Περίληψη

Ο κύριος σκοπός της παρούσας εργασίας είναι η λεπτομερής ανάλυση και αξιολόγηση διαφόρων μεθόδων παλινδρόμησης, όπως η απλή γραμμική μέθοδος των συνήθων ελαχίστων τετραγώνων (Ordinary Least Squares - OLS) , οι μέθοδοι παλινδρόμησης Ridge και LASSO (Least Absolute Shrinkage and Selection Operator). Αυτές οι μέθοδοι συνδυάζονται με τη σειρά τους με τεχνικές επαναδειγματοληψίας, όπως η διασταυρούμενη επικύρωση (cross validation), για καλύτερη εκτίμηση της απόδοσης των διάφορων μεθόδων. Επιπρόσθετος σκοπός είναι η εξοικείωση με τις μετρικές αξιολόγησης, όπως το μέσο τετραγωνικό σφάλμα (Mean Squared Error - MSE), το μέσο απόλυτο σφάλμα (Mean Absolute Error - MAE) και το μέσο απόλυτο ποσοστιαίο σφάλμα (Mean Absolute Percentage Error - MAPE), καθώς και η στρατηγική επιλογής της κατάλληλης μετρικής. Θα ασχοληθούμε επίσης με τις υπερπαραμέτρους που ορίζουν ένα πρόβλημα μηχανικής μάθησης, όπως το βάρος ομαλοποίησης και ο βαθμός της πολυωνυμικής συνάρτησης. Ιδιαίτερη σημασία στην μελέτη και την επιλογή των τιμών των υπερπαραμέτρων έχει η αποφυγή της υπερεκπαίδευσης (overfitting) του μοντέλου.

Συνήθης γραμμική παλινδρόμηση (OLS)

Το γραμμικό μοντέλο αποτελεί εδώ και χρόνια βασικό πυλώνα της στατιστικής και εξακολουθεί να παραμένει ένα από τα σημαντικότερα εργαλεία. Δεδομένου ενός διανύσματος χαρακτηριστικών $\mathbf{X^T=(X_1,X_2,....,Xp)}$, η έξοδος **Y** μέσω της συγκεκριμένης μεθόδου μπορεί να προβλεφθεί από την σχέση:

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j$$

όπου β_j οι συντελεστές ή τα βάρη των χαρακτηριστικών του προβλήματος.
Υπάρχουν πολλές διαφορετικές μέθοδοι για την προσαρμογή του γραμμικού μοντέλου σε ένα σύνολο δεδομένων, αλλά η μέθοδος των ελαχίστων τετραγώνων είναι η πιο δημοφιλής. Σε αυτή την προσέγγιση ελαχιστοποιείται το άθροισμα των τετραγώνων του υπολοίπου.

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - x_i^T \beta)^2.$$

Η παραπάνω εξίσωση μπορεί να γραφεί σε συμβολισμό πινάκων ως εξής:

$$RSS(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) \ .$$

Παλινδρόμηση Ridge (Ομαλοποίηση L2)

Η ομαλοποίηση Ridge είναι άλλη μία γνωστή μέθοδος παλινδρόμησης. Η μέθοδος Ridge μειώνει τους συντελεστές των χαρακτηριστικών του διανύσματος $\mathbf{X^T}$ εφαρμόζοντας την κατάλληλη ποινή(penalty). Οι συντελεστές Ridge δίνονται από την παρακάτω σχέση:

$$\hat{\beta}^{Ridge} = argmin \left(\sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ii} \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right) .$$

Το λ ορίζεται ως το μέγεθος που καθορίζει την ποινή που επιβάλλεται στα αντίστοιχα βάρη του κάθε χαρακτηριστικού. Γενικά, οι συντελεστές(βάρη) συρρικνώνονται προς το μηδέν (και μεταξύ τους). Η παραπάνω εξίσωση σε μορφή πίνακα δίνεται ως εξής:

$$\hat{\beta}^{Ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \ .$$

Παλινδρόμηση LASSO (Ομαλοποίηση L1)

Η τελευταία μέθοδος ομαλοποίησης που θα χρησιμοποιήσουμε είναι η LASSO και έχει αρκετές ομοιότητες με την μέθοδο Ridge. Η μέθοδος αυτή δίνει αραιές λύσεις και συχνά χρησιμοποιείτε ως μέθοδος επιλογής χαρακτηριστικών στην λογική της μείωσης των διαστάσεων ενός προβλήματος. Η παλινδρόμηση Lasso στοχεύει στον μηδενισμό των συντελεστών των χαρακτηριστικών του μοντέλου τα οποία δεν είναι ιδιαίτερα «επιδραστικά» στην τελική πρόβλεψη **Y**.

$$L_{lasso}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})^2 + \lambda\sum_{j=1}^{m}|\hat{\beta}_j| \ .$$

Η παραπάνω εξίσωση σε μορφή πίνακα δίνεται ως εξής:

$$C(\boldsymbol{X}, \boldsymbol{\beta}) = \{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\} + \lambda||\boldsymbol{\beta}||_1 \ .$$

Όπως και στην μέθοδο ομαλοποίησης L2 το λ ορίζει το βάρος ομαλοποίησης, με την διαφορά ότι κάνει διάκριση των χαρακτηριστικών τα οποία θεωρούνται σημαντικά στη τελική πρόβλεψη μηδενίζοντας τους συντελεστές των χαρακτηριστικών που δεν θεωρούνται σημαντικοί. Η αυστηρότητα με την οποία θα επιλεχθούν τα κατάλληλα χαρακτηριστικά καθορίζεται από τον χρήστη. Και στις 2 μεθόδους ομαλοποίησης, ιδιαίτερη σημασία στην τελική επιλογή της ποινής που θα επιβληθεί στα χαρακτηριστικά έχει η διαδικασία της επαναδειγματοληψίας.

Αξιολόγηση Αλγορίθμων

Για την εύρεση του βέλτιστου αλγορίθμου και κατά συνέπεια του βέλτιστου μοντέλου πρέπει να βρεθεί το κατά πόσο ο εκάστοτε αλγόριθμος προσφέρει ικανοποιητικά αποτελέσματα. Για την αξιολόγηση των αλγορίθμων παλινδρόμησης δεν υπάρχει ένα προφανές μέτρο εύρεσης της απόδοσής τους. Αντιθέτως υπάρχουν διάφορες μετρικές για την μέτρηση των σφαλμάτων μεταξύ πραγματικών και προβλεπόμενων τιμών όπως το μέσο απόλυτο σφάλμα, το μέσο τετραγωνικό σφάλμα και το μέσο απόλυτο ποσοστιαίο σφάλμα.

