Υπολογιστική Νοημοσύνη

4η εργασία

Επίλυση Προβλήματος Παλινδρόμησης με χρήση RBF δικτύου

Γεώργιος Τσουμπλέκας, gktsoump@ece.auth.gr, ΑΕΜ: 9359, ΝΠΣ

---

0 Εισαγωγή

Στην συγκεκριμένη εργασία καλούμαστε να δημιουργήσουμε και να μελετήσουμε την λειτουργία ενός RBF δικτύου για την πραγματοποίηση προβλέψεων στο Boston Housing dataset. Πιο συγκεκριμένα, στο 1ο σκέλος της εργασίας δημιουργούμε 3 διαφορετικά δίτυα RBF τα οποία διαφέρουν ως προς τον αριθμό πυρήνων στο κρυφό στρώμα και τα συγκρίνουμε ως προς την ικανότητα πρόβλεψης τους. Στο 2ο σκέλος της εργασίας καλούμαστε να κάνουμε fine-tuning ενός RBF δικτύου ως προς ορισμένες υπερ-παραμέτρους του. Το fine-tuning πραγματοποιείται με grid search και cross-validation στα μοντέλα που προκύπτουν και τέλος εξετάζεται η αποδοτικότητα του βέλτιστου μοντέλου.

1 Υλοποίηση του RBF δικτύου

Αρχικά, όλες οι υλοποιήσεις που θα παρουσιαστούν πραγματοποιήθηκαν με χρήση των frameworks Tensorflow και Keras σε συνδυασμό με κάποιες λειτουργίες από άλλες βιβλιοθήκες όπως η scipy και η sklearn.

Καθώς δεν υπάρχει στο Keras κάποιο έτοιμο μοντέλο ή στρώμα για RBF δίκτυα καλούμαστε να δημιουργήσουμε εμείς ένα custom layer που να υλοποιεί την λειτουργία του κρυφού στρώματος. Στην συγκεκριμένη υλοποίηση, προκειμένου να διευκολύνουμε την διαδικασία δημιουργίας του μοντέλου αποφασίστηκε να δημιουργηθεί ένα μοντέλο το οποίο θα περιέχει μόνο το στρώμα εξόδου (αξιοποιώντας την κλάση Layer του Keras) και του οποίου η είσοδος θα είναι η έξοδος που θα έδινε το κρυφό στρώμα αν υπήρχε εντός του μοντέλου. Έτσι, το κρυφό στρώμα ουσιαστικά υλοποιείται και εκπαιδεύεται εκτός του μοντέλου, δέχεται τα δεδομένα εισόδου του training set, παράγει την έξοδο του και στην συνέχεια αυτή τροφοδοτείται στο μοντέλο για να παραχθεί το τελικό αποτέλεσμα.

Ο λόγος που κάτι τέτοιο είναι εφικτό είναι ότι τα 2 στρώματα των RBF μοντέλων (κρυφό και εξόδου) εκπαιδεύονται με διαφορετικό τρόπο: το κρυφό στρώμα με χρήση κάποιου αλγορίθμου συσταδοποίησης (όπως ο KMeans) ενώ το στρώμα εξόδου με back-propagation. Έτσι, σε πρώτη φάση εφαρμόζουμε τον αλγόριθμο συσταδοποίησης στα δεδομένα εκπαίδευσης για

να δημιουργήσουμε τους πυρήνες του κρυφού στρώματος και στην συνέχεια εκπαιδεύουμε το στρώμα εξόδου όπως θα εκπαιδεύαμε ένα στρώμα ενός MLP δικτύου αλλά τροφοδοτώντας το με τα μετασχηματισμένα δεδομένα εκπαίδευσης τα οποία έχουν περάσει από τους πυρήνες του κρυφού στρώματος.

