Machine learning exam

This exam is timed and the duration of each question is displayed at the end of each question. A progress bar below each question also indicates the time remaining. **Questions are automatically skipped** as soon as the time runs out, and **there is no turning back.**

It is divided into 2 parts

1- **Oral exam:** you will read the questions on the screen and answer them orally to your examiner.

You will have a two-minute break in between to prepare.

2**- The written exam:** you will read the questions and answers on the screen, and you will have to answer via a google forms form by ticking the right answer(s).

ORAL EXAM 15 min

(2 min)

[redacted]

QUESTION 1

Name three types of supervised machine learning algorithms. (1 min)

QUESTION 2

How can overfitting be avoided? (1 min)

QUESTION 3

Provide an example of an application where an unsupervised algorithm can be used to improve a supervised model. (1min30)

QUESTION 4

Explain the trade-off between bias and variance in a machine learning model (1 min 30)

QUESTION 5

Linear regression algorithms are used to establish a relationship between input variables and an output variable, which can be either discrete or continuous.

True or False? Explain why if false. (1 min)

QUESTION 6

How can outliers in a dataset be detected and handled? (1 min 30)

QUESTION 7

Explain the difference between supervised and unsupervised machine learning. (1 min)

QUESTION 8

Explain the difference between linear regression and logistic regression. (1 min)

QUESTION 9

How does k-means clustering decide the number of clusters to form? *(1 min)*

QUESTION 10

Decision trees can handle both categorical and continuous features.

True or False? Explain why if false. *(30 s)*

QUESTION 11

Why and how would you use dimensionality reduction (e.g., PCA) in a machine learning project?
(1min)

QUESTION 12

Why is it important to split a dataset into training and test sets in a supervised ML task? (1min)

QUESTION 13

How do you choose the right number of neighbors in the KNN algorithm? *(1 min)*

QUESTION 14

How does cross-validation improve a model's performance? *(1 min)*

WRITTEN EXAM 15 min

(2 min)

[redacted]

QUESTION 1

Name 3 metrics that can be used to evaluate the output of a classification model? *(1 min)*

QUESTION 2

What is the formula for accuracy? *(1 min)*

QUESTION 3

What are the main steps of a machine learning project? (Provide only the names of the steps, without explanation.) *(2 min)*

QUESTION 4

Among the following scenarios, choose the one(s) that might be considered as a regression problem: (3min)

a) We collect a set of data on the top 500 firms in France. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary

b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charger for the product, marketing budget, competition price, and ten other variables.

c) We are interested in predicting the % change in the US dollar in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week record the % change in the dollar, the % change in the US market, the % change in the British market, and the % change in the German market

d) An emergency room in a hospital measures 10 variables (e.g. blood pressure, age, etc.) of newly admitted patients. Based on these variables, the hospital can determine the life-risk (high/low) of each patient. A decision has to be taken whether to put the patient in an intensive-care unit based on his life-risk level.

QUESTION 5

1. **To measure the effectiveness of a split while constructing a Decision Tree, we can use:** (1 min)

   a. The Classification error metric

   b. The Gini index metric

   c. The entropy metric

   d. None of the above

QUESTION 6

What does the term "cross-validation" mean? (1 min)

a) A technique to normalize data before training.

b) A method to evaluate the performance of a model by dividing the data into multiple subsets.

c) A technique to detect outliers in a dataset.

d) A method to balance classes in an imbalanced dataset.

QUESTION 7

Which technique can be used to handle missing data in a dataset? (1 min)

a) Removing columns with missing values.
b) Imputation using the mean or median.
c) Using predictive models to estimate the missing values.
d) All of the above.

QUESTION 8

What does bagging stand for? (1 min)

a) Boosted Aggregation
b) Bootstrap Aggregation
c) Bayesian Aggregation
d) Balanced Aggregation

