Final Integrated Exam - Part 2: Machine Learning (Version : 1)

TEST

Correct Answer

Answered in 31.516666666667 Minutes

Uploaded File : 1652478705-Final_Integrated_Exam_-_Part_2:_Machine_Learning-5665-Eric_Mbuthia.Zip

Question 1/53

Regression (20 Questions, 30 Marks)
Questions 1 - 20

For this section please follow the steps below before attempting the questions.

- Load the data set titled ‘rand-dollar.csv’ as follow:
  - pd.read_csv(‘rand-dollar.csv’, index_col=0)
- Separate the data set into X (features) and y (targets)
- Our target variable will be ZAR/USD, with all other variables being the predictors
- Create an 80/20 split between train and test sets
- The training data should be the first 80% of the data, with the final 20% being used in the test set:
  - train_test_split(X, y, test_size=0.2, shuffle=False)
- Train a simple linear regression model to predict the ‘ZAR/USD’ using only ‘Value of Exports (ZAR)’ as the predictor variable:
  - from sklearn.linear_model import LinearRegression
  - lm = LinearRegression( )
  - lm.fit(...)

What is the value of the intercept of the model?

- [ ] -1.24
- [ ] 8.67 e-5
- [ ] 3.99
- [x] 3.29

Question 2/53

How do we interpret the intercept?

- [ ] The value of ZAR/USD will be equal to zero when exports are equal to this value
- [x] The value of ZAR/USD will be equal to this value when exports are zero
- [ ] This is the average value of ZAR/USD
- [ ] The maximum value of exports is equal to this value

Question 3/53

What is the value of the slope of this model?

- [ ] 86 800 000
- [x] 8.68 e -5
- [ ] 3.29
- [ ] 8.68

Question 4/53

How do we interpret the slope of the model?

- [ ] A decrease of 1 unit in ZAR/USD results in an increase of this many units in exports

- [x] An increase of 1 unit in exports results in an increase of this many units in ZAR/USD
- [ ] A decrease of this many units in exports results in an increase of 1 unit in ZAR/USD
- [ ] An increase this many units in exports results in an increase of 1 unit in ZAR/USD

Question 5/53

What is the predicted value of the exchange rate in a month where exports total R100 000?

- [ ] $11.97 / R1.00
- [ ] R4.16 / $1.00
- [ ] R90.07 / $1.00
- [x] R11.97 / $1.00

Question 6/53

What is the MSE of the model on the test set?

- [x] 8.22
- [ ] -8.45
- [ ] 6

- [ ] 1.89

Question 7/53

What is the R-squared value of the model on the test set?

- [ ] -4.25
- [x] -8.45
- [ ] 8.22
- [ ] 9

Question 8/53

What would a negative R-squared test value imply?

- [ ] The data needed to be standardized before the model was trained
- [ ] R-squared has no implication when using a test set
- [x] The model is worse at predicting the target variable than if a constant line y = intercept was used
- [ ] The relationship between predictor and response is negative

Question 9/53

What is the predicted value for August 2017?

- [ ] $12.25 / R1.00
- [ ] R7.13 / $1.00
- [x] R12.25 / $1.00
- [ ] R16.51 / $1.00

Question 10/53

What is the absolute error for this prediction?

- [x] R0.98
- [ ] R0.49
- [ ] R1.98
- [ ] R2.42

Question 11/53

Now revert back to using the full original data set.

Use the df.corr() function to find the correlations between the predictors and the target variable.

Which variable has the weakest linear relationship with the ZAR/USD exchange rate?

- [ ] Lending Rate
- [ ] Consumer Price Index
- [ ] Claims on Non-residents

- [x] Savings Rate

Question 12/53

Which variable has the strongest linear relationship with the ZAR/USD exchange rate?

- [ ] Claims on Non-residents
- [ ] IMF Reserve Position (USD)
- [ ] Savings Rate
- [x] Consumer Price Index

Question 13/53

Before answering the following questions, make sure to perform the steps outlined below:

- Split the original DataFrame into X (features) and y (targets)
- Standardise the entire X matrix
- Create X_train, X_test, y_train, y_test using the same chronological 80/20 split as before
- Train two models, “ridge” and “lasso”, which use ridge regression and LASSO, respectively (in the case of the lasso model, set alpha=0.01, use default parameters for the ridge model)

What is the training MSE of the Ridge model?

