Εργασία 4
Διαγωνισμός Κατηγοριοποίησης

Εξόρυξη Γνώσης από Βάσεις Δεδομένων και τον Παγκόσμιο Ιστό

Γιάννης Νικολέντζος - nikolentzos@aueb.gr
Μιχάλης Βαζιργιάννης - mvazirg@aueb.gr

Δεκέμβριος, 2014

1 Περιγραφή

Το πρόβλημα που θα αντιμετωπίσετε στα πλαίσια της παρούσας εργασίας προέρχεται από τον ιστοχώρο Kaggle[1]. Το Kaggle είναι μια πλατφόρμα για την ανάλυση δεδομένων και την δημιουργία μοντέλων πρόβλεψης. Συγκεκριμένα, εταιρείες και ερευνητές δημοσιεύουν τα δεδομένα τους σε μορφή διαγωνισμών και άνθρωποι από όλο τον κόσμο οι οποίοι δουλεύουν στην επιστήμη των δεδομένων ανταγωνίζονται για να παράγουν τα καλύτερα μοντέλα. Με τον τρόπο αυτό, οι εταιρείες και οι ερευνητές μπορούν να διερευνήσουν ποιά μέθοδος είναι πιο αποτελεσματική για την αντιμετώπιση των προβλημάτων τους. Για να παρακινήσουν μεγαλύτερο αριθμό ανθρώπων να συμμετέχουν σε ένα διαγωνισμό, συχνά, οι εταιρείες δίνουν ένα χρηματικό έπαθλο στην ομάδα ή στις ομάδες που θα παρουσιάσουν τα καλύτερα μοντέλα.

Σκοπός της εργασίας είναι να προβλέψτε την επιβίωση των επιβατών του Τιτανικού. Το ναυάγιο του Τιτανικού είναι ίσως το πιο γνωστό ναυάγιο στην ιστορία. Στις 15 Απριλίου 1912, κατά τη διάρκεια του παρθενικού του ταξιδιού, ο Τιτανικός βυθίστηκε μετά από πρόσκρουση με παγόβουνο, σκοτώνοντας 1502 από τους 2224 επιβάτες και πλήρωμα. Το ναυάγιο συγκλόνισε τη διεθνή κοινότητα και οδήγησε σε καλύτερους κανονισμούς ασφαλείας για τα πλοία. Ένας από τους λόγους που διασώθηκε μόνο ένα τόσο μικρό ποσοστό ανθρώπων ήταν ότι δεν υπήρχαν αρκετές σωσίβιοι λέμβοι για όλους. Αν και έπαιζε ρόλο και η τύχη στο αν θα επιζούσε κάποιος από το ναυάγιο, κάποιες ομάδες ανθρώπων όπως για παράδειγμα οι γυναίκες, τα παιδιά και η ανώτερη τάξη είχαν περισσότερες πιθανότητες να επιβιώσουν σε σχέση με τους άλλους. Στην εργασία αυτή, σας ζητείται να εφαρμόσετε μεθόδους κατηγοριοποίησης για να προβλέψετε ποιοί επιβάτες επέζησαν από την τραγωδία.

2 Σύνολο Δεδομένων

Ως σύνολο δεδομένων σας δίνεται το αρχείο `train.csv` το οποίο αποτελεί το σύνολο εκπαίδευσης του διαγωνισμού. Κάθε γραμμή του αρχείου αντιστοιχεί σε έναν επιβάτη και κάθε επιβάτης χαρακτηρίζεται από τις εξής μεταβλητές:

1. **PassengerId:** Αναγνωριστικό επιβάτη
2. **Survived:** Εάν επέζησε ή όχι (0 = Όχι; 1 = Ναι)
3. **Pclass:** Οικονομική θέση επιβάτη (1 = Πρώτη; 2 = Δεύτερη; 3 = Τρίτη)

[1] http://www.kaggle.com/

4. **Name:** Όνομα
5. **Sex:** Φύλο
6. **Age:** Ηλικία
7. **SibSp:** Αριθμός από αδέλφια/συζύγους στο πλοίο
8. **Parch:** Αριθμός γονέων/παιδιών στο πλοίο
9. **Ticket:** Αριθμός Εισητηρίου
10. **Fare:** Τιμή Εισητηρίου
11. **Cabin:** Καμπίνα
12. **Embarked:** Λιμένας Επιβίβασης (C = Cherbourg; Q = Queenstown; S = Southampton)