Μέσο απόλυτο σφάλμα (MAE)

Το μέσο απόλυτο σφάλμα (Mean Absolute Error - MAE) εκφράζει ένα μέτρο της ακρίβειας της πρόβλεψης έναντι των πραγματικών τιμών. Συγκεκριμένα πρόκειται για το άθροισμα της απόλυτης τιμής των διαφορών μεταξύ πραγματικής και προβλεπόμενης τιμής διαιρεμένο με το πλήθος των

παρατηρήσεων/προβλέψεων. Όσο μεγαλύτερη είναι η τιμή του δείκτη τόσο μικρότερη προκύπτει η ακρίβεια της μεθόδου που εφαρμόστηκε. Υπολογίζεται όπως παρακάτω:

$$MAE = \frac{1}{n}\sum_{j=1}^{n} |y_j - \widehat{y_j}| \ .$$

Μέσο απόλυτο ποσοστιαίο σφάλμα (MAPE)

Ορισμένες φορές είναι πιο χρήσιμος ο υπολογισμός των σφαλμάτων πρόβλεψης σε καθαρά ποσοστιαία μορφή. Ο συγκεκριμένος στατιστικός δείκτης είναι ιδιαίτερα σημαντικός όταν οι πραγματικές τιμές (Y) είναι ιδιαίτερα υψηλές. Το μέσο απόλυτο ποσοστιαίο σφάλμα είναι εκφρασμένο επί τις εκατό και λαμβάνει τιμές μεγαλύτερες ή ίσες του μηδενός με τις μικρότερες τιμές να υποδηλώνουν και καλύτερη απόδοση του αλγορίθμου. Το μέσο απόλυτο ποσοστιαίο σφάλμα δίνεται ως εξής:

$$MAPE = \frac{100\%}{n}\sum_{j=1}^{n} \frac{|y_j - \widehat{y_j}|}{y_j} \ .$$

Μέσο τετραγωνικό σφάλμα (MSE)

Όπως και το μέσο απόλυτο σφάλμα είναι ένα μέτρο της ακρίβειας της πρόβλεψης το οποίο όμως δίνει πολύ μεγαλύτερο βάρος στα μεγάλα σφάλματα (αν αναλογιστούμε πως τα σφάλματα τετραγωνίζονται) και μικρότερο βάρος στα μικρά σφάλματα. Υπολογίζεται από τον παρακάτω τύπο:

$$MSE = \frac{1}{n}\sum_{j=1}^{n} (y_j - \widehat{y_j})^2 \ .$$

Ερώτημα 1

Αξιολογήστε την επίδοση της γραμμικής παλινδρόμησης ελαχίστων τετραγώνων (Ordinary Least Squares regression), καθώς και της γραμμικής παλινδρόμησης Ridge και LASSO. Πειραματιστείτε με διαφορετικές τιμές του βάρους ομαλοποίησης. Παρουσιάστε συνοπτικά τα αποτελέσματα της αξιολόγησης με βάση το μέσο τετραγωνικό σφάλμα (MSE), το μέσο απόλυτο σφάλμα (MAE) και το μέσο απόλυτο ποσοστιαίο σφάλμα (MAPE).

Μελέτη των Δεδομένων

Αρχικά διαβάσαμε το σύνολο δεδομένων που μας δίνεται, το αρχείο “Concrete_Data.xls”. Το σύνολο δεδομένων αποτελείται από 9 στήλες όπου αντιπροσωπεύουν τα χαρακτηριστικά για κάθε δείγμα σκυροδέματος και 1030 γραμμές, όσες και ο αριθμός των δειγμάτων.

```
Data.shape
```

```
(1030, 9)
```

Οι πρώτες 8 στήλες αντιστοιχούν σε μετρήσεις των ιδιοτήτων για κάθε δείγμα σκυροδέματος και ορίζουν τα διανύσματα χαρακτηριστικών. Η τελευταία στήλη αντιπροσωπεύει την αντοχή του σκυροδέματος (“Concrete compressive strength”) στην πίεση και δίνεται σε MPa. Πρόκειται ουσιαστικά

για την τιμή εκείνη που θέλουμε να προβλέπουμε εκπαιδεύοντας το μοντέλο. Όλες οι τιμές του συνόλου δεδομένων είναι συνεχείς τιμές (μη κατηγορικές). Στους παρακάτω πίνακες δίνονται όλες οι πληροφορίες για τα χαρακτηριστικά του προβλήματος.

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1030 entries, 0 to 1029
Data columns (total 9 columns):
 #   Column                                          Non-Null Count  Dtype
---  ------                                          --------------  -----
 0   Cement (component 1)(kg in a m^3 mixture)       1030 non-null   float64
 1   Blast Furnace Slag (component 2)(kg in a m^3 mixture)  1030 non-null   float64
 2   Fly Ash (component 3)(kg in a m^3 mixture)      1030 non-null   float64
 3   Water  (component 4)(kg in a m^3 mixture)       1030 non-null   float64
 4   Superplasticizer (component 5)(kg in a m^3 mixture)    1030 non-null   float64
 5   Coarse Aggregate  (component 6)(kg in a m^3 mixture)   1030 non-null   float64
 6   Fine Aggregate (component 7)(kg in a m^3 mixture)      1030 non-null   float64
 7   Age (day)                                       1030 non-null   int64
 8   Concrete compressive strength(MPa, megapascals)        1030 non-null   float64
dtypes: float64(8), int64(1)
memory usage: 72.5 KB
```

	count	mean	std	min	25%	50%	75%	max
Cement (component 1)(kg in a m^3 mixture)	1030.0	281.165631	104.507142	102.000000	192.375000	272.900000	350.000000	540.000000
Blast Furnace Slag (component 2)(kg in a m^3 mixture)	1030.0	73.895485	86.279104	0.000000	0.000000	22.000000	142.950000	359.400000
Fly Ash (component 3)(kg in a m^3 mixture)	1030.0	54.187136	63.996469	0.000000	0.000000	0.000000	118.270000	200.100000
Water (component 4)(kg in a m^3 mixture)	1030.0	181.566359	21.355567	121.750000	164.900000	185.000000	192.000000	247.000000
Superplasticizer (component 5)(kg in a m^3 mixture)	1030.0	6.203112	5.973492	0.000000	0.000000	6.350000	10.160000	32.200000
Coarse Aggregate (component 6)(kg in a m^3 mixture)	1030.0	972.918592	77.753818	801.000000	932.000000	968.000000	1029.400000	1145.000000
Fine Aggregate (component 7)(kg in a m^3 mixture)	1030.0	773.578883	80.175427	594.000000	730.950000	779.510000	824.000000	992.600000
Age (day)	1030.0	45.662136	63.169912	1.000000	7.000000	28.000000	56.000000	365.000000
Concrete compressive strength(MPa, megapascals)	1030.0	35.817836	16.705679	2.331808	23.707115	34.442774	46.136287	82.599225

Παρατηρούμε ότι το άνω και κάτω όριο και η μέση τιμή των μετρήσεων των χαρακτηριστικών για τα δείγματα σκυροδέματος διαφέρουν σημαντικά μεταξύ τους, εμφανίζοντας μεγάλη απόκλιση. Δηλαδή κάθε μέτρηση έχει διαφορετική κλίμακα. Επίσης, οι τιμές των χαρακτηριστικών εμφανίζουν μεγάλη κλίση (skewness). Πιο συγκεκριμένα, κανονική κλίση παρουσιάζουν τα χαρακτηριστικά (Cement, Fine Aggerate, Coarse Aggerate) , δεξιά κλίση τα χαρακτηριστικά (Age, Superplasticizer, Blast Furnace Slag) και αριστερή κλίση το χαρακτηριστικό «Water». Τέλος το χαρακτηριστικό «Age» έχει διαφορετική μονάδα μέτρησης από τα υπόλοιπα χαρακτηριστικά και θα μπορούσε να αποτελεί χαρακτηριστική μεταβλητή των υπόλοιπων χαρακτηριστικών σε μία βαθύτερη ανάλυση του προβλήματος.

Προεπεξεργασία των δεδομένων

Διαχωρίζουμε το σύνολο δεδομένων σε σύνολο εκπαίδευσης και σύνολο αξιολόγησης έτσι ώστε το 70% του συνόλου δεδομένων να αντιστοιχεί στο σύνολο εκπαίδευσης (train_set), με τη σειρά που δίνεται, και το υπόλοιπο 30% στο σύνολο αξιολόγησης(test_set).

