Κ23γ: Ανάπτυξη Λογισμικού για Αλγοριθμικά Προβλήματα
Χειμερινό εξάμηνο 2019-20

3η Προγραμματιστική Εργασία

Πρόγνωση της έντασης του ανέμου με προεκπαιδευμένο βαθύ νευρωνικό δίκτυο στη γλώσσα Python (3.7) με την χρήση του Keras API επί της πλατφόρμας μηχανικής μάθησης TensorFlow.

Η εργασία πρέπει να υλοποιηθεί σε σύστημα Linux και να υποβληθεί στις Εργασίες του e-class το αργότερο την Παρασκευή 17/01/2020 στις 23.59.

Περιγραφή της εργασίας

Θα πραγματοποιήσετε πειράματα με προεκπαιδευμένο βαθύ νευρωνικό δίκτυο για την πρόγνωση της έντασης του ανέμου. Τα πειράματα αυτά περιλαμβάνουν (α) συγκρίσεις των αποτελεσμάτων του νευρωνικού δικτύου με πραγματικά δεδομένα, καθώς και (β) εξαγωγή διανυσματικών αναπαραστάσεων της εισόδου του νευρωνικού από ενδιάμεσα στρώματα. Έπειτα, (γ) τα διανύσματα αυτά θα συσταδοποιηθούν στον χώρο $\Re^d$ βάσει της απόστασης Manhattan με σκοπό την αποτίμηση της φυσικής τους σημασίας.

Το νευρωνικό δίκτυο που παράγει προβλέψεις του ανέμου έχει εκπαιδευτεί χρησιμοποιώντας ως είσοδο αποτελέσματα από κλασικές μεθόδους αριθμητικής πρόγνωσης καιρού (Numerical Weather Prediction) και συγκεκριμένα από το μοντέλο Weather Research and Forecasting (WRF), με σκοπό τη βελτίωση των προγνώσεων αυτών. Η είσοδος του νευρωνικού αποτελείται από προγνώσεις για την ένταση του ανέμου, καθώς και άλλων σχετικών χαρακτηριστικών (θερμοκρασία, υγρασία, κλπ), σε 6ωρα χρονικά παράθυρα, με στόχο (έξοδο) την πραγματική τιμή έντασης ανέμου της τελευταίας (6ης) ώρας του παραθύρου. Η είσοδος και η έξοδος του νευρωνικού είναι κανονικοποιημένες στο [0, 1].

Καθώς το (βαθύ) νευρωνικό δίκτυο αποτελείται από πολλαπλά στρώματα (Convolutional, Recurrent, Fully Connected) με υψηλές υπολογιστικές απαιτήσεις, εμφανίζεται πλήρως μόνο υποσύνολο του δικτύου (N2) με το τελευταίο Fully Connected επίπεδο 64 κόμβων. Για τους σκοπούς της εργασίας, τα πρώτα στρώματα του δικτύου (N1) θεωρούνται ως Μαύρο Κουτί.

Τα δεδομένα που δίνονται για να χρησιμοποιηθούν ως είσοδο στο N2 είναι διανύσματα στον χώρο 128 διαστάσεων, τα οποία προκύπτουν από το τελευταίο επίπεδο 128 κόμβων του N1. Μαζί δίνονται και οι αληθινές τιμές του ανέμου για τη σύγκριση των αποτελεσμάτων του N2.

Α. Αφού τρέξετε το N2 νευρωνικό με είσοδο τα d-διάστατα διανύσματα (d=128) που δίνονται από το N1, θα συγκρίνετε την έξοδό του με τις αληθινές τιμές του ανέμου και θα υπολογίσετε το στατιστικό σφάλμα: μέσο απόλυτο σφάλμα (MAE), μέσο απόλυτο ποσοστό σφάλματος (MAPE).

Β. Χρησιμοποιώντας (μόνο) το πρώτο (ενδιάμεσο) επίπεδο του N2 (64 κόμβοι) θα δημιουργήσετε νέες d-διάστατες αναπαραστάσεις (d=64) της εισόδου του N2, εξάγοντας τα διανύσματα αυτά.

Γ. Οι αναπαραστάσεις από το Β θα χρησιμοποιηθούν για συσταδοποίηση (απόσταση Manhattan) με k=12 ή 4 ώστε να διαπιστωθεί αν διαχωρίζονται σε συστάδες ανά μήνα και εποχή του χρόνου αντίστοιχα. Η εν λόγω συσταδοποίηση θα συγκριθεί με τη συσταδοποίηση των διανυσμάτων 128 διαστάσεων που δίνεται στην είσοδο βάσει του δείκτη αξιολόγησης Silhouette. Η σύγκριση θα σχολιαστεί στην αναφορά που θα παραδοθεί.

ΕΙΣΟΔΟΣ

Αρχείο κειμένου `nn_representation.csv` διαχωρισμένο με κόμματα (comma-separated), το οποίο θα έχει την ακόλουθη γραμμογράφηση:

```
Timestamp1      x1|1        x1|2      ...    x1|128
     .            .           .       ...        .
TimestampN      xN|1        xN|2      ...    xN|128
```

όπου xij πραγματικός αριθμός εντός του [0,1] που αντιστοιχεί στην i συντεταγμένη της αναπαράστασης για το ωριαίο χρονικό παράθυρο που ξεκινά κατά το timestamp j.