- [ ] 0.047
- [x] 0.040
- [ ] 0.579
- [ ] 0.632

Question 14/53

What is the training MSE of the LASSO model?

- [ ] 0.632
- [ ] 0.040
- [x] 0.047
- [ ] 0.579

Question 15/53

What is the testing MSE of the Ridge model?

- [ ] 0.047
- [ ] 0.579
- [ ] 0.640
- [x] 0.632

Question 16/53

What is the testing MSE of the LASSO model?

- [ ] 0.632
- [ ] 0.640

- [ ] 0.047
- [x] 0.579

Question 17/53

Based on the values of the Ridge model's variable coefficients, which indicator is the best predictor of the target variable?

- [ ] Value of Imports (ZAR)
- [x] Value of Exports (ZAR)
- [ ] Government Bonds
- [ ] Liabilities to Non-residents (USD)

Question 18/53

Based on the values of the Ridge model's variable coefficients, which indicator is the worst predictor of the target variable?

- [x] Liabilities to Non-residents (USD)
- [ ] Value of Imports (ZAR)
- [ ] Value of Exports (ZAR)
- [ ] Government Bonds

Question 19/53

Based on the values of the LASSO model's variable coefficients, which indicator is the best predictor of the target variable?

- [ ] Government Bonds
- [ ] Value of Exports (ZAR)
- [ ] Liabilities to Non-residents (USD)
- [x] Value of Imports (ZAR)

Question 20/53

How many variables have coefficients equal to zero in the LASSO model?

- [ ] 3
- [x] 4
- [ ] 9
- [ ] 0

Question 21/53

Classification (21 Questions, 41 Marks) Questions 21 - 41

The next 10 questions (Question 21 - 30) are based on the medical claims dataset (claims_data.csv)

What proportion of individuals in this dataset would be classified as overweight or obese (BMI of greater than 25)?

- [ ] 73%

- [ ] 18%
- [x] 82%
- [ ] 85%
- [ ] 15%

Question 22/53

Is the Poisson distribution a good choice to model the distribution of the number of children in this dataset?

- [x] No, the variance is significantly higher than the mean, suggesting overdispersion relative to the Poisson distribution.
- [x] Yes, the Poisson is a good choice for count data such as the number of children in a family.
- [ ] No, the variance is significantly lower than the mean, suggesting underdispersion relative to the Poisson distribution.
- [ ] No, because the Poisson only applies to positive integers, so cannot accommodate observations with 0 children.
- [ ] No, the Poisson is inappropriate, as it is a continuous distribution while the number of children is a discrete variable.

Question 23/53

If we assumed that age of this group was normally distributed, then given the mean and standard deviation of age in the data set, calculate the number of individuals we would expect to be aged 60 or older.

- Use 60 exactly as the cutoff point on the distribution, and round to the nearest integer.
- Then compare this with the number actually aged 60 or older.

Which of the following is true?

- [x] There are 21 fewer individuals 60 or older than the normal distribution would suggest.
- [ ] There are 7 fewer individuals 60 or older than the normal distribution would suggest.
- [ ] There are 21 more individuals 60 or older than the normal distribution would suggest.
- [ ] The two are exactly equal!
- [ ] There are 7 more individuals 60 or older than the normal distribution would suggest.

Question 24/53

Create a joint plot on the age and BMI variables. What summarises best what you see?

- [ ] There is not an easily discernible pattern in the plot, but the correlation coefficient is 0.11, which is statistically significantly different from zero, suggesting that older people tend to have lower BMIs.
- [ ] There is not an easily discernible pattern in the plot, but the correlation coefficient is 0.11, which is statistically significantly different from zero, suggesting that older people tend to have higher BMIs.
- [ ] There is a clearly discernible pattern in the plot, with a tight clustering and downward

- [ ] trend showing that older people tend to have lower BMIs, confirmed by a correlation coefficient of 0.11, which is statistically significantly different from zero.

- [x] There is not an easily discernible pattern in the plot, and the correlation coefficient is 0.11, which is not statistically significantly different from zero.