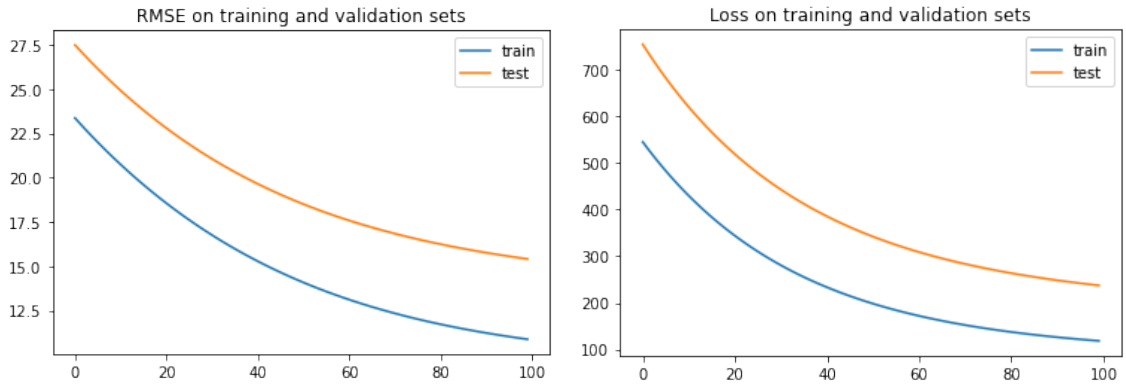
2 Απλή εφαρμογή σε RBF δίκτυο

Στην παρούσα ενότητα παρουσιάζονται 3 διαφορετικά μοντέλα RBF τα οποία θα συγκρίνουμε ως προς την ικανότητα προσαρμογής τους. Και για τα 3 δίκτυα χρησιμοποιούμε Γκαουσσιανούς πυρήνες $f(x) = \frac{e^{-\|x-c_i\|^2}}{2c_i^2}$ στο κρυφό στρώμα με κέντρα τα οποία επιλέγονται με χρήση του KMeans και διασπορά ίση με $\sigma = \frac{d_{max}}{\sqrt{2P}}$, όπου d_{max} η μέγιστη απόσταση μεταξύ των κέντρων των πυρήνων και P το πλήθος των δεδομένων εισόδου του training set. Το στρώμα εξόδου αποτελείται από 128 νευρώνες και για την εκπαίδευση του χρησιμοποιήθηκε ο SGD optimizer με ρυθμό μάθησης η=0.001 για 100 εποχές. Το batch size ορίστηκε στο 1 (online μάθηση) η οποία παρότι πιο αργή παράγει καλύτερα αποτελέσματα. Το 75% των δεδομένων χρησιμοποιείται για την εκπαίδευση των μοντέλων και το υπόλοιπο 25% για την αξιολόγησή τους. Επιπλέον, ένα 20% του συνόλου εκπαίδευσης χρησιμοποιείται για την επικύρωση κάθε μοντέλου. Ως συνάρτηση κόστους ορίζεται το μέσο τετραγωνικό σφάλμα ενώ ως δείκτες αξιολόγησης χρησιμοποιούνται το RMSE και το R^2. Το μοναδικό σημείο στο οποίο διαφέρουν τα 3 δίκτυα είναι ως προς το πλήθος των πυρήνων στο κρυφό τους στρώμα το οποίο ορίζεται ως ένα ποσοστό του πλήθους των δεδομένων του συνόλου εκπαίδευσης.

2.1 Μοντέλο 1

Το πλήθος των πυρήνων στο μοντέλο αυτό είναι ίσο με το 10% του πλήθους των δεδομένων εκπαίδευσης, δηλαδή 37 στο σύνολο. Επιπλέον, στο στρώμα εξόδου έχουμε 4864 βάρη προς εκπαίδευση. Οι αριθμοί αυτοί είναι σχτικά μικροί και έτσι αναμένουμε το μοντέλο να εκπαιδευτεί σχετικά γρήγορα αλλά με κόστος την μικρότερη αποδοτικότητα του σε σχέση με κάποιο πιο σύνθετο μοντέλο. Στα Σχ.1α,β φαίνονται οι καμπύλες μάθησης και RMSE για τα training και validation sets.