```
# Διαχωρισμός των δεδομένων σε train και test set
X_train= X[0:721]
y_train=y[0:721]
X_test= X[721:1030]
y_test=y[721:1030]
```

Πριν ξεκινήσουμε την εκπαίδευση του μοντέλου και την αξιολόγηση των διάφορων μεθόδων γραμμικής παλινδρόμησης κάνουμε κλιμακοποίηση των δεδομένων. Η κλιμακοποίηση των δεδομένων είναι απαραίτητη ειδικά στην περίπτωση που χρησιμοποιούμε τους αλγορίθμους Ridge και LASSO. Η μέθοδος κλιμακοποίησης που χρησιμοποιήσαμε είναι η MinMaxScaler και επιλέχθηκε για τους ακόλουθους λόγους:

i. η διακύμανση των τιμών που παίρνει κάθε χαρακτηριστικό (το άνω και κάτω όριο των μετρήσεων) είναι σημαντική.
ii. υπάρχει δυσαναλογία μεταξύ των τιμών των χαρακτηριστικών λόγω της διαφορετικής φύσης του κάθε χαρακτηριστικού.
iii. τα χαρακτηριστικά παρουσιάζουν δεξιά κλίση.

```
# Κλιμακοποίηση των δεδομένων μεσω της MinMaxScaler
scaler = MinMaxScaler()
scaler.fit(X_train)
scaler.fit(X_test)
X_train_scaled = scaler.transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

Εκπαίδευση του μοντέλου και αποτελέσματα

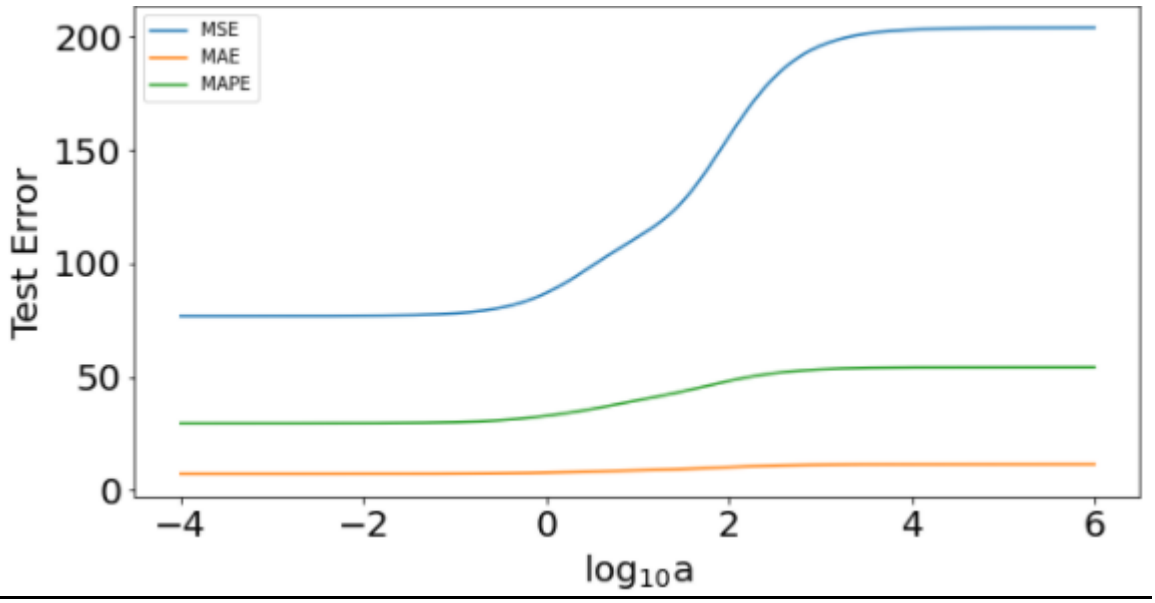
```
# Εκπαίδευση του μοντέλου
clf = skl.LinearRegression().fit(X_train_scaled, y_train)
y_pred_test_scaled = clf.predict(X_test_scaled)
y_pred_train_scaled = clf.predict(X_train_scaled)
#Αποτελέσματα στο σύνολο εκπαίδευσης (train set)
mse_train_split_scaled = mean_squared_error(y_train, y_pred_train_scaled)
mae_train_split_scaled = mean_absolute_error(y_train, y_pred_train_scaled)
mape_train_split_scaled= mape (y_train, y_pred_train_scaled)

#Αποτελέσματα στο σύνολο δοκιμής (test set)
mse_test_split_scaled = mean_squared_error(y_test, y_pred_test_scaled)
mae_test_split_scaled = mean_squared_error(y_test, y_pred_test_scaled)
mape_test_split_scaled= mape (y_test, y_pred_test_scaled)
```

Η πρώτη μέθοδος που χρησιμοποιήσαμε είναι η OLS και παρακάτω δίνονται οι μετρικές αξιολόγησης στο σύνολο εκπαίδευσης και στο σύνολο αξιολόγησης για μη κλιμακοποιημένα και κλιμακοποιημένα δεδομένα.

