Ecole Nationale Supérieure d'Ingénieurs de Tunis

EXAMEN

Classe: 3GMAM **Matière:** Data Mining
Enseignant~~(e)~~: Mohamed Koutheair Khribi

Date: 05/01/2019 **Durée:** 02h00
Nombre de pages: 10
Documents autorisés: Non ☑ Oui ☐

- *This exam has 10 pages, make sure you have all pages before you begin.*
- *There are 06 multipart questions, for a total of 20 points.*
- ***Please write your answer on the provided exam.***
- *Read all the questions before you start working and write concise answers.*
- *No materials or electronic devices shall be used.*
- *No partial credit on multiple-choice questions, the set of all correct answers must be checked.*

Good Luck!

Questions	Question I	Question II	Question III	Question IV	Question V	Question VI	Total
Points	7	3.5	3	3	2.5	1	20
Score							

Question I. (True/False) [7 points]:

1) In perceptron algorithm, choosing large learning rate will speed up the algorithm and let previously incorrect classifications becoming correct.	True	False
2) The perceptron algorithm does not converge if the training samples are not linearly separable.	True	False
3) In standardization, rescaled features are centered to the mean $\mu = 0$ with $\sigma = 1$.	True	False
4) In Min-Max and unit vector normalization, values are converted to the range [0,1].	True	False
5) In underfitting, the model suffers from low performance on training and test data.	True	False
6) In overfitting, the model performs well on training data but suffers from low performance on test data.	True	False
7) Adding a feature to a linear regression model can decrease model variance.	True	False
8) After applying a regularization penalty in linear regression, some of the coefficients can be zeroed out.	True	False
9) In regression analysis, residuals should be maximized to let the line fitting better the sample points.	True	False
10) LASSO Regularization can be used for variable selection in Linear Regression.	True	False
11) In optimization using gradient descent for linear regression, the cost function has many local optima and converges for one of them.	True	False
12) We can get multiple local optimum solutions if we solve a linear regression problem by minimizing the sum of squared errors using gradient descent.	True	False
13) In regularization, additional information is added in order to penalize extreme parameter weights.	True	False
14) When the feature space is larger, over fitting is more likely.	True	False
15) Increasing regularization parameter can help to fix high variance.	True	False
16) Decreasing regularization parameter can help to fix high bias.	True	False
17) Reducing the number of features can help in addressing high variance problem.	True	False
18) Zero correlation between two random variables implies that the two variables are independent.	True	False
19) Overfitting is more likely when we have huge amount of data to train.	True	False
20) kNN is a parametric method that performs well in case of high dimensional data.	True	False
21) Training a k-nearest-neighbors classifier takes more computational time than applying on new samples.	True	False
22) The more training examples, the more accurate the prediction of a k-nearest-neighbors.	True	False
23) k-nearest-neighbors cannot be used for regression.	True	False
24) Given n distinct classes in a dataset, all-vs-all classification builds $n(n-1)/2$ classifiers.	True	False
25) In regression analysis, logistic regression can be used as the learning function outputs continuous values.	True	False
26) Using slack variables can help to reduce overfitting in an SVM classifier.	True	False
27) Choosing large value of the tuning parameter C in SVM can lead to low bias and high variance.	True	False
28) Choosing small value of the tuning parameter C in SVM can lead to high bias and low variance.	True	False

Question II. *Select One or More Answer Choices [3.5 points]:*

1. Regularization: (Check all that apply)

   (a) Consists in penalizing the training error.

   (b) Used to avoid overfitting.

   (c) Equivalent to adjusting the bias/variance trade-off.

   (d) Tune the complexity of the model.

2. Lasso can be interpreted as least-squares linear regression where: *(Check all that apply)*

   (a) Weights are regularized with the L1 norm.

   (b) Weights are regularized with the L2 norm.

   (c) Weights often end up to being exactly zero.

   (d) Particular features are excluded from the model.

3. How does the bias-variance decomposition of a ridge regression estimator compare with that of ordinary simple linear regression? (Select one)

   (a) Ridge has larger bias, larger variance

   (b) Ridge has larger bias, smaller variance

   (c) Ridge has smaller bias, larger variance

   (d) Ridge has smaller bias, smaller variance

4. Assume we have N independent explanatory variables $(X_1, X_2 \ldots X_n)$ and a target dependent variable Y. When applying linear regression by fitting the best fit line using least square error on this data, we found that correlation coefficient for one of its variables (e.g. X_i) with Y is -0.95. Which of the following is true for X_i?