Όπως μπορούμε να δούμε από τα διαγράμματα, το μοντέλο φαίνεται με το πέρασμα των εποχών να προσαρμόζεται όλο και καλύτερα στα δεδομένα κάθως τόσο η τιμή της συνάρτησης κόστους όσο και του RMSE μειώνονται σταδιακά. Όπως ήταν αναμενόμενο και στα δυο διαγράματα η καμπυλη που αντιστοιχεί στο validation set βρίσκεται πάνω από την αντίστοιχη για το training set. Επιπλέον, το γεγονός ότι το κόστος για το validation set δεν αρχίζει να ξαναυξάνεται από κάποια εποχή και μετά δείχνει ότι δεν έχουμε εκπαιδεύσει για αρκετές εποχές το μοντέλο μας με αποτέλεσμα να έχουμε underfitting στα δεδομένα εκπαίδευσης. Τέλος, αξίζει να παρατηρήσουμε τις σχετικά μεγάλες τιμές τόσο για το RMSE όσο και για το κόστος οι οποίες μας υποδεικνύουν ότι η εκπαίδευση του μοντέλου δεν ήταν αρκετά αποτελεσματική.



(a) (b)

Figure 1: Καμπύλες RMSE και μάθησης για το μοντέλο 1

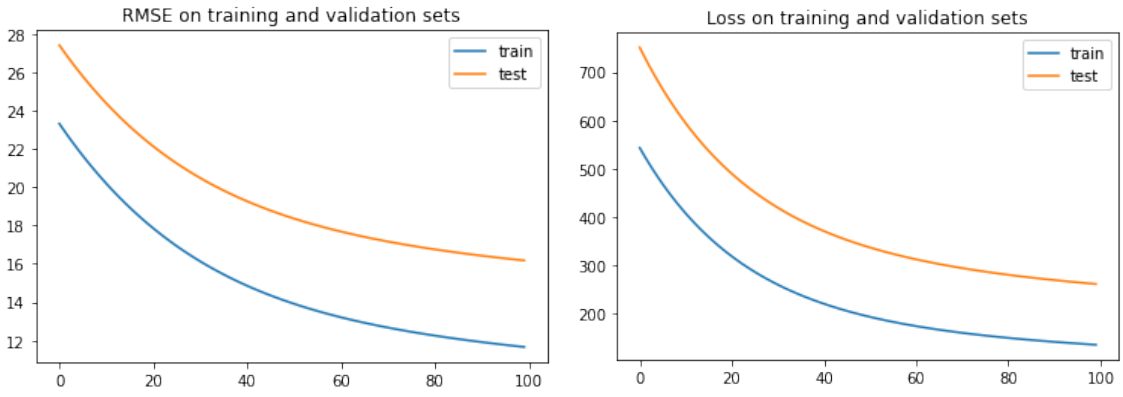
Οσον αφορά την αξιολόγηση του μοντέλου στο test set εχουμε ότι το RMSE = 11.99 το οποίο είναι καλύτερο από αυτό που είχαμε στο validation set αλλά προφανώς όχι τόσο καλό όσο αυτό που είχαμε στο training set. Βεβαια, για το R^2 έχουμε ότι $R^2 = -251$ το οποίο δεν είναι καθόλου ικανοποιητικό μιας και αρνητική τιμή του δείκτη αυτού δείχνει ότι οι προβλέψεις που πραγματοποιούμε με αυτό το μοντέλο ειναι "χειρότερες" από την πρόβλεψη με την μέση τιμή των δεδομένων εισόδου. Βέβαια, η τιμή αυτή μπορεί να είναι παραπλανητική καθώς το μοντέλο δίνει 128 εξόδους ενώ η πραγματική τιμή εξόδου είναι μόνο 1.

2.2 Μοντέλο 2

Το πλήθος των πυρήνων στο μοντέλο αυτό είναι ίσο με το 50% του πλήθους των δεδομένων εκπαίδευσης, δηλαδή 189 στο σύνολο. Επιπλέον, στο στρώμα εξόδου έχουμε 24320 βάρη προς εκπαίδευση. Τα νούμερα αυτά είναι σαφώς μεγαλύτερα σε σχέση με αυτά του μοντέλου 1 οπότε αναμένουμε και ο χρόνος εκπαίδευσης να είναι μεγαλύτερος. Στα Σχ.2α,β φαίνονται οι καμπύλες μάθησης και RMSE για τα training και validation sets.

