Τεχνικές Μηχανικής Μάθησης

Εργασία 2- Support Vector Machines

Ταμπάκη Ειρήνη- Μαρία (401)
Πασιοπούλου Ιωάννα (396)

24/5/2018

Α. Εισαγωγή

Η παρούσα εργασία πραγματεύεται τη δημιουργία μοντέλων ταξινόμησης με χρήση των Support Vector Machines για δεδομένο data set, πιο συγκεκριμένα για το Energy Efficiency Dataset όπως δόθηκε από την εκφώνηση της εργασίας.

Β. Δεδομένα

Το data set πραγματεύεται την ενεργειακή απόδοση των κτιρίων έχοντας ως δεδομένα διάφορες παραμέτρους τους (το πόσο συμπαγή είναι, το εμβαδό του κτιρίου, των τοίχων, της οροφής, το συνολικό ύψος, τον προσανατολισμό, το εμβαδό των τζαμιών και την κατανομή των τζαμιών, ενώ οι δύο εξαρτημένες είναι το φορτίο ψύξης και θέρμανσης). Έχει 768 δείγματα, 8 ανεξάρτητες μεταβλητές και δύο εξαρτημένες. Αποφασίσαμε να κάνουμε ταξινόμηση για την κατανάλωση ενέργειας ανεξαρτήτως του αν είναι θέρμανσης ή ψύξης.

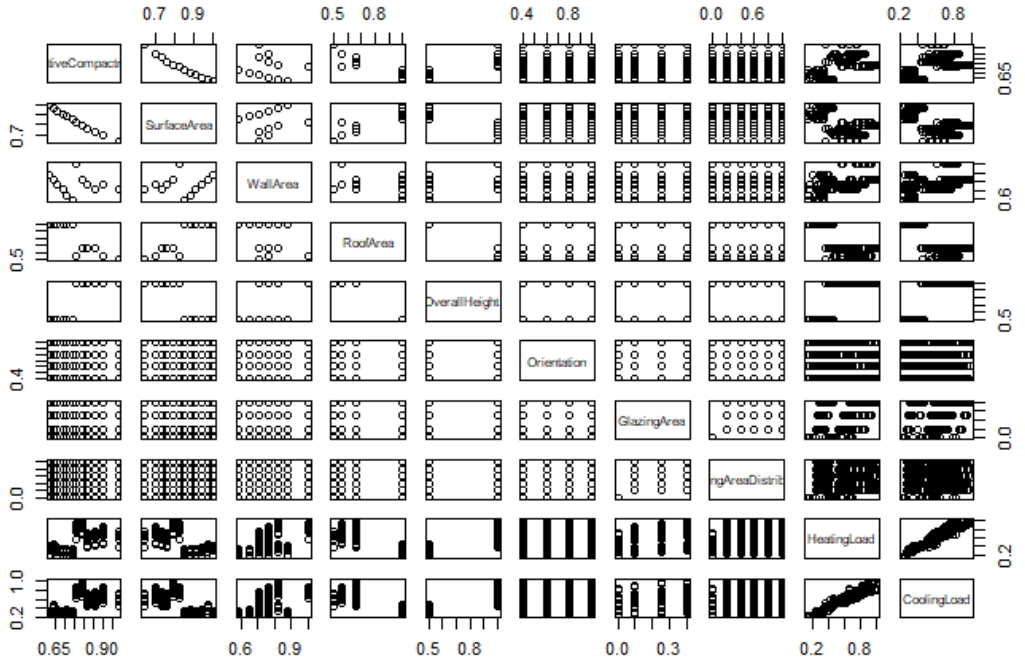
Αρχικά κάναμε μια επισκόπηση των δεδομένων:

```
> summary(en)
 RelativeCompactness  SurfaceArea       WallArea        RoofArea     OverallHeight
 Min.   :0.6200       Min.   :514.5   Min.   :245.0   Min.   :110.2   Min.   :3.50
 1st Qu.:0.6825       1st Qu.:606.4   1st Qu.:294.0   1st Qu.:140.9   1st Qu.:3.50
 Median :0.7500       Median :673.8   Median :318.5   Median :183.8   Median :5.25
 Mean   :0.7642       Mean   :671.7   Mean   :318.5   Mean   :176.6   Mean   :5.25
 3rd Qu.:0.8300       3rd Qu.:741.1   3rd Qu.:343.0   3rd Qu.:220.5   3rd Qu.:7.00
 Max.   :0.9800       Max.   :808.5   Max.   :416.5   Max.   :220.5   Max.   :7.00
 Orientation GlazingArea GlazingAreaDistribution  HeatingLoad     CoolingLoad
 2,00:192    0,00: 48    Min.   :0.000           Min.   : 6.01   Min.   :10.90
 3,00:192    0,10:240    1st Qu.:1.750           1st Qu.:12.99   1st Qu.:15.62
 4,00:192    0,25:240    Median :3.000           Median :18.95   Median :22.08
 5,00:192    0,40:240    Mean   :2.812           Mean   :22.31   Mean   :24.59
                         3rd Qu.:4.000           3rd Qu.:31.67   3rd Qu.:33.13
                         Max.   :5.000           Max.   :43.10   Max.   :48.03
```

Βλέποντας τις μεταβλητές κρίναμε ότι όλες χρειάζονται για καλύτερη πρόβλεψη και η μόνη επεξεργασία που κάναμε ήταν να τις κανονικοποιήσουμε. Οι “Orientation”, “GlazingArea”, “GlazingAreaDistribution” είναι κατηγορικές.