Για παράδειγμα, οι γραμμές που αντιστοιχούν στους επιβάτες με αναγνωριστικά 1 και 2 έχουν τη μορφή:

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C

Σημειώστε ότι στο σύνολο δεδομένων υπάρχουν τιμές οι οποίες για διάφορους λόγους λείπουν και θα πρέπει με κάποιον τρόπο να τις χειριστείτε. Μπορείτε για παράδειγμα να βάλετε στη θέση τους μια τιμή που υποδεικνύει ότι η τιμή αυτή λείπει. Μπορείτε επίσης να τις αντικαταστήσετε με τις μέσες τιμές των συγκεκριμένων χαρακτηριστικών. Για κατηγορικές μεταβλητές μπορείτε να τις αντικαταστήσετε με τις πιο κοινές τιμές. Μπορείτε ακόμη να χρησιμοπιήσετε ένα μοντέλο Γραμμικής Παλινδρόμησης για να προβλέψετε αυτές τις τιμές.

Επιπλέον, κάποια χαρακτηριστικά δεν προσφέρουν καμία πληροφορία για την κατηγοριοποίηση των επιβατών. Για παράδειγμα, το αναγνωριστικό επιβάτη απλώς αριθμεί τους επιβάτες από 1 μέχρι 891 και μπορούμε συνεπώς να το αγνοήσουμε. Με αυτόν τον τρόπο μπορούμε να διατηρήσουμε μόνο τα χαρακτηριστικά εκείνα που είναι σημαντικά για την ταξινόμηση. Δεδομένου ότι τα δεδομένα μας βρίσκονται σε έναν πίνακα X, μπορούμε να διαγράψουμε μια στήλη του πίνακα χρησιμοποιώντας την παρακάτω εντολή (τη στήλη 0 σε αυτή την περίπτωση):

```
X = delete(X,0,1)
```

Κάποια κατηγορικά χαρακτηριστικά είναι σε μορφή αλφαριθμητικών. Για να εφαρμόσετε τις περισσοτερες μεθόδους κατηγοριοποίησης θα πρέπει να αντικαταστήσετε τα χαρακτηριστικά αυτά με αριθμητικές τιμές. Για παράδειγμα, η τιμή άντρας του χαρακτηριστικού φύλο αναπαριστάται στα δεδομένα ως 'male', ενώ η τιμή γυναίκα του συγκεκριμένου χαρακτηριστικού αναπαριστάται ως 'female'. Μπορείτε να αλλάξετε τα παραπάνω αλφαριθμητικά με αριθμητικές τιμές (για παράδειγμα 1 για άντρα και -1 για γυναίκα) ώστε να μπορούν να τα διαχειριστούν οι μέθοδοί σας. Εάν τα δεδομένα στον πίνακα X είναι σε μορφή αλφαριθμητικών και εφόσον έχουμε μετατρέψει όλους τους χαρακτήρες των αλφαρθμητικών σε αριθμούς, μπορούμε να μετατρέψουμε τον X σε μορφή int ή float με τις παρακάτω εντολές:

```
X = X.astype(int)
```

```
X = X.astype(float)
```

Μπορείτε να εφαρμόσετε κάποια μέθοδο επιλογής χαρακτηριστικών στα δεδομένα ώστε να κρατήσετε μόνο ένα υποσύνολο από τα χαρακτηριστικά. Μπορείτε επίσης να δημιουργήσετε νέα χαρακτηριστικά από ήδη υπάρχοντα τα οποία δεν μπορούν να βοηθήσουν στην κατηγοριοποίηση. Για παράδειγμα, το όνομα του επιβάτη δεν δίνει κάποια σημαντική πληροφορία για την κατηγοριοποίηση. Ωστόσο, το γεγονός ότι κάθε όνομα περιλαμβάνει τον τίτλο του συγκεκριμένου ατόμου (Mr., Mrs., Miss., Master., Mlle., Dr., ...) μας δίνει τη δυνατότητα να δημιοργήσουμε ένα νέο χαρακτηριστικό το οποίο μπορεί να συνεισφέρει σημαντικά στην κατηγοριοποίηση. Αν έχετε ένα διάνυσμα u το οποιο αντιπροσωπεύει κάποιο νέο χαρακτηριστικό που έχετε δημιουργήσει και θέλετε να το προσθέσετε σαν στήλη στον πίνακα δεδομένων X μπορείτε να χρησιμοποιήσετε την παρακάτω εντολή:

```
X = column_stack((X,u))
```

Μπορείτε επιπλέον να πειραματιστείτε με κάποια μέθοδο μείωσης διάστασης και να διερευνήσετε αν η εφαρμογή της βελτιώνει το αποτέλεσμα της κατηγοριοποίησης.

