MoF Data Analytics and Machine Learning in Finance (Final)

Fällig am 18. Mai um 13:20 **Punkte** 70 **Fragen** 19
Verfügbar am 18. Mai um 12:00 – am 18. Mai um 13:20 etwa 1 Stunde
Zeitlimit Keine

Anweisungen

Examination in the Master of Science

Course title: Data Analytics and Machine Learning in Finance

Semester: SS 2021

Lecturer: Prof. Dr. Kornelia Fabisik

Examination date: 18.05.2021

Aids: Non-programmable calculator

The exam consists of **18** graded questions. You have **70** minutes to complete the examination plus **10** minutes reading time. The maximum of points to be reached is **70**.

You are allowed to use a non-programmable calculator. Programmable calculators or calculators connectible to the internet or to other calculators (e.g., via bluetooth) are NOT allowed. Usage of such calculators will result in failing the exam.

If you encounter any problems with tables, graphs, etc., please use the ***PDF version of the examination*** ***(https://frankfurtschool.instructure.com/courses/4186/files/236769/download?download_frd=1)*** *.*

Dieses Quiz wurde gesperrt um am 18. Mai um 13:20.

Versuchsverlauf

	Versuch	Uhrzeit	Punktzahl
NEUESTE	Versuch 1	80 Minuten	47 von 70

Punktzahl für dieses Quiz: **47** von 70
Abgegeben am 18. Mai um 13:20

Dieser Versuch dauerte 80 Minuten.

Frage 1

0 / 5 Pkte.

Consider the following line of Python code:

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.8, shuffle=True)

Hint: Do not look for errors in the syntax. The syntax is correct. Assume that we are preparing the data split for a supervised machine-learning algorithm. As usual, X is a matrix with features and y is a vector with labels.

Select all or only that apply:

□ Every time the code is run, the data is split in the same way, i.e., the split is reproducible across multiple function calls.

□ The code produces two new data frames.

□ None of the statements.

e antworteten

☑ If we use this code on a data set with 50,000 observations, 40,000 observations would be in the training set and 10,000 in the test set.

Richtig!

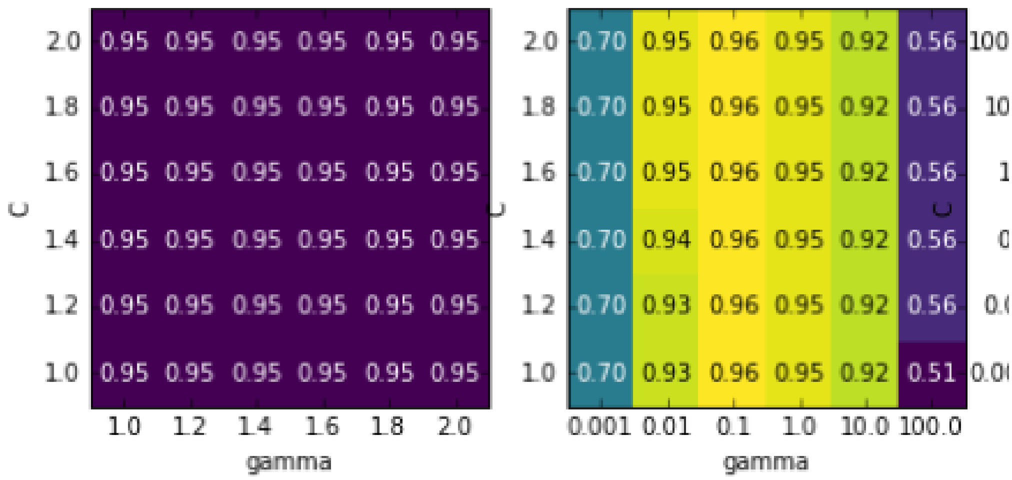
☑ The data is shuffled before splitting.

Frage 2

0 / 5 Pkte.

Assume that we have a binary classification problem and we are tuning two hyperparameters of the support vector classifier (SVC) model that employs a radial basis function (RBF) kernel. It has the C hyperparameter and the gamma hyperparameter. We are looking at a

two-dimensional grid of hyperparameters, which shows the accuracy scores as a heat map.