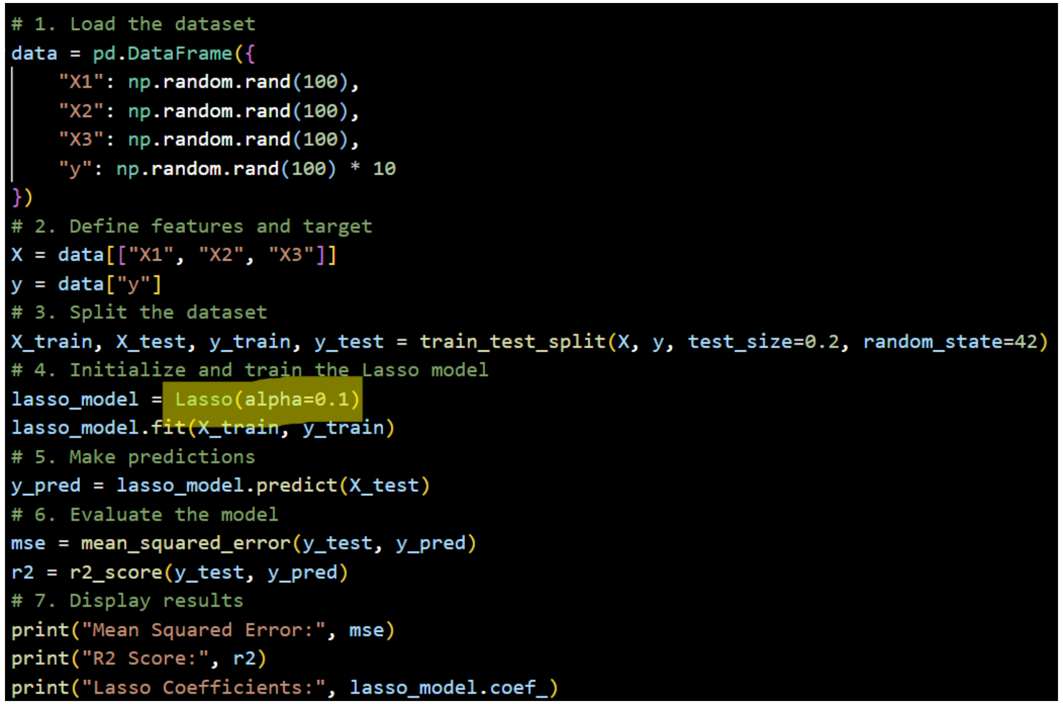
QUESTION 9

What is the purpose of Grid Search in machine learning? (1 min)

a) To visualize the decision boundary of a model
b) To perform hyperparameter tuning
c) To optimize the weights of a neural network
d) To identify outliers in the data

QUESTION 10

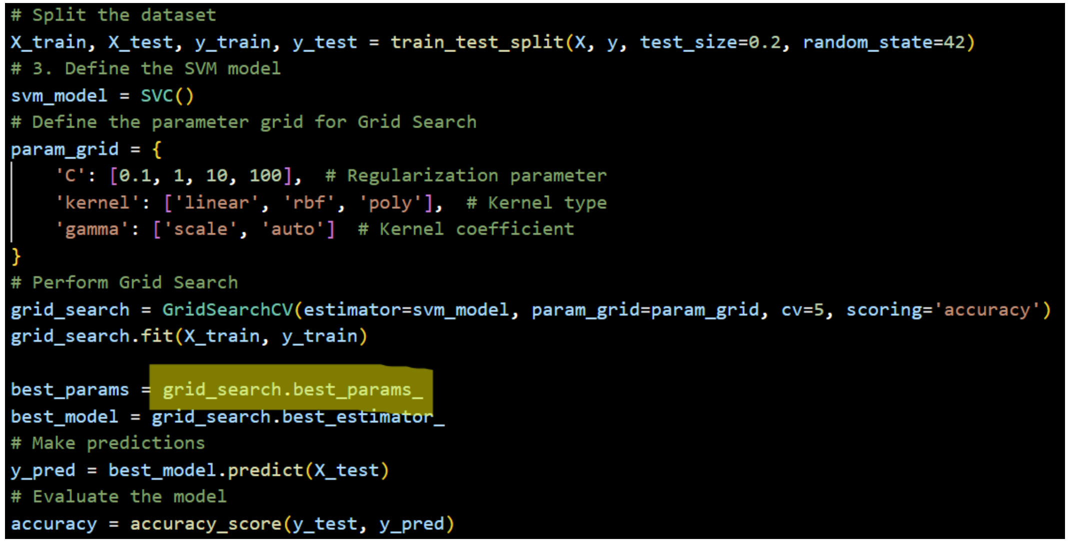
What does the alpha parameter in the Lasso model do? (1 min)



```
# 1. Load the dataset
data = pd.DataFrame({
    "X1": np.random.rand(100),
    "X2": np.random.rand(100),
    "X3": np.random.rand(100),
    "y": np.random.rand(100) * 10
})
# 2. Define features and target
X = data[["X1", "X2", "X3"]]
y = data["y"]
# 3. Split the dataset
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# 4. Initialize and train the Lasso model
lasso_model = Lasso(alpha=0.1)
lasso_model.fit(X_train, y_train)
# 5. Make predictions
y_pred = lasso_model.predict(X_test)
# 6. Evaluate the model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
# 7. Display results
print("Mean Squared Error:", mse)
print("R2 Score:", r2)
print("Lasso Coefficients:", lasso_model.coef_)
```

QUESTION 11

What does grid_search.best_params_ return? (2 min)



```
# Split the dataset
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# 3. Define the SVM model
svm_model = SVC()
# Define the parameter grid for Grid Search
param_grid = {
    'C': [0.1, 1, 10, 100],  # Regularization parameter
    'kernel': ['linear', 'rbf', 'poly'],  # Kernel type
    'gamma': ['scale', 'auto']  # Kernel coefficient
}
# Perform Grid Search
grid_search = GridSearchCV(estimator=svm_model, param_grid=param_grid, cv=5, scoring='accuracy')
grid_search.fit(X_train, y_train)

best_params = grid_search.best_params_
best_model = grid_search.best_estimator_
# Make predictions
y_pred = best_model.predict(X_test)
# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
```

END OF EXAM

(2 min)

[redacted]