Τέλος, μπορείτε να κανονικοποιήσετε τις τιμές των χαρακτηριστικών στον πίνακα X αφού κάποιοι κατηγοριοποιητές δουλεύουν καλύτερα με κανονικοποιημένα χαρακτηριστικά. Επίσης, μπορείτε να χρησιμοποιήσετε θορυβώδη ή ανούσια χαρακτηριστικά για να παράγετε νέα χαρακτηριστικά που παρέχουν μεγαλύτερα ποσοστά πληροφορίας.

3 Αξιολόγηση

Για την αξιολόγηση των μοντέλων που θα υλοποιήσετε θα χρησιμοποιηθεί διασταυρωμένη επικύρωση (cross validation). Η διασταυρωμένη επικύρωση είναι μια μέθοδος για την εκτίμηση της απόδοσης ενός μοντέλου σε ένα ανεξάρτητο σύνολο δεδομένων. Χρησιμοποιείται κυρίως σε προβλήματα κατηγοριοποίησης όπου κάποιος θέλει να ξέρει πώς ένα μοντέλο θα συμπεριφερθεί στην πράξη. Υποθέστε ότι έχουμε ένα μοντέλο με μια ή περισσότερες παραμέτρους και ένα σύνολο δεδομένων το οποίο μπορούμε να χρησιμοποιήσουμε για να εκπαιδεύσουμε το μοντέλο. Η διαδικασία εκπαίδευσης βελτιστοποιεί τις παραμέτρους του μοντέλου ώστε το μοντέλο να περιγράφει τα δεδομένα εκπαίδευσης όσο το δυνατόν ακριβέστερα. Αν στη συνέχεια χρησιμοποιήσουμε το μοντέλο σε ένα ανεξάρτητο σύνολο αξιολόγησης, θα παρατηρήσουμε ότι το μοντέλο δεν περιγράφει το σύνολο αξιολόγησης τόσο καλά όσο το σύνολο εκπαίδευσης. Το παραπάνω ονομάζεται υπερπροσαρμογή (overfitting) και είναι πολύ πιθανό να συμβεί όταν το σύνολο εκπαίδευσης είναι μικρό ή οταν το μοντέλο αποτελείται από πολλές παραμέτρους. Η διασταυρωμένη επικύρωση παρέχει έναν τρόπο για να εκτιμήσουμε την απόδοση που θα έχει ένα μοντέλο σε ένα υποθετικό σύνολο αξιολόγησης όταν αυτό δεν είναι διαθέσιμο.

Για να αξιολογήσουμε τους ταξινομητές που θα υλοποιήσετε θα χρησιμοποιήσουμε διασταυρωμένη επικύρωση k δειγμάτων (k-fold cross validation) με τιμή $k = 10$. Σύμφωνα με την παραπάνω μέθοδο, το αρχικό σύνολο δεδομένων διαιρείται τυχαία σε k υποσύνολα δεδομένων ίσου μεγέθους. Ένα από τα k υποσύνολα θεωρείται ως το σύνολο αξιολόγησης, ενώ τα υπόλοιπα $k-1$ υποσύνολα συνθέτουν το σύνολο εκπαίδευσης. Η διαδικασία επεναλαμβάνεται k φορές με καθένα από τα k υποσύνολα να χρησιμοποιείται ακριβώς μια φορά ως σύνολο αξιολόγησης και τις υπόλοιπες $k-1$ ως τμήμα του συνόλου εκπαίδευσης. Τα k αποτελέσματα μπορούν στη συνέχεια να συνδυαστούν για να παράγουν μια κοινή αξιολόγηση του μοντέλου.