Α. Το αρχείο `nn_representation.csv` δίνεται μέσω παραμέτρου στη γραμμή εντολών. Η εκτέλεση θα γίνεται μέσω της εντολής:

```
$python predict.py –i <input file>
```

Β. Η δημιουργία της νέας αναπαράστασης πραγματοποιείται με την εκτέλεση της εντολής:

```
$python new_representation.py –i <input file>
```

Γ. Το πρόγραμμα που αναπτύχθηκε στη 2η εργασία για τη συσταδοποίηση d-διάστατων διανυσμάτων θα χρησιμοποιηθεί για τη συσταδοποίηση των διανυσμάτων που θα προκύψουν από το ερώτημα Β, τα οποία θα έχουν αποθηκευτεί σε αρχείο CSV. Θα επιλεγεί η παραλλαγή του αλγορίθμου που επιτυγχάνει καλύτερη τιμή του δείκτη Silhouette. Η εκτέλεση θα γίνεται μέσω της εντολής:

```
$./cluster –i <new_representation.csv> –c cluster.conf –o
<output_file> -complete <optional>
```

ΕΞΟΔΟΣ

Α. Αρχείο κειμένου `predicted.csv` το όποιο θα έχει την ακόλουθη γραμμογράφηση:

```
MAE: <double>    MAPE: <double>  MSE: <double>
Timestamp1      c1|1        c1|2      ...    c1|7
     .            .           .        ...        .
TimestampN      cN|1        cN|2      ...    cN|7
```

όπου cij πραγματικός αριθμός εντός του [0,1] που αντιστοιχεί στο i κλιματικό χαρακτηριστικό (ένταση του ανέμου σε σημείου του χώρου, 1 <= i <=7) του οποίου γίνεται πρόγνωση για το ωριαίο χρονικό παράθυρο που ξεκινά κατά το timestamp j.

Η σύγκριση της πρόγνωσης με τις πραγματικές τιμές της έντασης του ανέμου που δίνονται στο αρχείο κειμένου actual csv χρησιμοποιείται για την εξαγωγή του μέσου απόλυτου σφάλματος (MAE), του μέσου απόλυτου ποσοστού σφάλματος (MAPE) και του μέσου τετραγωνικού σφάλματος (MSE). Τα αποτελέσματα θα γράφονται στην πρώτη γραμμή του αρχείου predicted.csv όπως φαίνεται παραπάνω.

Β. Ένα αρχείο κειμένου `new_representation.csv` το όποιο θα έχει την ακόλουθη γραμμογράφηση:

```
Timestamp1      n1|1        n1|2      ...    n1|64
     .            .           .        ...        .
TimestampN      nN|1        nN|2      ...    nN|64
```

όπου nij πραγματικός αριθμός εντός του [0,1] που αντιστοιχεί στην i συντεταγμένη της νέας αναπαράστασης για το ωριαίο χρονικό παράθυρο που ξεκινά κατά το timestamp j.

Γ. Ένα αρχείο κειμένου το οποίο περιλαμβάνει τις συστάδες των διανυσμάτων της νέας αναπαράστασης που παρήχθησαν από τον αλγόριθμο, τον χρόνο εκτέλεσης σε κάθε περίπτωση καθώς και τον δείκτη εσωτερικής αξιολόγησης της συσταδοποίησης **Silhouette** (k = 4 ή k = 12).

```
CLUSTER-1 {size: <int>, centroid: <timestamp_id> ή πίνακας με τις
συντεταγμένες του centroid στην περίπτωση k-means Update}
.     .     .     .     .     .     .
CLUSTER-K {size: <int>, centroid: <timestamp_id> ή πίνακας με τις
συντεταγμένες του centroid στην περίπτωση K-means Update }
clustering_time: <double> //in seconds
Silhouette: [s1,...,si,...,sK, stotal]

/* si=average s(p) of points in cluster i, stotal=average s(p) of points in
dataset */
```

```
/* Optionally with command line parameter –complete */
CLUSTER-1 {timestamp_idA, timestamp_idB, ..., timestamp_idC}
.      .      .      .      .      .      .

CLUSTER-K {timestamp_idR, timestamp_idT, ..., timestamp_idZ}
```

Επιπρόσθετες απαιτήσεις

1. Αρχείο (ή ενότητα στο Readme) που να σχολιάζει τα αποτελέσματα.
2. Το παραδοτέο πρέπει να είναι επαρκώς τεκμηριωμένο με πλήρη σχολιασμό του κώδικα και την ύπαρξη αρχείου Readme το οποίο περιλαμβάνει κατ' ελάχιστο: α) τίτλο και περιγραφή του προγράμματος, β) κατάλογο των αρχείων κώδικα και περιγραφή τους, γ) οδηγίες χρήσης του προγράμματος και δ) πλήρη στοιχεία των φοιτητών που το ανέπτυξαν.
3. Η υλοποίηση του προγράμματος θα πρέπει να γίνει με τη χρήση συστήματος διαχείρισης εκδόσεων λογισμικού και συνεργασίας (Git ή SVN) [ομάδες 2 ατόμων].