- [ ] There is a clearly discernible pattern in the plot, with a tight clustering and upward trend showing that older people tend to have higher BMIs, confirmed by a correlation coefficient of 0.11 which is statistically significantly different from zero.

Question 25/53

Use the appropriate model from the sklearn library (with default parameters unless specified otherwise) to fit a logistic regression model to the data, with insurance_claim as your target variable, using all other fields apart from claim_amount and creating dummy variables for the categorical variables in the data, dropping the first in each instance.

- Do a test-train split holding out 33% of the data for the test set, using a random seed of 42 for the split.
- Convert your target variable to a binary 0 or 1, where 1 indicates that there was a claim.

What proportion of claim indicators in the test set are correctly predicted?

- [ ] 86%

- [x] 85%

- [ ] 87%

- [ ] 81%

- [ ] 91%

Question 26/53

Now fit another logistic regression, this time using the statsmodels library to do so, with default parameters.

Be sure to add a constant to your X matrices, both train and test (you might want to check the statsmodels documentation for the add_constant function).

Which of the following best summarises the results?

- [ ] Age, sex, BMI, number of children and smoker status significantly affect the likelihood of an insurance claim. Number of steps is not significant. There appear to be some regional effects, but these are not strongly significant.
- [ ] Age, sex, BMI and smoker status significantly affect the likelihood of an insurance claim. Number of steps and number of children are not significant. There appear to be some regional effects, but these are not strongly significant.
- [ ] Apart from age, BMI and smoker status, none of the other features are statistically significant predictors of an insurance claim.
- [ ] All features seem to be significant predictors of the likelihood of a claim.
- [x] Age, BMI, number of children and smoker status significantly affect the likelihood of an insurance claim. Number of steps and sex are not significant. There appear to be some regional effects, but these are not strongly significant.

Question 27/53

What is the primary reason for using random forests instead of a single decision tree?

- [ ] There is much more firewood in a whole

forest than in a single tree.

- [ ] There is no good reason in principle to prefer a random forest over a single decision tree; it depends on the data.
- [ ] Decision trees suffer from high bias, and random forests reduce this, but at the expense of higher variance.
- [x] Decision trees suffer from high variance, and random forests reduce this variance by averaging multiple trees, each fitted to a subset of the observations and ensuring these trees are decorrelated by using only a subset of the available predictors.
- [ ] Decision trees suffer from high variance, and random forests reduce this variance by averaging multiple trees, each fitted to a subset of the observations.

Question 28/53

Now fit a random forest with 100 trees and a random seed of 101, and default parameters for the rest.

Which of the following sets out the number of false negatives and false positives in the confusion matrix on the test data?

- [ ] FN = 245, FP = 177
- [x] FN = 8, FP = 5
- [ ] FN = 5, FP = 8
- [ ] FN = 12, FP = 12
- [ ] FN = 177, FP = 245

Question 29/53

Fit Support Vector Machine models to the training data, using respectively the radial, sigmoid and linear kernels with default parameters.

Which model yields the best accuracy on test data?

- [x] Linear
- [ ] Radial
- [ ] Radial and linear yield very similar accuracies
- [ ] Sigmoid
- [ ] Sigmoid and linear yield very similar accuracies

Question 30/53

With respect to a SVM, which of the following is true?

- [ ] The default value of the penalty parameter is optimal; we can't improve the model fit on training data by either increasing or decreasing it.
- [x] Training accuracy can be improved by increasing the value of the penalty parameter.
- [ ] The penalty parameter cannot be varied using sklearn.
- [ ] Training accuracy can be improved by decreasing the value of the penalty

parameter.

- [ ] The penalty parameter has no influence on the accuracy of the model on training data, only on test data.

Question 31/53

The next 4 questions (Questions 31 - 34) are based on the IPL match data (matches.xlsx).

The indicator dl_applied refers to weather-shortened matches in which the Duckworth-Lewis method was applied to determine the winner.

In what proportion of matches did this happen?

- [ ] 50%
- [ ] 0.25%
- [x] 2.5%
- [ ] 25%
- [ ] 5%

Question 32/53

What proportion of matches was won by the team who batted first?

