ITESM

MAESTRÍA EN INTELIGENCIA ARTIFICIAL APLICADA

SEGUNDO TRIMESTRE

APUNTES DATA ANALYSIS CON PYTHON

MODULO 5

NOMBRE: VILLALPANDO GUERRERO JIRAM CESAR

MODULO 5

Evaluación del modelo

Hay distintos modelos de evaluación de modelos

Como In Sample: dice qué tan bien el modelo se adecua los datos

Es importante entrenar y setear las evaluaciones

Es decir, estimar el porcentaje para establecer las relaciones de estos y así poder entrenar al modelo adecuadamente.

Se puede usar la función: train:test:Split() desde sklearn

Ejemplo:

```
x_train, x_test, y_train, y_test = train_test_split(x_data, y_data, test_size=0.3, random_state=0)
```

Nos dará un resultado con os distintos scores.

También se debe de ser consiente de los errores de las distribuciones, para ello se deben de determinar al momento de realizar los análisis.

Para la evaluación de métricas se puede realizar una evaluación cruzada. Se puede usar la función: cross_val_score().

Ejemplo:

```
from sklearn.model_selection import cross_val_score


scores= cross_val_score(lr, x_data, y_data, cv=3)
```

Overfitting, underfitting y selección del modelo

Underfitting se refiere al escenario en el que un modelo de aprendizaje automático no puede generalizarse o encajar bien en un conjunto de datos invisible.

Al momento de elegir los modelos, hay que saber qué debemos de elegir para realizar un mejor análisis.

El overfitting o sobreajuste es un error de modelado en estadística que ocurre cuando una función está demasiado alineada con un conjunto limitado de puntos de datos. El resultado es un modelo que es útil solo cuando se refiere a su conjunto de datos inicial y no a cualquier otro conjunto de datos.

Regresión Ridge

La regresión de cresta es un método para estimar los coeficientes de modelos de regresión múltiple en escenarios donde las variables independientes están altamente correlacionadas. Se ha utilizado en muchos campos, incluidos la econometría, la química y la ingeniería.

Ejemplo:

$$\hat{y} = 1 + 2\,x - 3\,x^2 - 2x^3 - 12\,x^4 - 40x^5 + 80x^6 + 71x^7 - 141x^8 - 38x^9 + 75x^{10}$$

Alpha es un parámetro que se elige antes de empezar el modelo

Ejemplo:

Alpha	x	x^2	x^3	x^4	x^5	x^6	x^7	x^8	x^9	x^{10}
0	2	-3	-2	-12	-40	80	71	-141	-38	75
0.001	2	-3	-7	5	4	-6	4	-4	4	6
0.01	1	-2	-5	-0.04	0.15	-1	1	-0.5	0.3	1
1	0.5	-1	-1	-0.614	0.70	-0.38	-0. 56	-0.21	-0.5	-0.1
10	0	-0.5	-0.3	-0.37	-0.30	-0.30	-0.22	-0.22	-0.22	-0.17

Se puede hacer regresión con:

```
from sklearn.linear_model import Ridge
RigeModel=Ridge(alpha=0.1)
RigeModel.fit(X,y)
Yhat=RigeModel.predict(X)
```

Grid Search

En el aprendizaje automático, un hiperparámetro es un parámetro cuyo valor se utiliza para controlar el proceso de aprendizaje. Por el contrario, los valores de otros parámetros se obtienen a través del entrenamiento

Esto ayuda a elegir el mejor modelo que se requiera.