Μετά από κανονικοποίηση πήραμε τα παρακάτω:

```
> summary(enrg)
 RelativeCompactness  SurfaceArea       WallArea          RoofArea      OverallHeight
 Min.   :0.6200      Min.   :0.6364   Length:768         Min.   :0.5000   Min.   :0.50
 1st Qu.:0.6825      1st Qu.:0.7500   Class :character   1st Qu.:0.6389   1st Qu.:0.50
 Median :0.7500      Median :0.8333   Mode  :character   Median :0.8333   Median :0.75
 Mean   :0.7642      Mean   :0.8308                      Mean   :0.8009   Mean   :0.75
 3rd Qu.:0.8300      3rd Qu.:0.9167                      3rd Qu.:1.0000   3rd Qu.:1.00
 Max.   :0.9800      Max.   :1.0000                      Max.   :1.0000   Max.   :1.00
  Orientation     GlazingArea     GlazingAreaDistribution  HeatingLoad     CoolingLoad
 Min.   :0.40   Min.   :0.0000   Min.   :0.0000          Min.   :0.1394   Min.   :0.2269
 1st Qu.:0.55   1st Qu.:0.1000   1st Qu.:0.3500          1st Qu.:0.3015   1st Qu.:0.3252
 Median :0.70   Median :0.2500   Median :0.6000          Median :0.4397   Median :0.4597
 Mean   :0.70   Mean   :0.2344   Mean   :0.5625          Mean   :0.5176   Mean   :0.5119
 3rd Qu.:0.85   3rd Qu.:0.4000   3rd Qu.:0.8000          3rd Qu.:0.7347   3rd Qu.:0.6898
 Max.   :1.00   Max.   :0.4000   Max.   :1.0000          Max.   :1.0000   Max.   :1.0000
```



Αυτό που κάναμε στη συνέχεια ήταν να κατηγοροιοπηήσουμε τις εξαρτημένες μεταβλητές. Πιο συγκεκριμένα τις χωρίσαμε σε τρεις κλάσεις: χαμηλής, μέσης και υψηλής κατανάλωσης ισόποσα και για το φορτίο θέρμανσης και για το φορτίο ψύξης.

- LOW =< 1/3
- 1/3 < MEDIUM < 2/3
- HIGH >= 2/3

Επίσης δημιουργήσαμε μία νέα μεταβλητή που ισούται με τη μέγιστη κατανάλωση είτε σε θέρμανση, είτε σε ψύξη. Δηλαδή για κάθε εγγραφή κρατάει το μέγιστο των δύο αν είναι διαφοτερικά ή την τιμή τους αν είναι ίδια. Με αυτόν το τρόπο κατηγοριοποιούμε τα κτίρια σε χαμηλής, μέσης και υψηλής κατανάλωσης

ανεξαρτήτως του είδους του φορτίου και θεωρώντας ως δεδομένο ότι η θέρμανση και η ψύξη δε θα είναι ποτέ ταυτόχρονα μέγιστες και στην καλύτερη περίπτωση η ζήτηση του ενός δε θα είναι ταυτόχρονη με τη ζήτηση του άλλου.

Για πέντε διαφορετικά seed το dataset χωρίστηκε σε train και validation set και μοκιμαστήκανε τέσσερα διαφορετικά kernels: radial, polynomial, linear, sigmoidal για τα ίδια εύρη cost και c, ώστε να βρεθούν οι καλύτεροι παράμετροι, σε 10 fold cross validation:

	seed	kernel	cost	gamma	performance
1	2	radial	10	0,5	0,062402
2	2	polynomial	1	1	0,050423
3	2	linear	1	0,5	0,183696
4	2	sigmoid	0,1	0,5	0,378342
5	50	radial	100	0,5	0,062553
6	50	polynomial	0,1	2	0,046945
7	50	linear	10	0,5	0,140321
8	50	sigmoid	0,1	0,5	0,359316
9	356	radial	10	0,5	0,060678
10	356	polynomial	10	0,5	0,052027
11	356	linear	10	0,5	0,147641
12	356	sigmoid	0,1	0,5	0,338596
13	2002	radial	10	0,5	0,062492
14	2002	polynomial	0,1	2	0,041742
15	2002	linear	1	0,5	0,175348
16	2002	sigmoid	0,1	0,5	0,388838
17	12345	radial	10	0,5	0,074592
18	12345	polynomial	0,1	2	0,060738
19	12345	linear	10	0,5	0,159952
20	12345	sigmoid	0,1	0,5	0,373412

Για τα καλύτερα cost και c, δημιουργούμε τα 4 μοντέλα για κάθε kernel.