- [x] 45%
- [ ] 52%
- [ ] Impossible to determine from the data

- [ ] 55%
- [ ] 48%

Question 33/53

We define a close match as one which was won by 20 runs or less, or by 4 wickets or less.

We want to build a model to predict whether or not a game will be close based on the following three features:

- whether the match was played in the month of April, or not
- whether the toss winners chose to bat or field
- whether or not the Duckworth-Lewis method was applied

Create these features. Which of the following tuples correctly enumerates respectively the number of April games and choices to field first across the data set?

- [ ] (339, 363)
- [x] (297, 273)
- [ ] (339, 273)
- [ ] (297, 363)

Question 34/53

Build a decision tree classifier on these features, using a train-test split with a 75:25 weight and a random seed of 999.

Which of the following is the most accurate reflection of the confusion matrix on the test data?

- [ ] The model predicts that the vast majority of games will not be close wins. Hence we have few false negatives but many false positives.
- [ ] The model predicts that the vast majority of games will not be close wins. Hence we

have few false positives but many false negatives.

- [x] The model does a reasonably good job of predicting close wins, with relatively few false positives and false negatives.
- [ ] The model predicts that the vast majority of games will be close wins. Hence we have few false positives but many false negatives.
- [ ] The model predicts that the vast majority of games will be close wins. Hence we have few false negatives but many false positives.

Question 35/53

The next 4 questions (Questions 35 - 38) are based on the FIFA players dataset (football_players.csv).

Please note that you may have some difficulties importing this file into pandas, and you may need to do some research to figure out how to do so successfully.

What is the most common Overall score for players in the database?

- [ ] 93
- [ ] 67
- [x] 66
- [ ] 64
- [ ] 68

Question 36/53

Construct a dataset that is a subset of players who can play in central defence (i.e. who have 'CB' somewhere in their Preferred Positions field).

Split this group into three:

- World Class: overall score of 80 or more
- Good: overall score of 70-79
- Mediocre: overall score below 70

Now build a random forest classifier with default parameters apart from setting to 500 trees and setting the random seed to 1971, on ALL of the data for these central defenders, where the target variable is the classification into one of the three classes defined above, and the candidate features are all other numerical variables.

In descending order, which are the five most important features that emerge from this model?

Hint: Search sklearn random forest documentation for feature importance if you don't have any idea how to establish this.

- [ ] Composure, Sliding Tackle, Aggression, Short passing, Marking
- [x] Standing tackle, Marking, Interceptions, Sliding Tackle, Reactions
- [ ] Standing tackle, Sliding Tackle, Composure, Marking, Heading accuracy
- [ ] Composure, Standing Tackle, Marking, Heading accuracy, Sliding Tackle
- [ ] Marking, Reactions, Composure, Standing Tackle, Sliding Tackle

Question 37/53

Why do we generally not use all the data to fit models as we did in the previous question, but rather perform a train-test split or cross-validation?

- [ ] Although it's considered best practice to split the data into test and training sets, this is not actually required in many circumstances.

- [ ] To avoid underfitting to the data, and hence improve our chances of generalising to unseen data.
- [ ] Because Dewald gets mad if we don't.
- [ ] To simultaneously minimise bias and variance.
- [x] To avoid overfitting to the data, and hence improve our chances of generalising to unseen data.

Question 38/53

Split the data into test and training sets, with 33% of the data reserved for the test set and a random seed of 911.

Compare k nearest neighbours (KNN) models with k varying from 1 to 5. Which k gives rise to the best F1 score for the world class good groups respectively?

- [ ] World class: k=5, Good: k=5
- [x] World class: k=5, Good: k=4
- [ ] World class: k=1, Good: k=2
- [ ] World class: k=1, Good: k=4
- [x] World class: k=4, Good: k=5

Question 39/53

Which of the following is an accurate description of logistic regression?

- [ ] Logistic regression fits a linear model to

- [ ] the odds ratio, which is the probability of being in a class as a proportion of the probability of not being in that class.

- [ ] A coefficient of below 1 in a logistic regression implies a negative contribution to the probability of being in the target class

- [ ] Logistic regression fits a linear model to the log odds ratio, which is the probability of being in a class as a proportion of the probability of not being in that class.

- [ ] Logistic regression is not a linear model, unlike linear regression.