   (a) Relation between the X_i and Y is weak

   (b) Relation between the X_i and Y is strong

   (c) Relation between the X_i and Y is neutral

   (d) Correlation can't judge the relationship

5. Given two variables V_1 and V_2 following the below two characteristics.
   - If V_1 increases, then V_2 also increases
   - If V_1 decreases, then V_2 behavior is unknown.

   Which of the following option is the correct for Pearson correlation between V_1 and V_2?

   (a) Pearson correlation will be close to 1.

   (b) Pearson correlation will be close to -1.

   (c) Pearson correlation will be close to 0.

   (d) None of these.

6. The coefficients of linear regression can be computed with the help of an analytical method called "Normal Equation" characterized by the following:
    1. The learning rate has not to be chosen.
    2. The method becomes slow when number of features is very large.
    3. There is no need to iterate.

    Which of the following is/are true?

    (a) 1 and 3.

    (b) 2 and 3.

    (c) 1,2 and 3.

7. Suppose we have fitted a Ridge regression with penalty ε model on a dataset. Choose the option which describes bias in best manner:

    (a) In case of very large ε; bias is low.

    (b) In case of very large ε; bias is high.

    (c) We can't say about bias.

    (d) None of these.

8. When we increase the size of the training dataset, what do we expect to get as for bias and variance:

    (a) Bias increases and variance increases.

    (b) Bias decreases and variance increases.

    (c) Bias decreases and variance decreases.

    (d) Bias increases and variance decreases.

9. Assume we train a hard-margin linear SVM on $n > 100$ data points in R^2, yielding a hyperplane with exactly 2 support vectors. If we add one more data point and retrain the classifier, what is the maximum possible number of support vectors for the new hyperplane (assuming the n+1 points are linearly separable)?

    (a) 2 (b) 3 (c) n (d) $n + 1$

10. Suppose we have got to add a variable in variable space such that this added feature is very important. Which of the following options would we expect to observe in such case?

    (a) Training Error will decrease and validation error will increase

    (b) Training Error will increase and validation error will increase

    (c) Training Error will increase and validation error will decrease

    (d) Training Error will decrease and validation error will decrease

    (e) None of the above

11. Which of the following methods can achieve zero training error on any linearly separable dataset?

(a) Adaline (b) Hard-margin SVM (c) 15-nearest neighbors (d) Perceptron

12. We run gradient descent for 15 iterations with $\eta = 0.3$ and compute the cost function $J(w)$ after each iteration, we find that its value decreases slowly and it is still decreasing after 15 iterations. Based on this, which of the following conclusions seems to be more plausible:

(a) We should try a larger value of η (e.g. $\eta = 1.0$) rather than the current value.

(b) The current value is an effective choice.

(c) It would be more promising to try smaller value of η (e.g. $\eta = 0.1$) rather than the current value.

13. Assume we have been given the following scenario for training and validation error for linear regression:

Scenario	*Learning rate*	*Number of iterations*	*Training error*	*Validation error*
1	0.1	1000	100	110
2	0.2	600	90	105
3	0.3	400	110	110
4	0.4	300	120	130
5	0.4	250	130	150

Which of the following scenario would give us the right hyperparameter:

(a) 1 (b) 2 (c) 3 (d) 4 (e) 5

14. *Considering a classifier trained till convergence on some training data D_{train}, and tested on a separate test set D_{test}. We look at the test error and find that it is very high. We then compute the training error and find that it is close to 0. Which of the following is expected to help? (Select all that apply).*

(a) Increase the training data size.

(b) Decrease the training data size.

(c) Increase model complexity (e.g. in case of an SVM, we use a more complex kernel).

(d) Decrease model complexity.

(e) None of the above

Question III. [3 points]:

(a) The following table gives the decision table for "OR" function. x_1 and x_2 are the inputs, and Y the output. Can this function be represented by a perceptron? Explain your answer.

x_1	x_2	Y
0	0	0
0	1	1
1	0	1
1	1	1

.....

