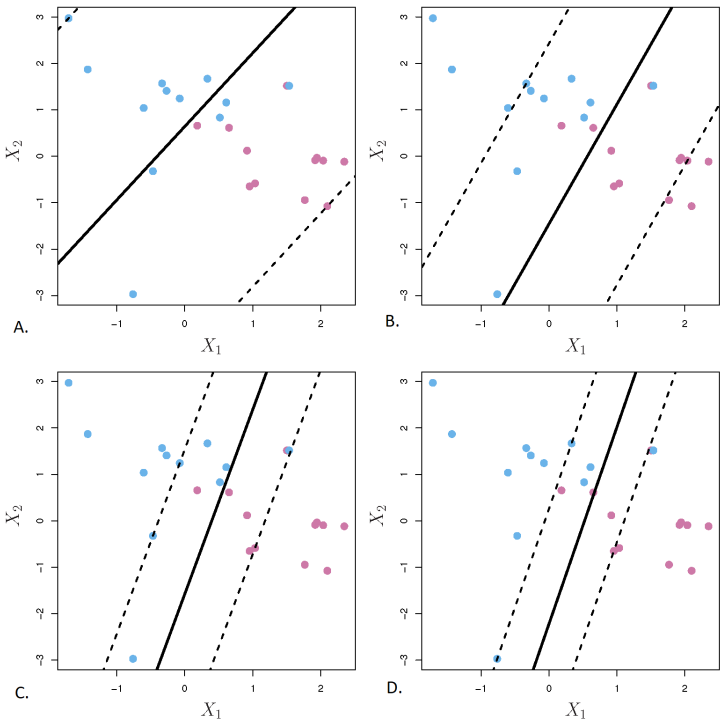
(f) (2 points) A decision tree classifier suffers from high variance, i.e., different training set can result in different trees. Provide a learning method that has lower variance in comparison with the decision tree classifier.

6. (10 points) This question tests your understanding of SVM.

   (a) (2 points) Give a data set shown in Figure below, draw the separating hyperplane resulted from an SVM classifier and circle all the support vectors.

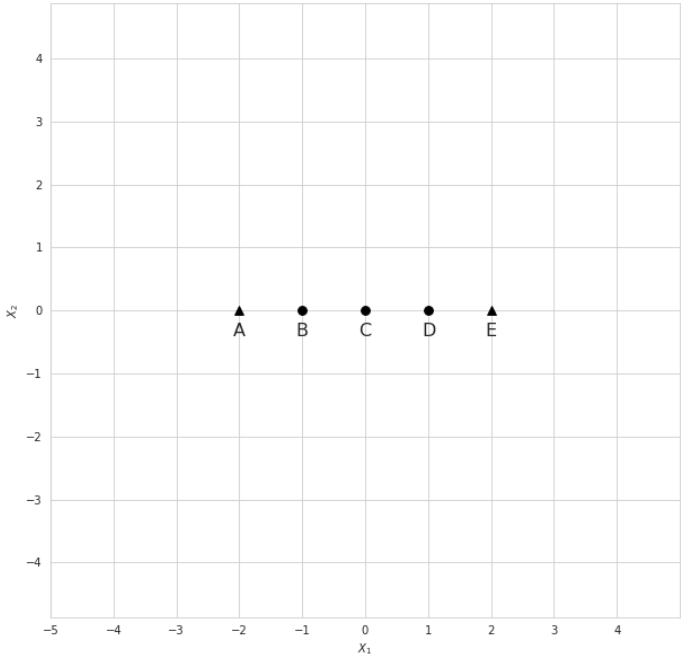


   (b) (2 points) The plots below show a Support Vector Classifier that was fit on the same data using four different values for the tuning parameter C. Which of the classifiers has the smallest bias but potentially largest variance?



(c) (3 points) Show that $K(u, v) = (u^T \cdot v + 1)^2$ is a kernel that corresponds to the mapping $\phi : x \mapsto (x^2, \sqrt{2}x, 1)^T$, where x, u and v are all vectors in 1D space. In other words, show $K(u, v) = \phi(u)^T \cdot \phi(v)$.

(d) (3 points) In the figure below, sketch the corresponding images (i.e., $\phi(x)$ for point x) of the points A-E in the 1D space in the 2D space using the mapping $\phi : x \mapsto (x^2, \sqrt{2}x)^T$. **In the 2D space**, sketch the separating hyperplane and circle all the support vectors.



7. (15 points) This problem tests your understanding of the **K-Means Clustering**.

   (a) (2 points) Generally an unsupervised learning algorithm aims to solve an optimization problem. What objective function does a K-means clustering seek to minimize?

   (b) (3 points) Sketch the K-means algorithm below and mark the key steps and termination condition(s).

   (c) (3 points) Does the K-means algorithm always converge? Explain why?

(d) (4 points) In this problem, you will perform K-means clustering manually, with $K = 2$, on a small data set with $n = 5$ observations and $p = 2$ features. The observations are as follows.

Obs.	X_1	X_2
1	1	4
2	1	3
3	0	4
4	5	1
5	4	0

You label the two clusters with + and −. In a particular run of the K-means algorithm, you assign observation 1 to cluster + and observation 2 to cluster −.

i. In the following table, assign a cluster label to all the observations.

Obs.	X_1	X_2	Cluster
1	1	4	+
2	1	3	−
3	0	4	
4	5	1	
5	4	0	

ii. after step i, the coordinates of the new centroid for cluster + will be

(_____, _____).

(e) (3 points) List three caveats of the k-means algorithm.