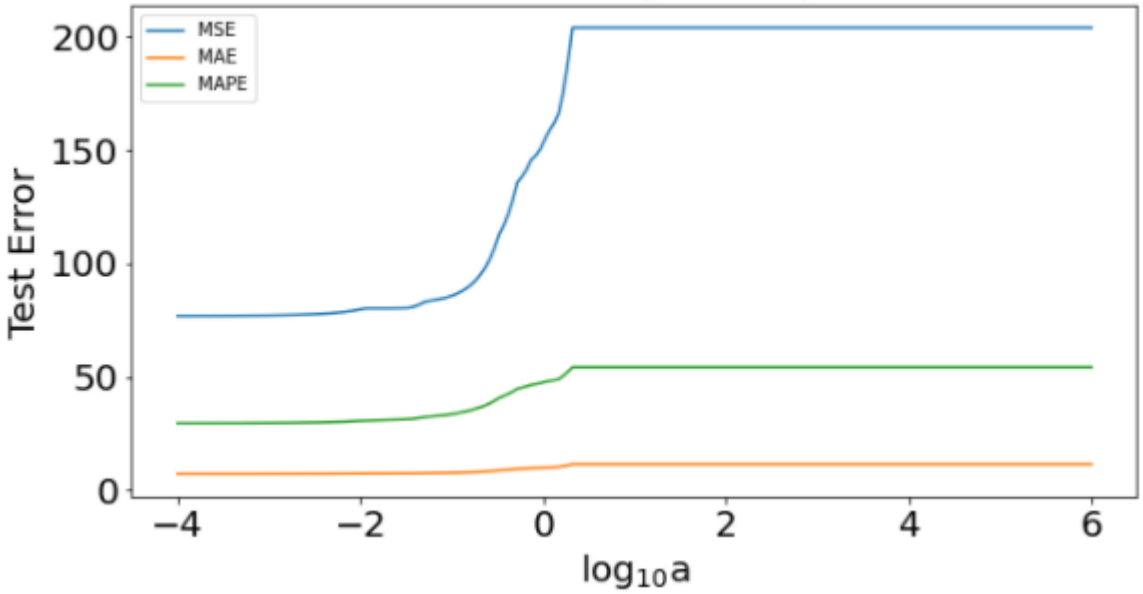
```
Αποτελέσματα της μεθόδου (OLS) σε μη κλιμακοποιημένα δεδομένα
------------------------------------------------------
Mean squared error of OLS on train data: 124.2541
Mean absolute error of OLS train data: 8.9899
Mean absolute percentage error of OLS on train data: 33.3565

Mean squared error of OLS on testing data: 73.2000
Mean absolute error of OLS on testing data: 6.6089
Mean percentage absolute error of OLS on testing data: 25.9313
```

```
Αποτελέσματα της μεθόδου (OLS) σε κλιμακοποιημένα δεδομένα
------------------------------

Mean squared error of OLS on training data: 124.2541
Mean absolute error of OLS on training data: 8.9899
MAPE of OLS on training data: 33.3565

Mean squared error of OLS on test data: 76.7813
Mean absolute error of OLS on test data: 7.0366
MAPE of OLS test data: 29.3867
```

Παρατηρούμε ότι δεν υπάρχει διαφορά στα αποτελέσματα του συνόλου εκπαίδευσης μετά την κλιμακοποίηση των δεδομένων για τη μέθοδο ελαχίστων τετραγώνων, ενώ η απόκλιση μεταξύ των σφαλμάτων στο σύνολο αξιολόγησης είναι σχετικά μικρή. Επίσης, το μέσο τετραγωνικό σφάλμα είναι δυσανάλογα μεγαλύτερο από τα υπόλοιπα, γεγονός που μαρτυρά ότι το μοντέλο κάνει μερικές προβλέψεις που διαφέρουν πολύ από την πραγματική τιμή. Το μεγάλο τετραγωνικό σφάλμα είναι αποτέλεσμα των ακραίων σφαλμάτων.
Στη συνέχεια παρουσιάζονται τα σφάλματα για διάφορες τιμές του βάρους ομαλοποίησης alpha χρησιμοποιώντας τις μεθόδους Ridge και Lasso για την εκπαίδευση του μοντέλου.



Διάγραμμα1: Τα σφάλματα για την μέθοδο Ridge σε συνάρτηση του βάρους ομαλοποίησης



Διάγραμμα2:Τα σφάλματα για την μέθοδο Lasso σε συνάρτηση του βαρους ομαλοποίησης

Από τα παραπάνω διαγράμματα συμπεραίνουμε ότι μία καλή επιλογή της τιμής της υπερπαραμέτρου είναι η **a=0.01.**

Ερώτημα 2

Στο προηγούμενο ερώτημα επιλέξατε το βάρος ομαλοποίησης (υπερπαράμετρος alpha στο scikit-learn) εξετάζοντας τα αποτελέσματα στο σύνολο αξιολόγησης. Ποιο μειονέκτημα έχει αυτή η προσέγγιση; Μπορείτε να προτείνετε άλλες στρατηγικές επιλογής της υπερπαραμέτρου;

Παρατηρούμε ότι για μικρές τιμές της υπερπαραμέτρου alpha τα αποτελέσματα είναι όμοια με τα αντίστοιχα της απλής γραμμικής παλινδρόμησης ενώ για μηδενική τιμή του alpha, οι ποινές στα βάρη μηδενίζονται και έχουμε ταύτιση με την περίπτωση της απλής γραμμικής παλινδρόμησης. Όσο η τιμή του βάρους ομαλοποίησης αυξάνεται οδηγούμαστε σε υψηλότερα σφάλματα καθώς οι ποινές που επιβάλλονται στους συντελεστές των χαρακτηριστικών, είτε τους μειώνουν σημαντικά (L2 ομαλοποίηση) είτε τους μηδενίζουν (L1 ομαλοποίηση) , με συνέπεια να αφαιρούν σημαντική πληροφορία κατά την εκπαίδευση του μοντέλου και να οδηγούν σε λιγότερο ακριβείς προβλέψεις. Αυτό συμβαίνει διότι τα χαρακτηριστικά του προβλήματος είναι σχετικά λίγα συγκριτικά με το σύνολο των δειγμάτων. Επομένως η χρησιμοποίηση των αλγορίθμων Lasso και Ridge δεν διαφοροποιεί σημαντικά τα αποτελέσματα. Το γεγονός αυτό ίσως είναι αποτέλεσμα της στρατηγικής διαχωρισμού στο σύνολο εκπαίδευσης και αξιολόγησης που χρησιμοποιήσαμε(με την σειρά).
Μία ενναλακτική επιλογή της υπερπαραμέτρου είναι να χρησιμοποιήσουμε ένα σύνολο επικύρωσης πριν την τελική αξιολόγηση του μοντέλου(cross validation). Επίσης, θα μπορούσαμε να κάνουμε πιο εκφραστικό το μοντέλο, μεγαλώνοντας την χωρητικότητα του. Μία μέθοδος είναι να εισάγουμε πολυωνυμικούς όρους στα χαρακτηριστικά. Με αυτόν τρόπο προκύπτει μία καινούρια υπερπαράμετρος η οποία αντιστοιχεί στον βαθμό του πολυωνύμου.