.....

.....

(b) If yes, give a perceptron that predicts this function (give one possible solution determining w_0, w_1, w_2). Plot the points and the decision boundary.

.....

.....

.....

.....

.....

(c) Consider the function given in the following table. x_1 and x_2 are the inputs, and Y the output. Can this function be represented by a perceptron? Explain your answer.

x_1	x_2	Y
0	0	1
0	1	1
1	0	0
1	1	1

.....

.....

.....

(d) If yes, draw a perceptron that represents this function (give one possible solution determining w_0, w_1, w_2). Plot the points and the decision boundary.

.....

.....

.....

.....

.....

.....

.....

(e) Consider the function given in the following table. Can this function be represented by a perceptron? Explain your answer.

X_1	X_2	Y
0	0	0
0	1	1
1	0	1
1	1	0

Question IV. [3 points]:

(a) Consider fitting a linear regression model for these data:

x	y
-1	1
0	-1
2	1

Fit $y_i = w_0 + \varepsilon_i$ model (intercept only), find w_0 using MSE as cost function (Hint: derive the function and solve the equation).

(b) Fit $y_i = x * w_1 + \varepsilon_i$ model (linear regression without the intercept term), find w_1 using MSE as cost function (Hint: derive the function and solve the equation).

(c) Assume we are training a model on a training dataset containing 165 examples, using a binary classifier, we've got 100 true positives predictions, 50 true negatives, 10 false positives, and 5 false negatives. Complete the confusion matrix below:

N = ...	Predicted Yes	Predicted No
Actual Yes	....	....
Actual No	....	....

(d) Based on the above confusion matrix, compute the following evaluation metrics:

Accuracy rate: .....

Error rate: .....

False positive rate: .....

True negative rate: .....

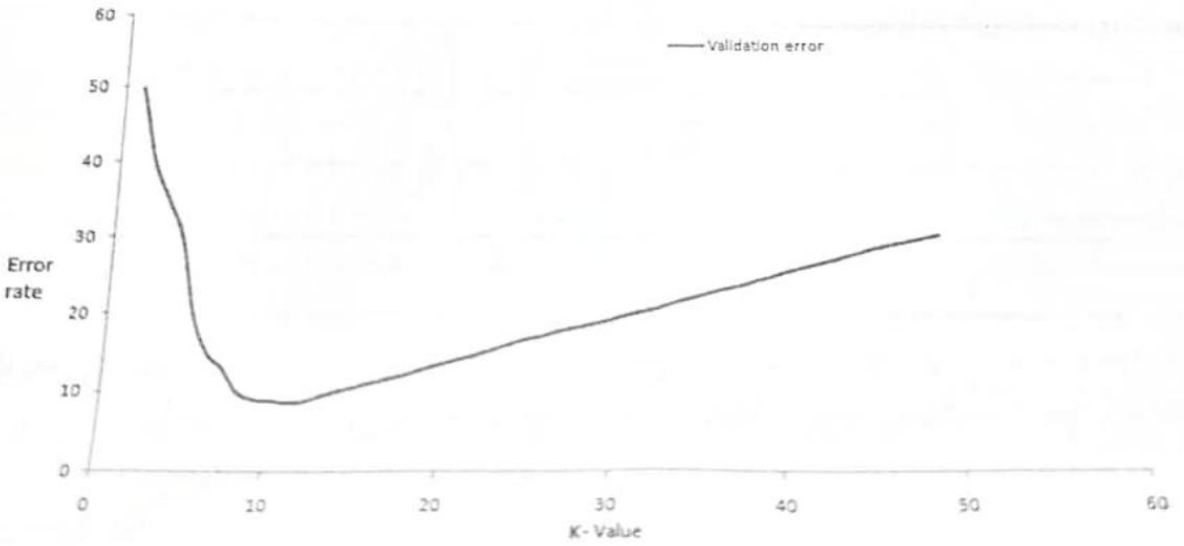
Recall: .....

Precision: .....