C \ gamma	1.0	1.2	1.4	1.6	1.8	2.0
2.0	0.95	0.95	0.95	0.95	0.95	0.95
1.8	0.95	0.95	0.95	0.95	0.95	0.95
1.6	0.95	0.95	0.95	0.95	0.95	0.95
1.4	0.95	0.95	0.95	0.95	0.95	0.95
1.2	0.95	0.95	0.95	0.95	0.95	0.95
1.0	0.95	0.95	0.95	0.95	0.95	0.95

C \ gamma	0.001	0.01	0.1	1.0	10.0	100.0
2.0	0.70	0.95	0.96	0.95	0.92	0.56
1.8	0.70	0.95	0.96	0.95	0.92	0.56
1.6	0.70	0.95	0.96	0.95	0.92	0.56
1.4	0.70	0.94	0.96	0.95	0.92	0.56
1.2	0.70	0.93	0.96	0.95	0.92	0.56
1.0	0.70	0.93	0.96	0.95	0.92	0.51

Select all or only that apply:

e antworteten

☑ Heat maps are the outcome of cross-validation, which is a method that can help us pick the best model hyperparameters.

Richtig!

☑ The heat map on the right suggests that we might be able to achieve better accuracy by e.g., increasing gamma and re-running the grid search again.

☐ None of the statements.

☐ The heat map on the left is ideal. We learn from it the direction for how to improve the model further.

e antworteten

☑ The heat map in the middle tells us that the range for the C parameter was selected well. It is the gamma hyperparameter range that we should adjust.

Frage 3

5 / 5 Pkte.

Consider the following statements about **accuracy**:

Select all or only that apply:

Richtig!

☑ It is not suitable for imbalanced data sets.

□ Together with R-squared, they are the most commonly-used evaluation metrics for classification problems.

□ It is the best measure to use when the cost of false positives (also known as Type I error) is high.

□ It is computed as (TP + FP)/(TP + FP + TN + FN) where TP stands for true positives, FP stands for false positives, TN stands for true negatives, and FN stands for false negatives.

□ None of the statements.

Frage 4

3 / 3 Pkte.

Assume that you want to use Python to download data from Yahoo Finance. Assume that you have a new computer (just taken out of the box) and have freshly installed Python. If you run this line of code: **df_yahoo = yf.download('AAPL', start='2020-01-01', end='2020-12-31')** do you expect the code to run without any troubles (i.e., error messages)? Hint: Do not look for errors in the syntax. The syntax is correct.

```
df_yahoo = yf.download('AAPL', start='2020-01-
```

○ Yes.

○ No. But once I install yfinance, the code would surely download the data I need.

Richtig!

● No, I would first need to install yfinance. Afterwards, I would have to add a line in the Python script that says: import yfinance as yf. I would add this line above the line that says: df_yahoo = yf.download('AAPL', start='2020-01-01', end='2020-12-31').

Frage 5

5 / 5 Pkte.

Consider the following statements about **support vector machines (SVMs)**:

Select all or only that apply:

Richtig!

☑ None of the statements.

□ Always when we apply kernelized support vector machines, we use kernels to project the original data into a lower dimension. This transformation can help us classify the data better.

□ If the number of input features is 3 (i.e., in a 3-D space), then the separating hyperplane is just a line.

□ SVMs can only be used for classification problems. In finance, such problems involve e.g., the decision to approve credit application, the decision to buy/sell/hold a stock, and many others.

□ All of the training points always lie on the decision boundary. To make a prediction for a new point, the distance to each of the training points has to be measured.

Frage 6

0 / 5 Pkte.

Consider the following statements about **naive Bayes classifiers:**

Select all or only that apply:

□ In probability theory, the binomial distribution is a generalization of the multinomial distribution.

chtige Antwort

□ None of the statements.

e antworteten

☑ Despite the name, these models can also be used for regression problems.

e antworteten

☑ The three naive Bayes classifiers covered in class differ mainly by the assumptions they make regarding the “prior” in the Bayes’ theorem.

□ Of the three naive Bayes models covered in class, Gaussian naive Bayes models are the most preferred for text data classification.

Frage 7

0 / 5 Pkte.

Consider the following statements about **tree-based models**:

Select all or only that apply:

□ Gradient boosting is considered a gradient descent algorithm. The key ingredient in gradient boosted trees is randomization.

□ Some of the possible pre-pruning criteria involve specifying the minimum depth of the tree, limiting the maximum number of leaves, or requiring a minimum number of points in a node to continue splitting it further.

e antworteten

☑ The process of adjusting the learning rate (the rate at which the next tree corrects the mistake of the previous one) in random forest models is known as hyperparameter tuning.

□ CART stands for Clustering And Regression Trees.

chtige Antwort

□ None of the statements.

Frage 8

2 / 2 Pkte.