Ερώτημα 3

Επαναλάβετε το βήμα 1, με τη διαφορά ότι η αξιολόγηση θα γίνει ως εξής: Επιλέγετε με τυχαίο τρόπο το 70% του συνόλου δεδομένων για εκπαίδευση και το υπόλοιπο 30% για αξιολόγηση και υπολογίζετε τις μετρικές αξιολόγησης. Η διαδικασία αυτή επαναλαμβάνεται 10 φορές και ως αποτέλεσμα δίνετε το μέσο όρο και την τυπική απόκλιση της κάθε μετρικής. Συμφωνούν τα αποτελέσματα με αυτά του βήματος Β;

Για τον τυχαίο διαχωρισμό του συνόλου δεδομένων χρησιμοποιήσαμε την συνάρτηση **ShuffleSplit** από την βιβλιοθήκη sklearn και για τον υπολογισμό του μέσου όρου και της τυπικής απόκλισης των εκάστοτε σφαλμάτων χρησιμοποιήσαμε την συνάρτηση cross_val_score από την βιβλιοθήκη sklearn.

Μέθοδος Παλινδρόμησης	OLS	Ridge	Lasso
Mean_MSE	10.79	11.00	10.35
Standard Deviation MSE	0.40	0.36	0.37
Mean_MAE	2.92	2.93	2.96
Standard_Deviation_MAE	0.06	0.07	0.04

Πίνακας1: Μέση τιμή και τυπική απόκλιση των σφαλμάτων MSE και MAE για τις μεθόδους OLS, Ridge και Lasso.

Παρατηρούμε ότι τα σφάλματα διαφέρουν αισθητά από εκείνα που υπολογίσαμε στο προηγούμενο ερώτημα και είναι μικρότερα. Ωστόσο, δεν παρουσιάζεται σημαντική διαφοροποίηση μεταξύ των τριών μεθόδων.

Επειδή τα χαρακτηριστικά του προβλήματος εμφανίζουν θετική κλίση (positive Skewness**),** ο τυχαίος διαχωρισμός των δεδομένων βοήθησε στην καλύτερη εκπαίδευση του μοντέλου. Επίσης το μέσο τετραγωνικό σφάλμα έχει εμφανίσει σημαντική πτώση, που σημαίνει ότι τα μεγάλα σφάλματα μειώθηκαν σε σημαντικό βαθμό.

Σημείωση: Λόγω αδυναμίας της έκδοσης δεν έγινε ο υπολογισμός της Mape score.

Ερώτημα 4

Δεδομένης της μη γραμμικότητας της συνάρτησης που προσπαθούμε να μοντελοποιήσουμε, αξίζει να αξιολογήσουμε και πιο εκφραστικά μοντέλα γραμμικής παλινδρόμησης με πολυωνυμικούς όρους των χαρακτηριστικών.

Υλοποιήστε συνάρτηση ***test_poly_regression(X_train, y_train, X_test, y_test, n=2)*** *Η οποία θα δέχεται ως είσοδο ένα σύνολο εκπαίδευσης (X_train πίνακας σχεδιασμού του συνόλου εκπαίδευσης και y_train η εξαρτημένη μεταβλητή), ένα σύνολο αξιολόγησης (X_test, y_test), και έναν βαθμό πολυωνύμου* $n \geq 1$*. Η συνάρτηση θα πρέπει να δημιουργεί ένα νέο σύνολο χαρακτηριστικών που αποτελείται από τα αρχικά χαρακτηριστικά και εκδοχές τους υψωμένες σε δυνάμεις έως* n*. Συγκεκριμένα αν* $\boldsymbol{X}$ *είναι το αρχικό σύνολο, η συνάρτηση δημιουργεί το σύνολο* $\boldsymbol{Xn} = [\boldsymbol{X}\ \boldsymbol{X2} \ldots \boldsymbol{Xn}]$ *Αυτό συμβαίνει τόσο για το σύνολο εκπαίδευσης όσο και για το σύνολο αξιολόγησης. Έπειτα, η συνάρτηση εκπαιδεύει και αξιολογεί μοντέλα γραμμικής παλινδρόμησης στα σύνολα που προκύπτουν. Εκτελέστε τη διαδικασία αυτή για* $n = 1$ *έως* $n = 10$*. Μπορείτε να χρησιμοποιήσετε οποιονδήποτε από τους τρόπους αξιολόγησης των προηγούμενων υποερωτημάτων*

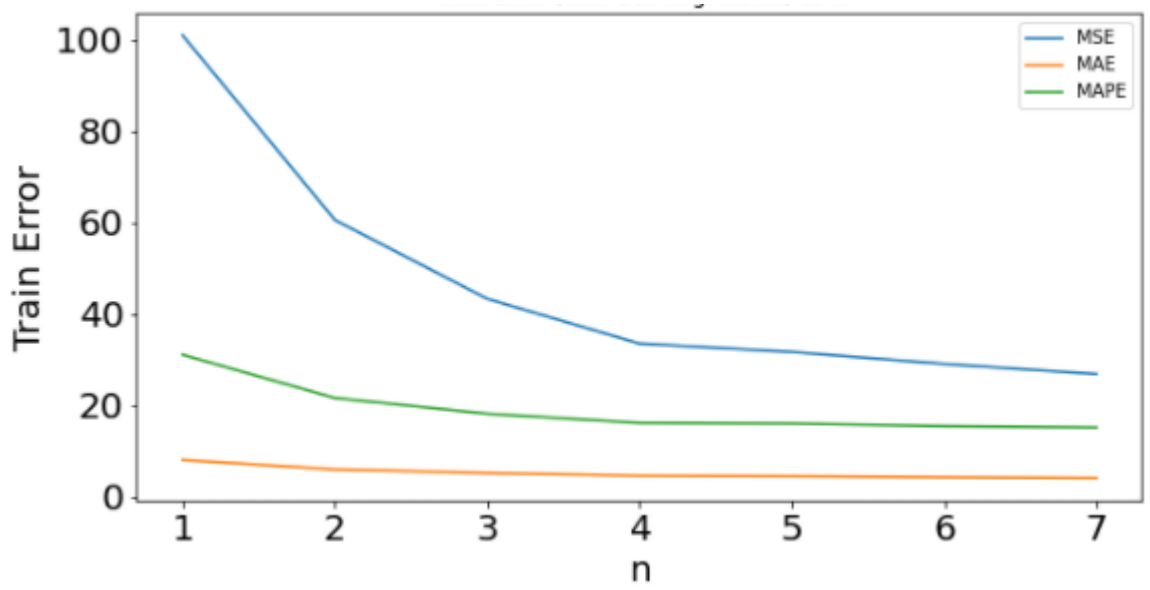
(σταθερό σύνολο εκπαίδευσης ή τυχαία επιλογή επαναληπτικά). Τι παρατηρείτε; Με βάση αυτά τα αποτελέσματα, ποιο μοντέλο θα επιλέγατε για πρακτική εφαρμογή; Επίσης, ποια είναι τα πλεονεκτήματα και ποια τα μειονεκτήματα των μοντέλων με υψηλότερους βαθμούς πολυωνύμου, n.

