

Given the following code that trains a linear regressor, complete the code to evaluate the trained model on the test set. You need to prepare the test set, evaluate, and print the RMSE.

```
# import and initialization code is not shown

from sklearn.model_selection import train_test_split
housing_train, housing_test = train_test_split(housing, test_size=0.2, random_state=42)

X_train = housing_train.drop("median_house_value", axis=1)
y_train = housing_train["median_house_value"].copy()
X_train_prepared = full_pipeline.fit_transform(X_train)

reg = LinearRegression()
reg.fit(X_train_prepared, y_train)
```



Given the output of the following confusion matrix, what is the precision?

```
from sklearn.metrics import confusion_matrix
```

```
confusion_matrix(y_train_5, y_train_pred)
```

```
array([[4, 2],  
       [1, 3]])
```



Answer:

The following are main reasons why AI is succeeding now:

Select one or more:

- More users are becoming connected through the internet.
- Internet communication is becoming faster.
- Developers are developing better algorithms.
- Problems are becoming easier to solve.
- Computers are becoming faster.
- Training datasets are becoming more available.

Assume that you want to solve a linear regression problem using the batch gradient descent algorithm. How many multiplication operations are required in each training step to calculate the gradient vector of the MSE cost function? Given that the dataset has 1,000 training instances and each instance has 9 features.


Answer:

POWERUNIT

Which of the following are unsupervised learning tasks?

Select one or more:

- Sorting
- Linear Algebra
- Regression
- Classification
- Anomaly detection
- Clustering

POWERUNIT 

Assume that you have a very small dataset. What is the best validation scheme among the following?

Select one:

- Cross validation (with 70% for training and 30% for testing)
- Cross validation (with 95% for training and 5% for testing)
- Stratified split (85% for training and 15% for testing)
- 10-fold cross validation
- Cross validation (with 80% for training and 20% for testing)

Given the following code that loads the Diabetes dataset, complete the code to evaluate the 5-nearest neighbors regressor using the 3 fold cross validation method and print the average RMSE.

```
import numpy as np
from sklearn.neighbors import KNeighborsRegressor
from sklearn.model_selection import cross_val_score
from sklearn import datasets
X, y = datasets.load_diabetes(return_X_y=True)
reg = KNeighborsRegressor(n_neighbors=5)
```

Assume that you want to solve a linear regression problem using the mini-batch gradient descent algorithm. How many multiplication operations are required in each training step to calculate the gradient vector of the MSE cost function? Given that the dataset has 1,000 training instances, each instance has 9 features, and the mini batch size is 100.

Answer:



Assume that you have a very small dataset. What is the best validation scheme among the following?

Select one:

- Cross validation (with 70% for training and 30% for testing)
- Cross validation (with 95% for training and 5% for testing)
- Stratified split (85% for training and 15% for testing)
- 10-fold cross validation
- Cross validation (with 80% for training and 20% for testing)

POWERUNIT



Given the output of the following confusion matrix, what is the recall?

```
from sklearn.metrics import confusion_matrix
```

```
confusion_matrix(y_train_5, y_train_pred)
```

```
array([[4, 2],  
       [1, 3]])
```

Answer:

.75

Assume that you have a very large dataset. What is the best validation scheme among the following?

Select one:

- 10-fold cross validation
- Cross validation (with 70% for training and 30% for testing)
- Stratified split (85% for training and 15% for testing)
- Cross validation (with 95% for training and 5% for testing)
- Cross validation (with 80% for training and 20% for testing)

Given the following code and output, write your Python code below to handle the data.

Time left 0:11:21

```
import pandas as pd
data = pd.read_csv("filename.csv")
data.info()
<class 'pandas.core.frame.dataframe'="">
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 3 columns):
column_1 100000 non-null float64
column_2 100000 non-null float64
column_3    10 non-null float64
dtypes: float64(3)
memory usage: 2.5 MB
```

Which of the following are supervised learning tasks?

Select one or more:

- Regression
- Sorting
- Clustering
- Linear Algebra
- Classification
- Anomaly detection

POWERUNIT

Assume that you have a very small dataset. What is the best validation scheme among the following?

Select one:

- Cross validation (with 70% for training and 30% for testing)
- Cross validation (with 95% for training and 5% for testing)
- Stratified split (85% for training and 15% for testing)
- 10-fold cross validation
- Cross validation (with 80% for training and 20% for testing)

[Clear my choice](#)