- [x] Logistic regression fits a linear model to the log odds ratio, which is the log of the probability of being in a class as a proportion of the probability of not being in that class.

Question 40/53

Which of the following are true of the k-nearest neighbours (KNN) algorithm applied to an n-dimensional feature space?

i. For a new test observation, the algorithm looks at the k training observations closest to it in n-dimensional space and assigns it to the majority class among those k observations.

ii. For a new test observation, the algorithm looks at the k training observations closest to it in n-dimensional space and assigns it proportionally to each class represented in those k observations.

iii. KNN models tend to perform poorly in very high dimensions.

iv. KNN models are well-suited to very high-dimensional data.

v. The K in KNN stands for Kepler, the scientist who first proposed the algorithm.

- [ ] ii and iv

- [ ] i, iv and v

- [ ] i, iii and v
- [x] i and iii
- [ ] i only

Question 41/53

What is a hyperparameter?

- [ ] Machine learning terminology for a model parameter.
- [ ] A model parameter which has more than one dimension.
- [x] A parameter whose value is set before the model-fitting process begins.
- [ ] A parameter which can only be set by grid search.

Question 42/53

NLP (12 Questions, 22 Marks)
Questions 42 - 53

The practical questions of the next 12 questions should be answered using the Essay_data.csv file.

This CSV file contains a personality profile, together with an essay written by an individual with that specific personality type.

Once you have imported the data frame, use the dropna() function to remove rows containing missing values and reset the index.

How many bi-grams can be created from the following sentence after performing the following steps in the correct order:

- Remove stopwords
- Remove punctuation and replace by a single white space
- Convert all text to lower case

I’m a part-time student @explore-software.

- [ ] 6
- [ ] 8
- [x] 4
- [x] 5

Question 43/53

Which ratio best relates the number of intuitive (N) students to the number of sensing (S) students?

- [ ] 6:4
- [x] 7:3
- [ ] 3:7
- [ ] 4:6

Question 44/53

Remove all punctuation from the essays and convert it to lower case.

What is the 10th character in the first essay?

- [ ] ‘t’
- [x] ‘4’

- [ ] ‘ ’
- [x] ‘filled’

Question 45/53

Tokenise the essays.

How many tokens are in the 17th essay?

- [x] 440
- [ ] 94
- [ ] 337
- [ ] 128

Question 46/53

How does lemmatisation differ from stemming?

- [ ] Lemmatisation cuts off part of the word while stemming considers the morphological analysis of the words.
- [x] Lemmatisation always results in actual words, while stemming can result in non-existent words.
- [ ] Stemming always results in actual words, while lemmatisation can result in non-existent words.
- [ ] They operate the same way – they just

- [ ] come from different libraries (scikit-learn and NLTK).

Question 47/53

Use the SnowballStemmer to stem the word experiences.

What is the output?

- [x] ‘experi’
- [ ] ‘experiences’
- [ ] ‘experience’
- [ ] ‘experienc’

Question 48/53

Remove all the stop words.

What is the 24th token in the 81st essay?

- [x] ‘selfconfidence’
- [ ] ‘working’
- [ ] ‘times’
- [ ] ‘evidently’

Question 49/53

How many unique words are in these essays (after we have removed the stopwords)?

- [ ] 4087
- [ ] 3494
- [ ] 3812
- [x] 3406

Question 50/53

Create a bag of words and use it to determine how many times ‘time’ was mentioned in the 56th essay?

- [x] 2
- [ ] 1
- [x] 3
- [ ] 0

Question 51/53

Words that appear at least twice account for what percentage of the total number of words in the essays?

- [ ] 81%
- [ ] 62%
- [x] 90%
- [x] 48%

Question 52/53

What is the most commonly mentioned word by ENFJ personalities?

- [ ] ‘data’
- [x] ‘team’
- [ ] ‘group’
- [ ] ‘academy’

Question 53/53

Create a new column in the data frame containing the bi-grams from each essay.

What is the 109th bi-gram in the 70th essay?

- [ ] (‘may’, ‘better’)
- [ ] (‘work’, ‘quite’)
- [x] (‘quite’, ‘well’)
- [ ] (‘better’, ‘certain’)