Αξιολόγηση των μοντέλων γραμμικής παλινδρόμησης εισάγοντας πολυωνυμικούς όρους στα χαρακτηριστικά:

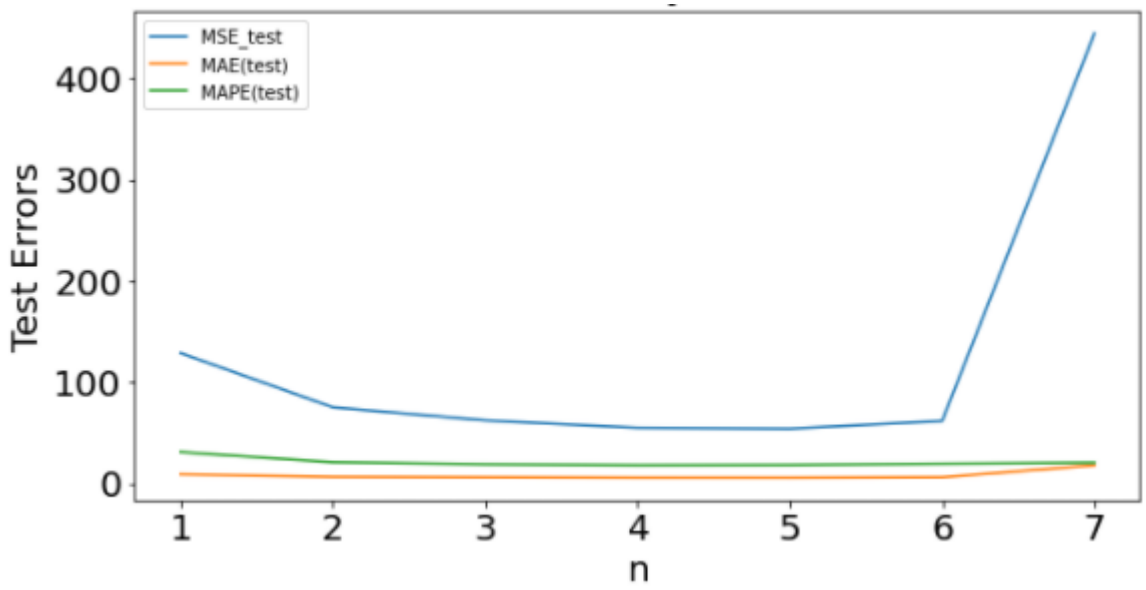
Ο τρόπος αξιολόγησης που θα χρησιμοποιήσουμε είναι σε σταθερό σύνολο αξιολόγησης. Ο διαχωρισμός των δεδομένων γίνεται με τυχαίο τρόπο έτσι ώστε το σύνολο αξιολόγησης να αποτελεί το 30% του συνόλου δεδομένων. Στη συνέχεια κατασκευάζουμε την συνάρτηση **test_poly_regression** ως εξής:

```
# Ορισμός πολυωνυμικής Συνάρτησης
def test_poly_regression(X_train,y_train,X_test,y_test,n,model):

    X_train_new = X_train.copy()
    X_test_new =X_test.copy()
    y_train_new = y_train.copy()
    y_test_new= y_test.copy()
    #Κατασκευή των πολυωνυμικών όρων για δωσμένο n
    for p in range(2, n+1):
        X_p = X_train**p
        X_train_new = np.hstack((X_train_new, X_p))
        Xt_p = X_test**p
        X_test_new = np.hstack((X_test_new, Xt_p))
        y_p = y_train**p
        y_train_new = np.hstack((y_train_new, y_p))
        yt_p = y_test**p
        y_test_new = np.hstack((y_test_new, yt_p))
    #Κλιμακοποίηση του τελικού συνόλου
    scaler = MinMaxScaler()
    scaler.fit(X_train_new)
    scaler.fit(X_test_new)
    X_train_scaled = scaler.fit_transform(X_train_new)
    X_test_scaled = scaler.fit_transform(X_test_new)
    #Εκπαίδευση του μοντέλου επί του τελικού συνόλου
    clf1 = model.fit(X_train_scaled, y_train)
    # Προβλέψεις
    y_pred_test_scaled = clf1.predict(X_test_scaled)
    y_pred_train_scaled = clf1.predict(X_train_scaled)
    #Αποτελέσματα στο σύνολο αξιολόγησης
    mse_test = mean_squared_error(y_test, y_pred_test_scaled)
    mae_test = mean_absolute_error(y_test, y_pred_test_scaled)
    mape_test=mape(y_test, y_pred_test_scaled)
    #Αποτελέσμα στο σύνολο εκπαίδευσης
    mse_train = mean_squared_error (y_train, y_pred_train_scaled)
    mae_train = mean_absolute_error (y_train, y_pred_train_scaled)
    mape_train = mape (y_train, y_pred_train_scaled)

    # Τα σφάλματα που επιστρέφει η συνάρτηση
    return mse_test,mae_test,mape_test, mse_train,mae_train,mape_train
```

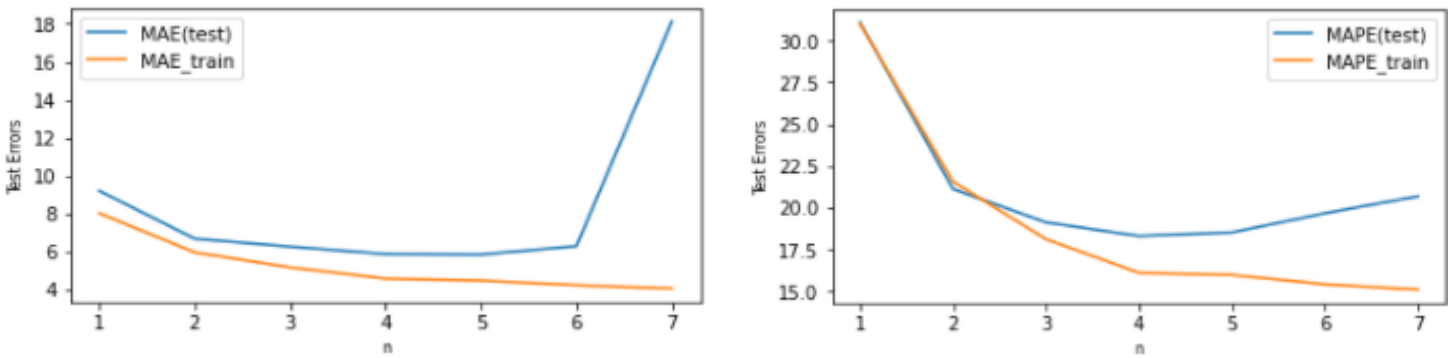
Απλή Γραμμική Παλινδρόμηση:



Γράφημα 1:Τα σφάλματα επί του συνόλου εκπαίδευσης για τους βαθμούς του πολυωνύμου στην μέθοδο OLS