What is bootstrapping? Describe the method in 2 sentences.

Ihre Antwort:

Bootstrapping is a test that uses randomized sampling with replacement.

So for example from our n samples data points we draw examples randomly with replacement , creating a dataset as big as the original one but some points will be missing and replaced.

Frage 9

5 / 5 Pkte.

Consider the following statements about **cross-validation**:

Select all or only that apply:

□ When using time-series data, the order of data matters. There does not exist any cross-validation method that we could apply. We therefore never perform cross-validation on time-series data.

□ The following Python code specification means that 360 models would have to be trained: param_grid = {'C': [0.001, 0.01, 0.1, 1, 10], 'gamma': [0.001, 0.01, 0.1, 1, 10]} grid_search = GridSearchCV(SVC(), param_grid, cv=10). Hint: Do not look for errors in the syntax. The syntax is correct.

□ None of the statements.

Richtig!

☑ It is most commonly-used for assessing the generalization performance of supervised machine-learning models.

□ Leave-one-out cross-validation is like a k-fold cross-validation where each fold is a single sample. Each split, you pick a single data point to be the training set and the remaining data points constitute the test set.

Frage 10

4 / 4 Pkte.

How many weights would the algorithm have to compute in a feed-forward neural network (FFNN) with 570 features and 3 hidden layers with 203 hidden units each? Assume one output node.

Richtig!

198.331

chtige Antwort

198.331

Frage 11

0 / 0 Pkte.

To allow for some partial credit, provide the number of weights that would need to be computed in the **first step** of determining the network complexity. Write your answer here: 115710

To allow for some partial credit, provide the number of weights that would need to be computed in the **last step** of determining the network complexity. Write your answer here: 203

The direction is from the input to the output.

You can skip this question if you are confident in your answer to the question about the complexity of a feed-forward neural network (FFNN).

Antwort 1:

e antworteten

115710

chtige Antwort

variables * units

Antwort 2:

e antworteten

203

chtige Antwort

units

Frage 12

2 / 2 Pkte.

We speak of a deep neural network when it has ___ or more layers. Your answer should be an integer. The answer has to be based on the rule provided in class.

Richtig!

2

chtige Antworten 2

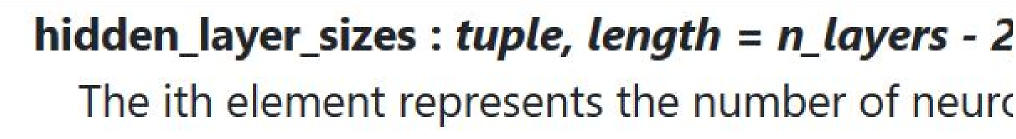
Frage 13

3 / 5 Pkte.

Consider the following Python code and answer the following:

MyMLModel = MLPClassifier(max_iter=1000, hidden_layer_sizes=(10,), random_state = 37)

- Which supervised machine-learning model are we specifying? (1P)
- What does “max_iter=1000” do? Why do we use it? (1P)
- What does “hidden_layer_sizes=(10,)” do? What will the model look like when we use this parameter specification? You can use the below figure as a hint. (2P)



hidden_layer_sizes : *tuple, length = n_layers - 2*

The ith element represents the number of neurc

- Why do we use the parameter “random_state”? (1P)

Ihre Antwort:

1. MLPClassifier is a Neural Network

2. max_iter does give a parameter how many iterations it takes for the solver to converge. In the case of the MLPClassifier it is basically the number of epochs run.

3. It is a parameter that every perceptron or layer has the size of 10. So basically 10 units per layer to feed information forward to next layer.

4. A random state of 0 makes sure we get the same output if we run the same function several times, making the outcome deterministic. So here basically it means we don't want the output to be determinstic.

Frage 14

2 / 2 Pkte.

Imagine that you are asked to predict sentiment based on a corpus that contains finance news (positive or negative). Assume you have a labelled data set on which you can train an algorithm and you have reached the **inference** step.

- Could you use *naive Bayes classifier?*
- Could you use *Latent Dirichlet Allocation (LDA)?*

Comment on the viability of each method. Start your answer with the word "**YES**" or "**NO**", followed by a brief **explanation**. Write maximum 2 sentences.

Ihre Antwort:

1. YES. Since the data is labelled and I can train the algorithm the MultinomialNB is a very good fit for this supervised ML task as it can produce reasonable accuracy using simple assumptions.

2. NO. LDA is predominantly used for topic modeling which is an unsupervised NLP method to reveal atttribution for words.