Για το κάθε μοντέλο και για κάθε seed υπολογίσαμε τα accuracy, precision, recall και f-measure. Με βάση το accuracy, το recall, αλλά και τη διασπορά του για τα διαφορετικά seed που θέσαμε επιλέξαμε τα καλύτερα μοντέλα προς την κατεύθυνση του καλύτερου συνδυασμού για να έχουμε υψηλά τα πρώτα και χαμηλό το δεύτερο. Στου δύο παρακάτω πίνακες φαίνεται το accuracy για κάθε μοντέλο και seed και ο μέσος όρος του recall για κάθε μοντέλο.

seed	Radial		Polynomial		Linear		Sigmoid	
	Train	Test	Train	Test	Train	Test	Train	Test
2	0.9896	0.9531	0.9757	0.9792	0.8524	0.9219	0.5955	0.5885
50	0.9983	0.9219	0.9757	0.9375	0.8802	0.8490	0.6094	0.5990
356	0.9913	0.9271	0.9878	0.9479	0.8594	0.8229	0.6302	0.5573
2002	0.9913	0.9427	0.9809	0.9375	0.8611	0.8698	0.5938	0.5677
12345	0.9896	0.9583	0.9740	0.9635	0.8403	0.7604	0.5972	0.6510
σ^2:	0.0000	0.0003	0.0000	0.0003	0.0002	0.0035	0.0002	0.0013
Mean:	0.9920	0.9406	0.9788	0.9531	0.8587	0.8448	0.6052	0.5927

		Recall
Radial	Train	0.991777
	Test	0.938001
Polynomial	Train	0.978546
	Test	0.950119
Linear	Train	0.860528
	Test	0.847609
Sigmoid	Train	0.616689
	Test	0.612106

Με βάση τη διασπορά επιλέγουμε το καλύτερο μοντέλο ως αυτό που έχει τη μικρότερη για τα διαφορετικά seed, καθώς θέλουμε το μοντέλο να λειτουργεί το ίδιο ανεξαρτήτως του πως του δίνουμε τα δεδομένα. Αλλά θέλοντας να αποφύγουμε το overfitting και να έχουμε καλύτερη απόδοση του μοντέλου στα δεδομένα ελέγχου διαλέγουμε ως καλύτερο το πολυωνυμικό, καθώς έχει καλύτερο accuracy για τα δεδομένα ελέγχου και η διαφορά του με αυτό των δεδομένων εκπαίδευσης είναι μικρότερη από αυτή του radial μοντέλου. Επίσης, είναι καλύτερο και στη μετρική Recall καθώς δίνει καλύτερη τιμή για το σετ ελέγχου. Οι πλήρεις μετρικές για το polynomial:

	Accuracy		Precision					
seed	Train	Test	Train			Test		
			HIGH	MEDIUM	LOW	HIGH	MEDIUM	LOW
2	0.9757	0.9792	0.9950	0.9646	0.9669	1.0000	0.9595	0.9815
50	0.9757	0.9375	0.9949	0.9739	0.9527	1.0000	0.9571	0.8421
356	0.9878	0.9479	1.0000	0.9822	0.9810	0.9857	0.9333	0.9149
2002	0.9809	0.9375	0.9949	0.9731	0.9742	1.0000	0.9091	0.9000
12345	0.9740	0.9635	0.9948	0.9682	0.9573	1.0000	0.9500	0.9268
σ^2:	0.0000	0.0003	0.0000	0.0000	0.0001	0.0000	0.0004	0.0025
Mean:	0.9800	0.9505	0.9962	0.9735	0.9687	0.9964	0.9398	0.9096

	Recall						F measure					
seed	Train			Test			Train			Test		
	HIGH	MEDIUM	LOW	HIGH	MEDIUM	LOW	HIGH	MEDIUM	LOW	HIGH	MEDIUM	LOW
2	0.9900	0.9732	0.9605	1.0000	0.9861	0.9464	0.9925	0.9689	0.9637	1.0000	0.9726	0.9636
50	0.9899	0.9655	0.9724	1.0000	0.8816	0.9412	0.9924	0.9697	0.9625	1.0000	0.9178	0.8889
356	0.9897	0.9866	0.9873	1.0000	0.9333	0.8958	0.9948	0.9844	0.9841	0.9928	0.9333	0.9053
2002	0.9949	0.9775	0.9679	0.9848	0.9333	0.8824	0.9949	0.9753	0.9711	0.9924	0.9211	0.8911
12345	0.9896	0.9638	0.9691	1.0000	0.9620	0.9048	0.9922	0.9660	0.9632	1.0000	0.9560	0.9157
σ^2:	0.0000	0.0001	0.0001	0.0000	0.0015	0.0008	0.0000	0.0001	0.0001	0.0000	0.0006	0.0009
Mean:	0.9912	0.9757	0.9720	0.9962	0.9336	0.9164	0.9937	0.9746	0.9703	0.9963	0.9362	0.9122

Σε συνημμένο υπαρχουν δύο αρχεία .xlsx. Το “myparameters” περιέχει τα αποτελέμσατα του tuning για τα cost και c. Το “mydata” περιέχει τα αποτελέσματα των μετρικών για τα τέσσερα μοντέλα.