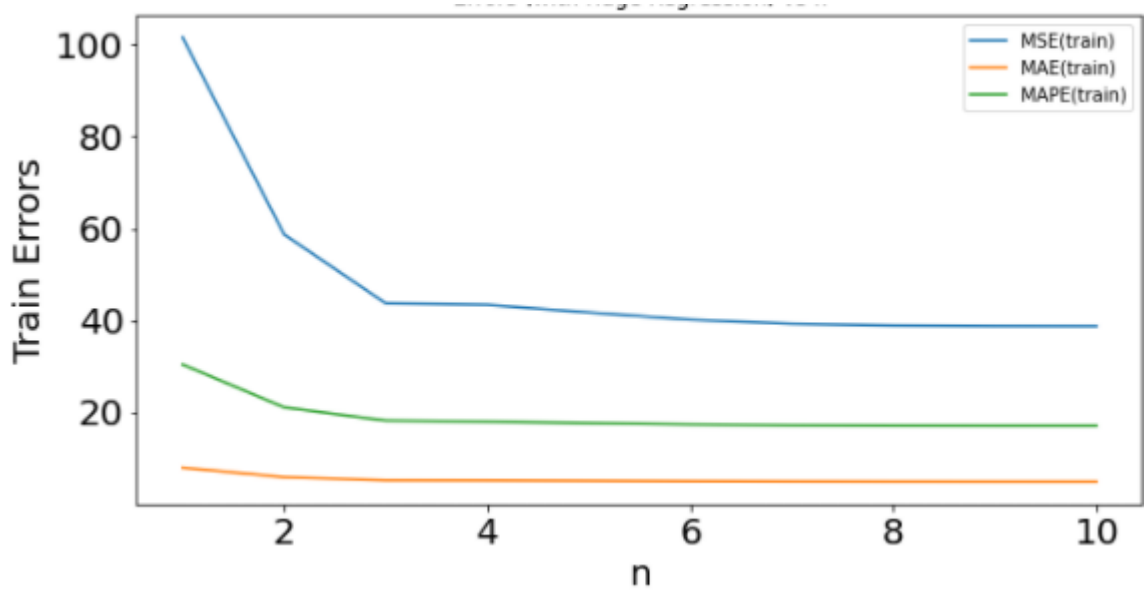


Γράφημα 2:Τα σφάλματα επί του συνόλου αξιολόγησης για την μέθοδο OLS σε συνάρτηση του βαθμού πολυωνύμου

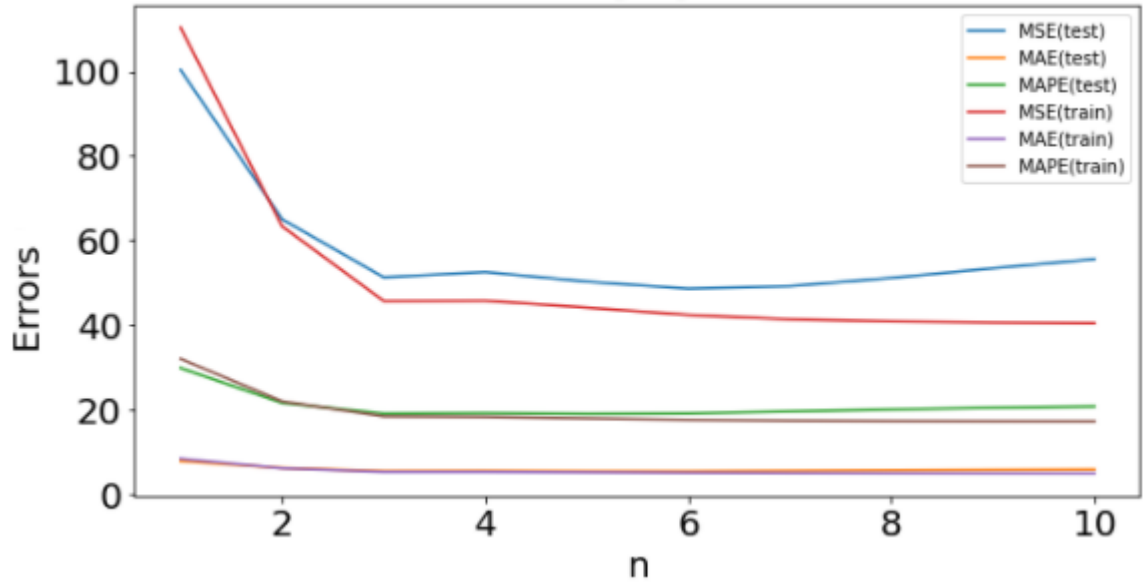


Γράφημα 3:Σύγκριση των test και train errors για την μέθοδο OLS συναρτήσει του πολυωνυμικού όρου

Γραμμική παλινδρόμηση Ridge:

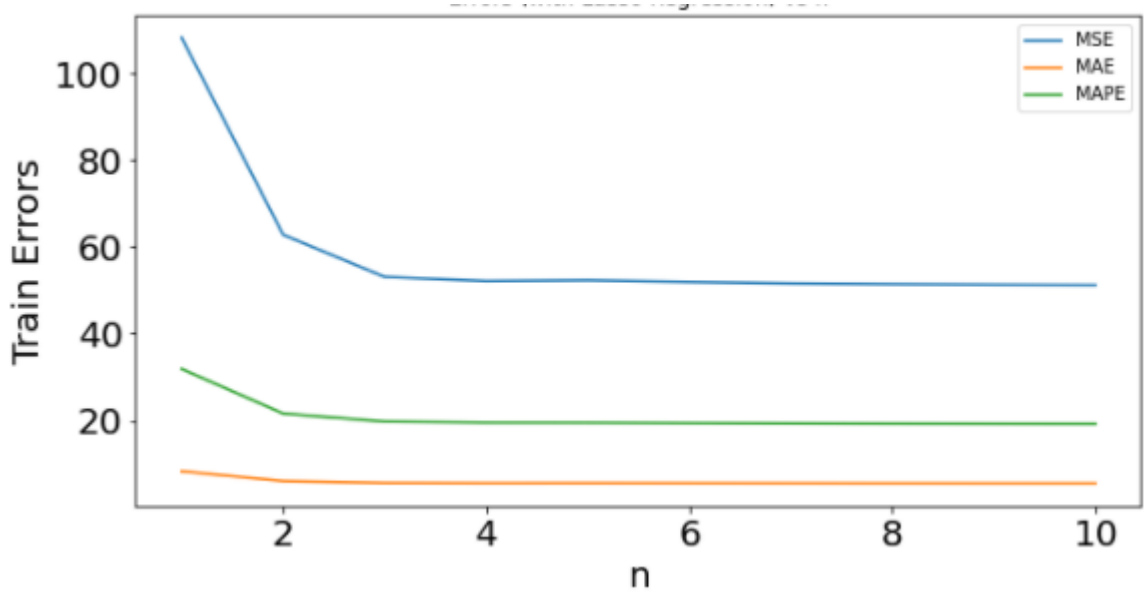


Γράφημα 4: Τα σφάλματα επί του συνόλου εκπαίδευσης για τους βαθμούς του πολυωνύμου στην μέθοδο Ridge