Η αξιολόγηση θα γίνει με βάση την ακρίβεια δηλαδή τα στιγμιότυπα των συνόλων αξιολόγησης που κατηγοριοποιήθηκαν σωστά προς το συνολικό αριθμό τους. Ένα ιδανικό (τέλειο) μοντέλο θα επέστρεφε ακρίβεια ίση με 1. Στην πράξη, είναι πολύ δύσκολο ένα μοντέλο να προσεγγίσει το ιδανικό, οπότε τα μοντέλα που σχεδιάζουμε επιτυγχάνουν χαμηλότερες ακρίβειες.

Εάν επιθυμείτε να δείτε τον τρόπο με τον οποίο λειτουργεί η πλατφόρμα Kaggle, μπορείτε να δημιουργήσετε λογαριασμό στην πλατφόρμα και να κατεβάσετε το σύνολο αξιολόγησης το οποίο διατίθεται στον

ιστοχώρο του διαγωνισμού[2]. Στη συνέχεια, μπορείτε να χρησιμοποιήσετε το μοντέλο σας για να προβλέψετε ποιά άτομα του συνόλου αυτού επέζησαν, να υποβάλλετε το αρχείο με τα αποτελέσματα στην πλατφόρμα και να δείτε την ακρίβεια που επιτύχατε και την θέση που έχετε καταλάβει σε σχέση με τους άλλους διαγωνιζόμενους.

4 Αρχικός Κώδικας

Κάνοντας μια απλή διερευνητική ανάλυση (exploratory analysis) στα δεδομένα, μπορείτε να παρατηρήσετε ότι το μεγαλύτερο ποσοστό των γυναικών επέζησε, ενώ το μέγαλύτερο ποσοστό των αντρών δεν κατάφερε να επιζήσει. Σας δίνεται ένας αρχικός Python κώδικας ο οποίος προβλέπει ότι όλες οι γυναίκες επέζησαν ενώ όλοι οι άντρες δεν επέζησαν. Με εφαρμογή διασταυρωμένης επικύρωσης 10 δειγμάτων, το παραπάνω μοντέλο δίνει ακρίβεια ίση με 0.787. Στα πλαίσια της παρούσας εργασίας καλείστε να τροποποιήσετε τον κώδικα της `classify.py` και να χρησιμοποιήσετε κάποια μέθοδο κατηγοριοποίησης για να προβλέψετε ποιοι επιβάτες επέζησαν.

4.1 Περιγραφή Κώδικα `main.py`

Στο σημείο αυτό θα δώσουμε μια περιγραφή του κώδικα της `main.py`.

```
# Load data
csv_file_object = csv.reader(open('train.csv', 'rb')) # Load in the csv file
header = csv_file_object.next()                       # Skip the fist line as it is a header
data=[]                                               # Create a variable to hold the data

for row in csv_file_object: # Skip through each row in the csv file,
    data.append(row[0:])    # adding each row to the data variable
X = array(data)             # Then convert from a list to an array.
```

Οι παραπάνω εντολές φορτώνουν τα δεδομένα από το αρχείο *train.csv* στη μεταβλητή X. Προσέξτε ότι τα δεδομένα του πίνακα X είναι σε μορφή *string*. Μπορείτε να μετατρέψετε τα δεδομένα του πίνακα σε μορφή *int* ή *float* με τις εντολές $X = X.astype(int)$ και $X = X.astype(float)$ αντίστοιχα. Βέβαια, για να μπορεί να πραγματοποιηθεί αυτό, θα πρέπει όλα τα στοιχεία του πίνακα να μπορούν να μετατραπούν σε αυτούς τους τύπους δηλαδή να περιέχουν μόνο αριθμητικούς χαρακτήρες.

```
y = X[:,1].astype(int) # Save labels to y

X = delete(X,1,1) # Remove survival column from matrix X
```

Αποθηκεύουμε την στήλη *Survived* του πίνακα X στη μεταβλητή y σε μορφή *int*. Μπορούμε να το κάνουμε αυτό αφού η στήλη αυτή περιλαμβάνει αλφαριθμητικά με τιμές 0 ή 1. Στη συνέχεια, διαγράφουμε από τον X την παραπάνω στήλη.

```
# Initialize cross validation
kf = cross_validation.KFold(X.shape[0], n_folds=10)
```

Χρησιμοποιούμε τη συνάρτηση $KFold()$ η οποία εφαρμόζει διασταυρωμένη επικύρωση 10 δειγμάτων και επιστρέφει για κάθε μια από τις 10 περιπτώσεις ποιές γραμμές του συνόλου δεδομένων ανήκουν στο σύνολο εκπαίδευσης και ποιές στο σύνολο αξιολόγησης.

```
for trainIndex, testIndex in kf:
    trainSet = X[trainIndex]
    testSet = X[testIndex]
```