Frage 15

2 / 2 Pkte.

Explain the difference between supervised and unsupervised machine learning. Write maximum 4 sentences.

Ihre Antwort:

The difference is that supervised ML is used for labeled data and the inputs and outputs are given, while this is not the case for unsupervised ML since there the outputs are missing and the data is not labeled. In the latter we predominantly aim for dimensionality reduction of the given data which means we try to reduce the numbers or features in the dataset without reducing model performance or for clustering which tries detect hidden patterns in the data.

Frage 16

6 / 6 Pkte.

Name 6 uses of supervised machine learning in finance. Your answers should be sufficiently distinct from each other rather than mere nuances of the same application. Describe each use in 1 sentence.

Ihre Antwort:

Stock price prediction: Trying to predict stock prices by using ML to get some regressions going with the help of historic data.

Credit default protection: risk measurement in order to try to predict for example which credit card owners are likely to default.

Robo advisory: Having an AI as the advisor using NLP and tending to customers needs.

Fraud prevention: Use of captchas in order to prevent hacks or fraudulent misuse with bots.

High frequency trading: trading too fast to properly do for humans and better done by machines

Money laundering prevention: Automation of compliance based checks and detection of signs of money landering.

Frage 17

3 / 3 Pkte.

Name and briefly explain 3 clustering techniques. Use maximum 2 sentences for your description of each technique.

Ihre Antwort:

k-means clustering: Most well known clustering technique, centroid or distance based algorithm. Tries to find cluster centers representative od certain regions in the data.

Hierarchical Clustering: involves creating clusters with predominant ordering. No need to specify number of clusters, does it by itself. 2 Approaches, Agglomerative, Divisive. First bottom up approach, second top down, decides direction of clustering process.

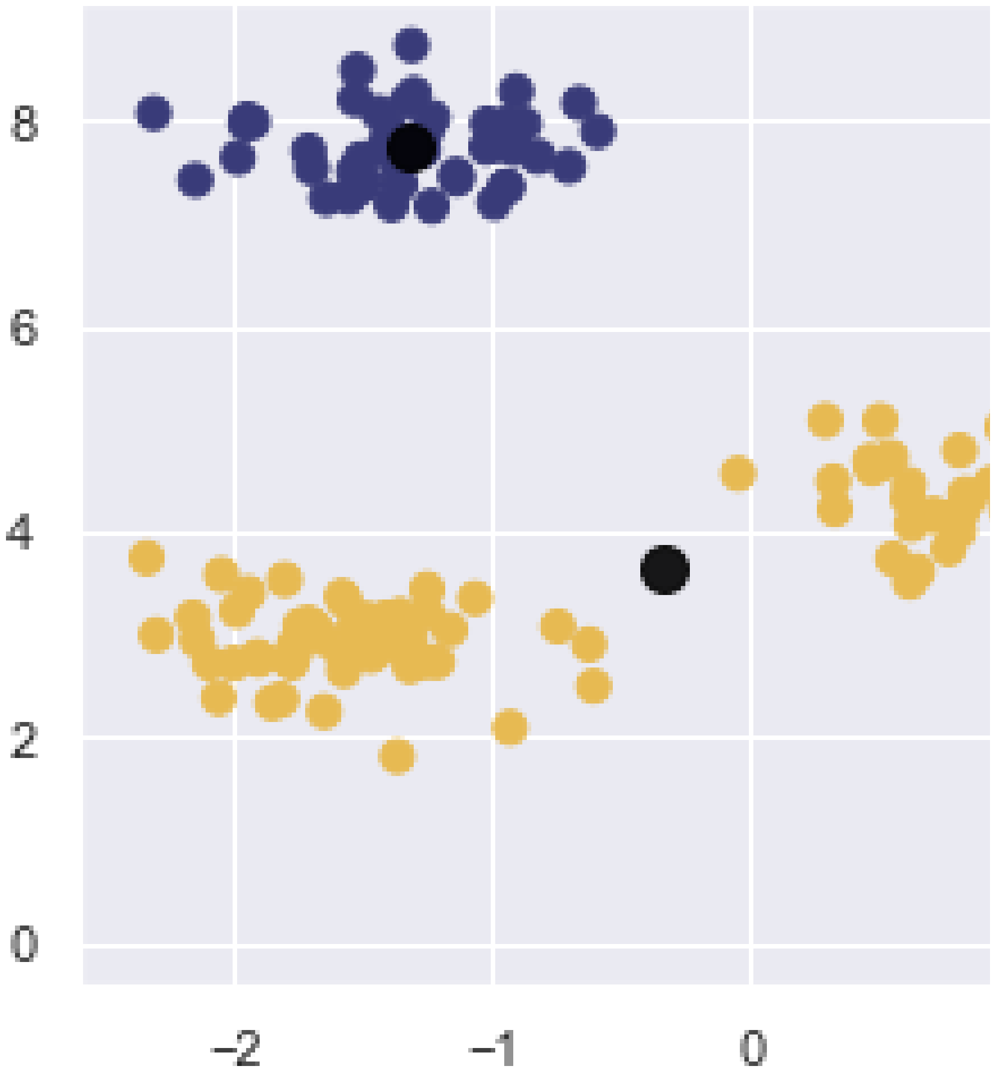
Affinity propagation clustering: creates clusters by sending messages between sample pairs until convergence. Inference done from small number of exemplars, which are most representative of other sampels.