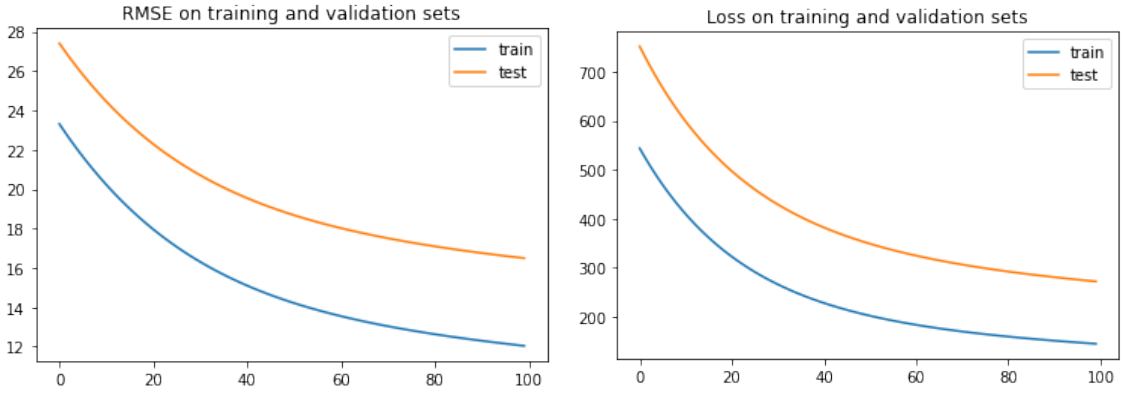
Όπως φαίνεται από τα διαγράμματα, αυτά έχουν αντίστοιχη μορφή με αυτά του μοντέλου 1. Έτσι, λοιπόν, μπορούμε και εδώ να προχωρήσουμε στα ίδια συμπεράσματα. Αυτό που αξίζει να σημειώσουμε εδώ είναι ότι καθόλη την διάρκεια της εκπαίδευσης, αν συγκρίνουμε τις τιμές του MSE και του RMSE για αντίστοιχες εποχές στα 2 μοντέλα θα δούμε ότι αυτές του μοντέλου 2 είναι μεγαλύτερες. Αυτό σημαίνει ότι το μοντέλο 2 δεν προσαρμόζεται τόσο καλά στα δεδομένα όσο το μοντέλο 1, κάτι το οποίο μας προκαλεί εντύπωση, λόγω του γεγονότος ότι ακριβώς επειδή έχει περισσότερους πυρήνες (πιο σύνθετο μοντέλο) θα περιμέναμε να έχει και μεγαλύτερη προσαρμοστική ικανότητα.

Αφού το μοντέλο 2 δεν προσαρμόζεται τόσο καλά στα δεδομένα εκπαίδευσης όσο το μοντέλο 1, είναι λογικό να υποθέσουμε ότι και η απόδοση του για πρόβλεψη στο test set δεν θα είναι τόσο καλή όσο του 1 (δεδομένου ότι το training set είναι αντιπροσωπευτικό ως σύνολο). Η υπόθεση αυτή φαίνεται να επιβεβαιώνεται αφού για το μοντέλο 2 έχουμε ότι RMSE = 12.54



(a) (b)

Figure 2: Καμπύλες RMSE και μάθησης για το μοντέλο 2



(a) (b)

Figure 3: Καμπύλες RMSE και μάθησης για το μοντέλο 3

και $R^2 = -277$ που είναι μεγαλύτερα από τα αντίστοιχα του μοντέλου 1 για το test set.

2.3 Μοντέλο 3

Το πλήθος των πυρήνων στο μοντέλο αυτό είναι ίσο με το 90% του πλήθους των δεδομένων εκπαίδευσης, δηλαδή 340 στο σύνολο. Επιπλέον, στο στρώμα εξόδου έχουμε 43776 βάρη προς εκπαίδευση. Έτσι, σε σχέση και με τα 2 προηγούμενα μοντέλα, αναμένουμε το μοντέλο αυτό να έχει την πιο χρονοβόρα εκπαίδευση μιας και είναι το πιο σύνθετο. Στα Σχ.3α,β φαίνονται οι καμπύλες μάθησης και RMSE για τα training και validation sets.