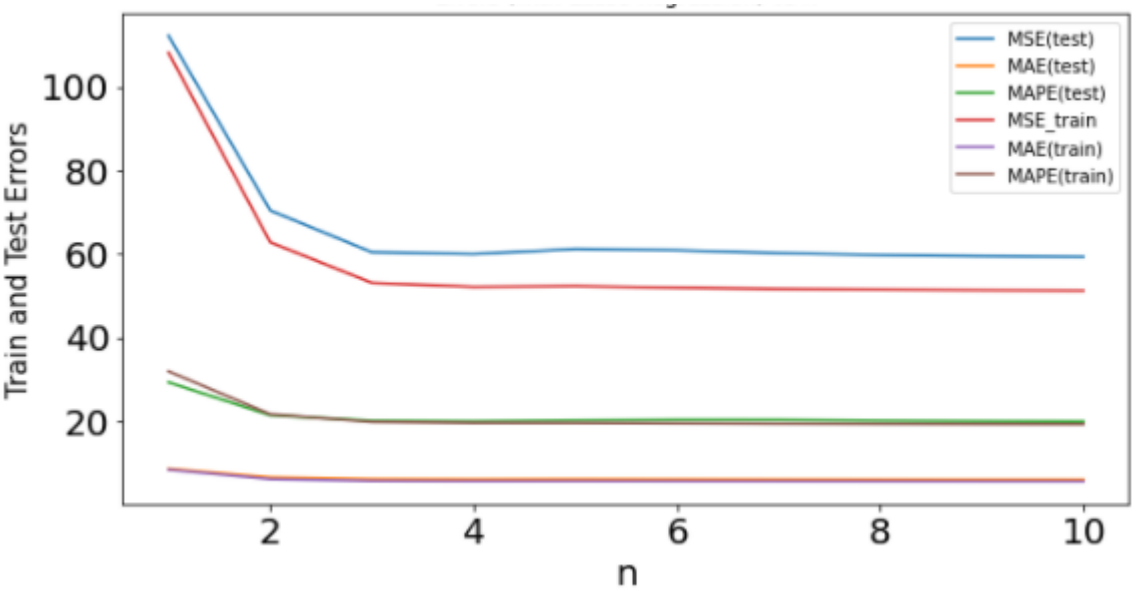


Γράφημα 5: Σύγκριση των test και train errors για την μέθοδο Ridge συναρτήσει του πολυωνυμικού όρου

Γραμμική Παλινδρόμηση Lasso:



Γράφημα 6:Τα σφάλματα επί του συνόλου εκπαίδευσης για τους βαθμούς του πολυωνύμου στην μέθοδο Lasso



Γράφημα 7: Σύγκριση των test και train errors στην μέθοδο Lasso συναρτήσει του βαθμού πολυωνύμου

Σχολιασμός των Αποτελεσμάτων:

Τα μοντέλα με μεγαλύτερη χωρητικότητα, δηλαδή τα μοντέλα όπου εισάγουμε πολυωνυμικούς όρους στα χαρακτηριστικά τους είναι πιθανό να οδηγηθούν σε υπερ-εκπαίδευση(overfitting), όπως φαίνεται ξεκάθαρα στα γραφήματα 2,3,5 και 7. Ωστόσο, η επιλογή της κατάλληλης τιμής της υπερπαραμέτρου (βαθμός πολυωνύμου) βελτιώνει σημαντικά το μοντέλο και αποτρέπει την υπο-εκπαίδευσή (underfitting) του, όπως φαίνεται ξεκάθαρα στα προηγούμενα γραφήματα από την κατακόρυφη πτώση της καμπύλης που εκφράζει το αντίστοιχο σφάλμα ακόμα και για μικρούς βαθμούς πολυωνύμου.

Η στρατηγική αυτή είναι απαραίτητη όταν ο αριθμός των χαρακτηριστικών του προβλήματος είναι μικρός συγκριτικά με τον αριθμό των δειγμάτων. Ένα ακόμα χαρακτηριστικό που πρέπει να λάβουμε υπόψιν είναι ότι επιδιώκουμε την μεγαλύτερη απόδοση για όσο το δυνατόν μικρότερη χωρητικότητα.

Συγκρίνοντας τα παραπάνω διαγράμματα γίνεται αντιληπτό ότι η εισαγωγή πολυωνυμικών όρων στα χαρακτηριστικά βελτίωσε την απόδοση του μοντέλου. Σύμφωνα με τα αποτελέσματα, δεν παρατηρούνται σημαντικές διαφορές μεταξύ της απλής γραμμικής παλινδρόμησης και της μεθόδου Ridge, σε αντίθεση με την μέθοδο Lasso που έχει την μικρότερη απόδοση. Η καταλληλότερη επιλογή της υπερπαραμέτρου n(βαθμός πολυωνύμου) είναι n_{Ridge}=4 και n_{OLS}=3, για την αποφυγή της υπερ-εκπαίδευσης (overfitting) ή της υπο-εκπαίδευσης (underfitting). Για τις ίδιες τιμές της υπερπαραμέτρου n η μέθοδος Lasso δεν έχει τόσο καλή απόδοση. Αυτό αποδίδεται στο γεγονός ότι σχεδόν όλα τα χαρακτηριστικά έχουν υψηλή συσχέτιση με την πρόβλεψη.

Συμπεράσματα

- Τα δεδομένα θα πρέπει να κλιμακοποιηθούν κατάλληλα, λαμβάνοντας υπόψιν την κλίση που παρουσιάζει το κάθε χαρακτηριστικό και την ύπαρξη ακραίων τιμών.
- Ο διαχωρισμός στα σύνολα εκπαίδευσης και αξιολόγησης πρέπει να γίνεται με τυχαίο τρόπο.
- Συνυπολογίζοντας τα αποτελέσματα της ανάλυσης και τα χαρακτηριστικά του συνόλου δεδομένων, συμπεραίνουμε ότι η καταλληλότερη επιλογή για την πρόβλεψη της αντοχής του σκυροδέματος φαίνεται να είναι η μέθοδος Ridge για a=0.01.
- Για αποφυγή της επανεκπαίδευσης πρέπει να αυξηθεί η χωρητικότητα του μοντέλου, με μία πολυωνυμική συνάρτηση με βαθμό πολυωνύμου n=4.
- Η κατάλληλη μετρική αξιολόγησης προκύπτει ότι είναι η MAPE διότι ταιριάζει καλύτερα με την κλίμακα των δεδομένων και δεν έχει ισχυρή ευαισθησία στα μεγάλα σφάλματα όπως η MSE. Η τιμή της για την μέθοδο Ridge(a=0.01, n_{Ridge}=4) και για την μέθοδο OLS για n_{OLS}=3 κυμαίνεται κάτω από 20%, το οποίο σημαίνει ότι το μοντέλο δουλεύει ικανοποιητικά.