Frage 18

5 / 5 Pkte.

Consider the following statements about **k-means clustering**:

Select all or only that apply:



□ In k-means clustering, the centroid has to be one of the data points.

□ The shown figure suggests that we have the optimal number of clusters.

Richtig!

☑ Because extreme values in a data set can disrupt a clustering solution significantly, we could instead use k-medoids as way for overcoming this problem.

□ None of the statements.

□ The number of clusters is a hyperparameter that is internal to the model, i.e., one does not have to determine it prior to running the algorithm.

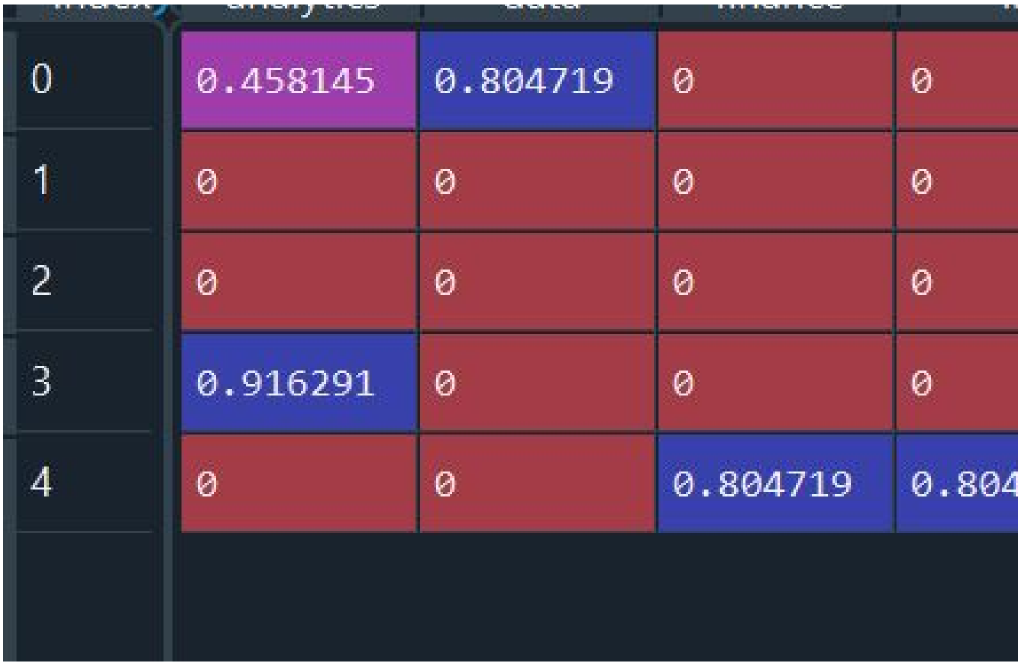
Frage 19

0 / 1 Pkte.

Imagine that you are building a TF-IDF (Term Frequency - Inverse Document Frequency) model from scratch and plan to use it as a document term matrix for training your machine-learning model. Recall Lab 3 during which we spent 40 minutes on the code.

Assume that your TF-IDF looks as follows:

Index	analytics	data	finance	i



0	0.458145	0.804719	0	0
1	0	0	0	0
2	0	0	0	0
3	0.916291	0	0	0
4	0	0	0.804719	0.804

How many documents is this corpus composed of? Your answer should be an integer. Hint: There are no more rows or columns in the matrix beyond those that you see displayed.

e antworteten 4

chtige Antworten 5 (mit Marge: 0)

Quizpunktzahl: **47** von 70