En Python se puede usar lo siguiente:

```
from sklearn.linear_model import Ridge
from sklearn.model_selection import GridSearchCV
```

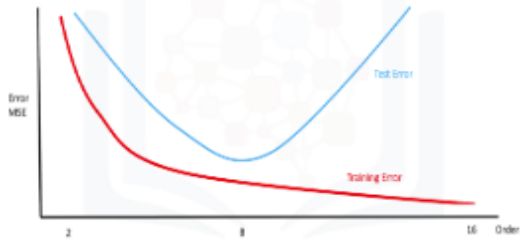
Graded Review Questions

4. Check your grades in the course at any time by clicking on the "Progress" ta

Question 1

1/1 point (graded)

In the following plot, the vertical axis shows the mean square error and the horizontal axis represents the order of the polynomial. The red line represents the training error the blue line is the test error. What is the best order of the polynomial given the possible choices in the horizontal axis?



- 2
- 8
- 16

Save | Show answer

Submit You have used 1 of 2 attempts

Correct (1/1 point)

Question 2

1/1 point (graded)

What is the correct use of the "train_test_split" function such that 40% of the data samples will be utilized for testing the parameter "random_state" is set to zero and the input variables for the features and targets are_data, y_data respectively?

- train_test_split(x_data, y_data, test_size=0, random_state=0.4)
- train_test_split(x_data, y_data, test_size=0.4, random_state=0)
- train_test_split(x_data, y_data)

Save | Show answer

Submit You have used 1 of 2 attempts

Correct (1/1 point)

Questions 3

1/1 point (graded)

What is the output of `cross_val_score(lre, x_data, y_data, cv=2)`?

- The predicted values of the test data using cross-validation.
- The average R^2 on the test data for each of the two folds.
- This function finds the free parameter alpha.

Save | Show answer

Submit You have used 1 of 2 attempts

Correct (1/1 point)

Question 4

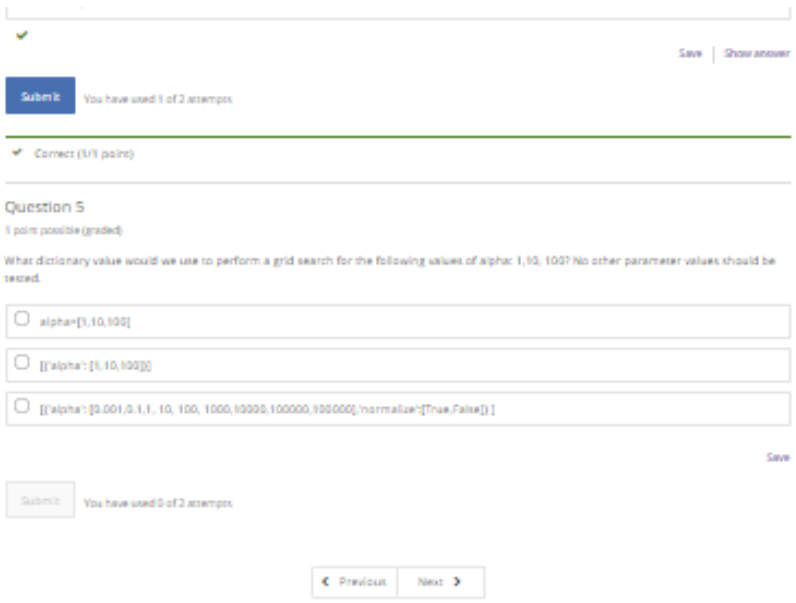
1/1 point (graded)

What is the code to create a ridge regression object "RR" with an alpha term equal 10?

- RR=LinearRegression(alpha=10)
- RR=Ridge(alpha=10)
- RR=Ridge(alpha=1)

Save | Show answer

Submit You have used 1 of 2 attempts


Save | Show answer
Submit
You have used 1 of 2 attempts
Correct (1/1 point)
Question 5
1 point possible (graded)
What dictionary value would we use to perform a grid search for the following values of alpha: 1, 10, 100? No other parameter values should be tested.
alpha=[1,10,100]
[{'alpha': [1,10,100]}]
[{'alpha': [0.001,0.1,1, 10, 100, 1000,10000,100000,1000000],'normalize':[True,False]}]
Save
Submit
You have used 0 of 2 attempts
Previous
Next