[2]http://www.kaggle.com/c/titanic-gettingStarted/data

```
    trainLabels = y[trainIndex]
    testLabels = y[testIndex]

    predictedLabels = classify(trainSet, trainLabels, testSet)

    correct = 0
    for i in range(testSet.shape[0]):
        if predictedLabels[i] == testLabels[i]:
            correct += 1

    print 'Accuracy: ' + str(float(correct)/(testLabels.size))
    totalCorrect += correct
    totalInstances += testLabels.size
print 'Total Accuracy: ' + str(totalCorrect/float(totalInstances))
```

Για καθεμία από τις 10 περιπτώσεις της διασταυρωμένης επικύρωσης, αποθηκεύουμε στις μεταβλητές *trainSet* και *trainLabels* τα δεδομένα του συνόλου εκπαίδευσης και τις κατηγορίες στις οποίες ανήκουν, και στις μεταβλητές *testSet* και *testLabels* τα δεδομένα του συνόλου αξιολόγησης και τις κατηγορίες στις οποίες αυτά ανήκουν. Χρησιμοποιούμε τη συνάρτηση $classify()$ για να προβλέψουμε τις κατηγορίες των στιγμιοτύπων του συνόλου αξιολόγησης και τις συγκρίνουμε με τις πραγματικές κατηγορίες που βρίσκονται στη μεταβλητή *testLables* για να υπολογίσουμε την ακρίβεια καθεμιάς από τις 10 περιπτώσεις καθώς και τη συνολική ακρίβεια.

4.2 Περιγραφή Κώδικα `classify.py`

Στο σημείο αυτό θα δώσουμε μια περιγραφή του κώδικα της `classify.py`.

```
def classify(trainSet, trainLabels, testSet):

        predictedLabels = zeros(testSet.shape[0])

        for i in range(testSet.shape[0]):
                if testSet[i,3] == 'female':
                        predictedLabels[i] = 1

        return predictedLabels
```

Αρχικά, αρχικοποιούμε τη μεταβλητή *predictedLabels* η οποία θα αποθηκεύσει την κατηγορία στην οποία ανήκει κάθε στιγμιότυπο του συνόλου αξιολόγησης και στη συνέχεια, για κάθε στιγμιότυπο του συνόλου αξιολόγησης εξετάζουμε αν είναι άντρας ή γυναίκα. Στην περίπτωση που πρόκειται για γυναίκα του αναθέτουμε την κατηγορία 1 (επέζησε). Διαφορετικά αφήνουμε τη τιμή που υπάρχει ήδη στην μεταβλητή *predictedLabels* (0 = δεν επέζησε).

5 Παράδοση Εργασίας

Το παραδοτέο της εργασίας θα είναι ένα συμπιεσμένο αρχείο το οποίο θα παραδοθεί μέσω e-class και θα περιλαμβάνει:

- Ένα αρχείο `report.pdf` το οποίο θα περιέχει μια σύντομη αναφορά. Μέσα στην αναφορά δεν αρκεί να αναφέρετε τις μεθόδους που χρησιμοποιήσατε, αλλά θα πρέπει να τις περιγράψετε και να εξηγήσετε εάν υπήρχε κάποιος λόγος που σας οδήγησε σε αυτή την επιλογή. Θα πρέπει να συμπεριλάβετε και τα βήματα προεπεξεργασίας που ακολουθήσατε. Η προσέγγισή σας θα πρέπει να φαίνεται ξεκάθαρα από τα αρχεία και την αναφορά που θα παραδώσετε ώστε να παρουσιάσετε καταλλήλως τον κύκλο

εργασιών σας. Αυτό σημαίνει ότι πρέπει να συμπεριλάβετε πέρα από τον αλγόριθμο που σας έδωσε τα καλύτερα αποτελέσματα και οτιδήποτε άλλο δοκιμάσατε και θεωρείτε πως αξίζει να το αναφέρετε.

- Το αρχείο `classify.py` καθώς και τυχόν άλλα αρχεία που αυτό χρησιμοποιεί για να προβλέψει την κατηγορία στην οποία ανήκουν τα στιγμιότυπα του συνόλου αξιολόγησης.