Όπως φαίνεται από τα διαγράμματα, αυτά έχουν αντίστοιχη μορφή με αυτά των μοντέλων 1 και 2. Έτσι, λοιπόν, μπορούμε και εδώ να προχωρήσουμε στα ίδια συμπεράσματα με πριν. Κάτι το οποίο δεν φαίνεται εύκολα με το μάτι από τα διαγράμματα αλλά μπορούμε να διαπιστώσουμε αν μελετήσουμε τις μετρικές αξιολόγησης και την τιμή του κόστους για αντίστοιχες εποχές

στα μοντέλα 2 και 3 είναι ότι αυτές του μοντέλου 3 είναι λίγο μεγαλύτερες. Αυτό ουσιαστικά σημαίνει ότι παρότι το μοντέλο 3 είναι πιο σύνθετο από το μοντέλο 2 φαίνεται να προσαρμόζεται στα δεδομένα λίγο χειρότερα από αυτό.

Αφού λοιπόν, το μοντέλο 3 δεν προσαρμόζεται τόσο καλά στα δεδομένα όσο το μοντέλο 2 και θεωρώντας πως το σύνολο εκπαίδευσης είναι αντιπροσωπευτικό αναμένουμε η απόδοση του μοντέλου 3 στο test set να είναι χειρότερη από αυτή του μοντέλου 2. Αυτό φαίνεται να επιβεβαιώνεται αφού έχουμε ότι RMSE $= 13.1$ και $R^2 = -302.57$ που είναι μεγαλύτερα από τα αντίστοιχα του μοντέλου 2.

2.4 Σύγκριση μοντέλων

Όπως έχει ήδη φανεί από την ανάλυση που προηγήθηκε, το μοντέλο 1 είναι το καλύτερο ακολουθόμενο από το μοντέλο 2, με το μοντέλο 3 να είναι το λιγότερο αποδοτικό από τα 3. Γενικά, ενώ αναμέναμε πως η αύξηση του πλήθους των πυρήνων θα οδηγούσε σε μεγαλύτερη προσαρμοστική και προβλεπτική ικανότητα, αυτό φαίνεται να μην ισχύει εν προκειμένω με αποτέλεσμα όσο λιγότεροι οι πυρήνες στο κρυφό στρώμα, τόσο καλύτερα τα αποτελέσματα. Πάντως, φαίνεται ότι και στα 3 μοντέλα έχουμε underfitting ενώ και η υλοποίηση με τους 128 νευρώνες στο στρώμα εξόδου φαίνεται να είναι προβληματική όπως υποδεικνύουν και οι αρνητικές τιμές του δείκτη R^2. Τέλος, στον Πιν.1 φαίνονται συγκεντρωτικά οι τιμές του RMSE και R^2 των 3 μοντέλων για το test set.

Μοντέλο	**RMSE**	$\mathbf{R^2}$
1	11.99	-251
2	12.54	-277
3	13.1	-302.57

Table 1: Μετρικές αξιολόγησης για πρόβλεψη στο test set

3 Fine-tuning RBF δικτύου

3.1 Το μοντέλο

Στην προηγούμενη ενότητα μελετήθηκαν και συγκρίθηκαν ως προς την αποδοτικότητά τους 3 διαφορετικά μοντέλα τα οποία διέφεραν ως προς το πληθος των πυρήνων στο κρυφό τους στρώμα. Στην παρούσα ενότητα, αυτό που μας ενδιαφέρει είναι πώς μπορούμε να ρυθμίσουμε εμείς τις διάφορες υπερπαραμέτρους ενός RBF μοντέλου προκειμένου να πετύχουμε όσο το δυνατόν καλύτερη απόδοση στο τελικό μοντέλο.