F1-score: .....

estion V. Interpreting figures [2.5 points]:

(a) Consider the following figure showing the plot of the validation error rate in function of the value of k when applying kNN algorithm:

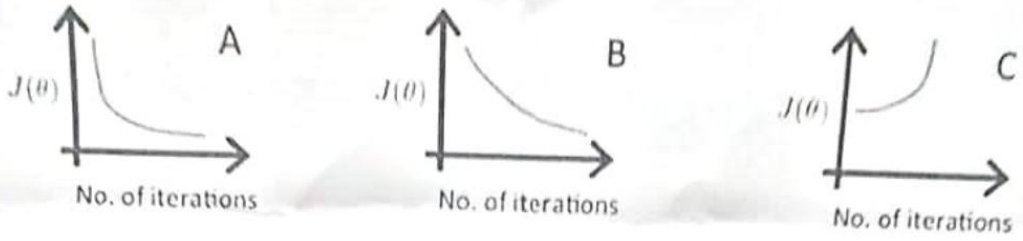


What which would be the best value for k? Explain.

.....

.....

(b) Consider the following plots showing the cost J in function of the number of iterations:



Suppose η_1, η_2 and η_3 are the three learning rates for A, B and C respectively, compare the values of the learning rates η_1, η_2 and η_3. Explain.

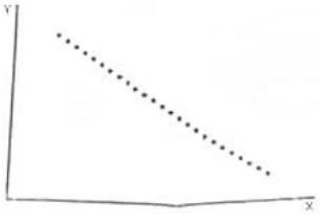
.....

.....

.....

.....

(c) Consider the following scatterplot:



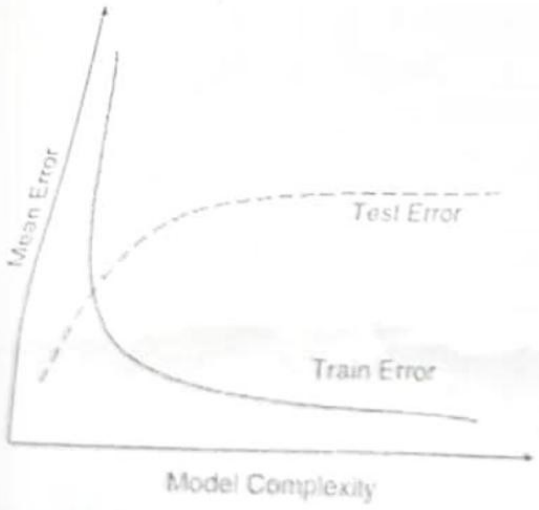
What would be the mean squared errors for this data if we run a linear regression model of the form $y = w_0 + x * w_1$? Explain.

.....

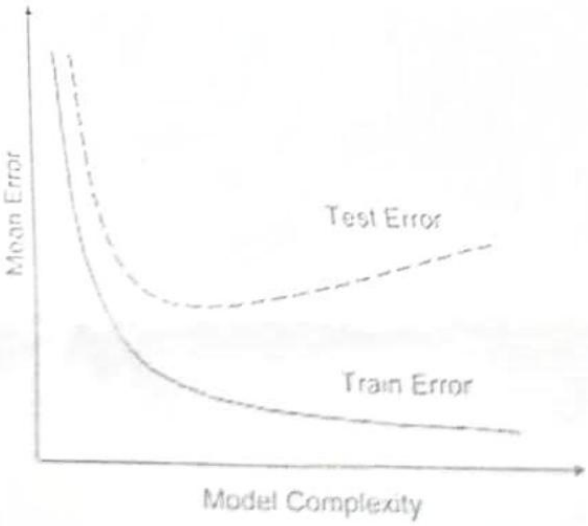
.....

.....

(d) Assume that we will debug a classifier using train and test errors. We found that the test error is very high, and the training error is close to zero. Which of the following plots showing the mean error as function of the model complexity is the plot expected to look like. Explain.



(a)



(b)

.....

.....

.....

.....

.....

.....

Question VI. Gift [1 point]:

(a) How difficult was this exam?

Easy									*Hard*
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*

(b) Do you have any comments on the number of questions or their quality?

.....

.....

.....

.....

(c) How difficult is this course?

Easy									*Hard*
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*

(d) Do you have any comments on the given lectures, the material covered, or on the course more generally?

.....

.....

.....

.....

Thanks and Good Luck !