Πιο συγκεκριμένα, θέλουμε να δημιουργήσουμε ένα RBF δίκτυο του οποίου το κρυφό στρώμα θα αποτελειται από Γκαουσσιανές RBFs και θα το εκπαιδεύσουμε με τον αλγόριθμο KMeans,

ενώ το στρώμα εξόδου θα εκπαιδευτεί για 100 εποχές με την μέθοδο του Stochastic Gradient Descent (SGD) με ρυθμό μάθησης η=0.001. Επιπλέον, όσον αφορά το στρώμα εξόδου, τα βάρη αρχικοποιούνται βάσει του κανόνα του LeCun για αποφυγή του φαινομένου των vanishing ή exploding gradients ενώ επίσης εφαρμόζουμε και dropout. Τέλος, ως συνάρτηση κόστους θεωρούμε το MSE ενώ ως μετρική αξιολόγησης το RMSE. Από όλα τα παραπάνω, προκύπτει ότι οι υπερπαράμετροι προς βελτιστοποίηση είναι το πλήθος πυρήνων του κρυφού στρώματος, το πλήθος νευρώνων του στρώματος εξόδου και η πιθανότητα dropout.

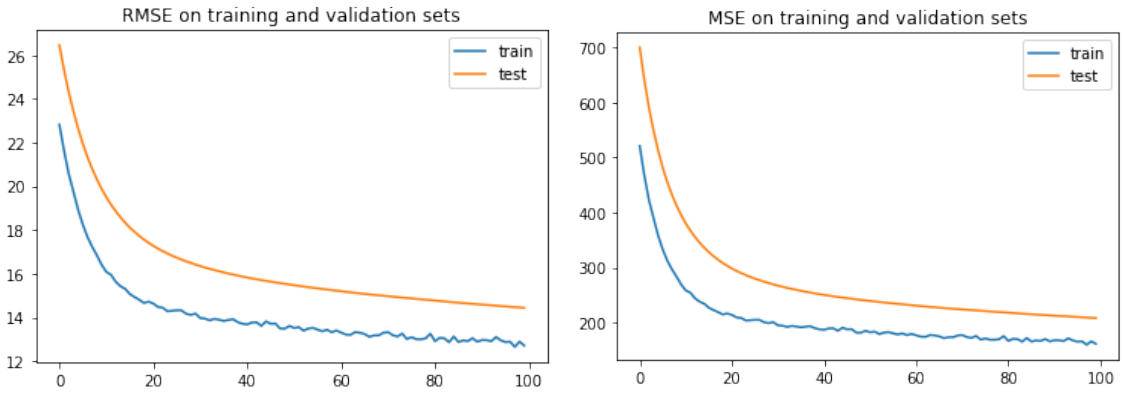
Για την επιλογή, του βέλτιστου μοντέλου δημιουργούμε ουσιαστικά ένα πλέγμα τιμών για τις υπερπαραμέτρους του μοντέλου και εκτελούμε εξαντλητικό grid search αξιολογώντας όλα τα πιθανά μοντέλα και επιλέγοντας τελικά το καλύτερο βάσει του RMSE. Επιπλέον εφαρμόζουμε 5-fold cross-validation κατά την αξιολόγηση κάθε μοντέλου προκειμένου τα συμπεράσματά μας να είναι πιο robust, θυσιάζοντας όμως ως προς τον χρόνο εκτέλεσης. Τέλος, κατά την διάρκεια εκπαίδευσης κάθε μοντέλου εφαρμόζουμε early stopping με patience 20 εποχών προκειμένου να επιταχύνουμε την διαδικασία και να αποφύγουμε την επιλογή μοντέλων που υπόκεινται σε overfitting. Από την διαδικασία αυτή τελικά προκύπτει ότι το βέλτιστο από τα 48 υποψήφια μοντέλα έχει 189 πυρήνες στο κρυφό στρώμα, 32 νευρώνες στο στρώμα εξόδου και πιθανότητα dropout p=0.2.

Βλέπουμε ότι, σε αντίθεση με αυτά που αναφέρθηκαν στην ενότητα 1, εδώ το βέλτιστο μοντέλο έχει τον μέγιστο αριθμό πυρήνων στο κρυφό στρώμα ενώ στο στρώμα εξόδου έχει τον ελάχιστο αριθμό πυρήνων. Αυτό μας κάνει να υποθέσουμε ότι η σκέψη μας ότι μεγαλύτερο πλήθος πυρήνων στο κρυφό στρώμα ισοδυναμεί με καλύτερη απόδοση να μην ήταν και τόσο λάθος. Το πρόβλημα κατά πάσα πιθανότητα έγγειται στο πλήθος των νευρώνων στο στρώμα εξόδου αφού εδώ βλέπουμε ότι ελλατώνοντάς το πετυχαίνουμε καλύτερη απόδοση. Τέλος, όσον αφορά την πιθανότητα dropout, βλέπουμε ότι η τιμή που επιλέχθηκε είναι η μικρότερη από τις υποψήφιες υποδεικνύοντας πως μάλλον δεν έχουμε overfitting του μοντέλου για να είναι απαραίτητη κάποια μέθοδος κανονικοποίησής του.

3.2 Απόδοση του μοντέλου

Αφού, επιλεχθεί το κατάλληλο μοντέλο, αυτό εκπαιδεύεται με τα δεδομένα του training set προκειμένου να αξιολογηθεί μετέπειτα. Όπως και κατά την εκπαίδευση των υποψήφιων δικτύων στο grid search, έτσι και τώρα χρησιμοποιούμε πάλι early stopping για να αποφύγουμε το overfitting και να επιταχύνουμε την διαδικασία. Αρχικά, θα μελετήσουμε την απόδοση του μοντέλου ως προς την ικανότητα προσαρμογής του στα δεδομένα εκπαίδευσης. Στα 4α,β μπορούμε να δούμε τις καμπύλες μάθησης και RMSE για τα training και validation sets.

Όπως μπορούμε να δούμε από τα διαγράμματα, το μοντέλο φαίνεται να προσαρμόζεται στα δεδομένα με το πέρασμα των εποχών αλλά όχι αρκετά καλά αφού τα MSE και RMSE παραμένουν υψηλά ακόμα και στις τελευταίες εποχές εκπαίδευσης. Αυτό σε συνδυασμό με το γεγονός ότι δεν είχαμε early stopping και ότι οι καμπύλες που αντιστοιχούν στο validation set συνεχώς φθίνουν κάνουν ξεκάθαρο ότι στην περίπτωση αυτή έχουμε underfitting του μοντέλου. Αξίζει επίσης να παράτηρήσουμε και την μορφή των καμπυλών για το training set που εμφανίζουν



(a) (b)

Figure 4: Καμπύλες RMSE και μάθησης του τελικού μοντέλου

κυματισμούς, κάτι το οποίο είναι χαρακτηριστικό όταν έχουμε dropout αφού ουσιαστικά προσθαφαιρούνται νευρώνες από το μοντέλο σε διαδοχικές εποχές εκπαίδευσής του επηρεάζοντας με μη ομαλό τρόπο την απόδοση του.

Τέλος, θα μελετήσουμε την ικανότητα γενίκευσης του μοντέλου αξιολογώντας την ικανότητα πρόβλεψής του στο test set. Προκύπτει ότι MSE $= 128.52$ και RMSE $= 11.33$ τιμές οι οποίες είναι αρκετά υψηλές και δείχνουν ότι το μοντέλο μας δεν είναι ιδιαίτερα αποδοτικό. Παρόλα αυτά, συγκρίνοντας την τιμή του RMSE που έχουμε εδώ με αυτήν που είχαν τα 3 μοντέλα που μελετήθηκαν στην Ενότητα 1 για το ίδιο test set βλέπουμε ότι όντως αυτό το μοντέλο είναι το βέλτιστο. Σε μελλοντικό επίπεδο, αυτό που θα μπορούσε να γίνει για την περαιτέρω βελτιστοποίηση του μοντέλου θα ήταν η επιλογή τιμών για τις υπερπαραμέτρους σε ένα ευρύτερο grid τιμών, πιθανώς με την χρήση κάποιου tuner για την επιτάχυνση της όλης διαδικασίας και η εκπαίδευση του μοντέλου για περισσότερες εποχές αφού δεν διατρέχουμε για τον υπάρχον αριθμό εποχών κίνδυνο overfitting του μοντέλου.