Πανεπιστήμιο Πατρών
Τμήμα Οικονομικών Επιστημών
Πρόγραμμα Μεταπτυχιακών Σπουδών «Εφαρμοσμένη Οικονομική και Ανάλυση Δεδομένων»

Ακαδημαϊκό έτος 2022- 2023

2η Εργασία μαθήματος «Διαχείριση Μεγάλων Δεδομένων»

Γιατί η λογοκρισία είναι κακό πράγμα:
Original: https://www.youtube.com/watch?v=3e7yYBDHOgg
Λογοκριμένο: https://www.youtube.com/watch?v=B-Wd-Q3F8KM

Εισαγωγή

Σκοπός της εργασίας είναι η εξοικείωση με τις μεθόδους εκτίμησης συντελεστών γραμμικών μοντέλων παλινδρόμησης τόσο με τη μέθοδο Ελαχίστων Τετραγώνων (OLS) όσο και με τη μέθοδο της Σταδιακής Καθόδου (Gradient Descent) και των διαφόρων εκδοχών αυτής στα περιβάλλοντα R και Python.

Υπενθυμίζεται, προκειμένου να εξοικειωθείτε με τη χρήση της R και της Python για την ανάγνωση αρχείων, επεξεργασία δεδομένων με τις ειδικές βιβλιοθήκες που αναφέρονται στην εκφώνηση καθώς επίσης για την δημιουργία συναρτήσεων (functions) στα περιβάλλοντα αυτά, σας συστήνεται να μελετήσετε **οπωσδήποτε** τα αρχεία που αναφέρονται στον παρακάτω πίνακα, που θα τα βρείτε στο ιστότοπο του μαθήματος στο eclass και τα οποία δίνουν απλά και συνοπτικά σχετικά παραδείγματα επεξεργασίας των δεδομένων με τον ζητούμενο τρόπο:

	R	**Python**
Ανάγνωση αρχείων csv	R-ReadingCSVFiles.R.rar και R-example.R	workingWithPandas.rar
Δημιουργία και χειρισμός διανυσμάτων	R-IntroVarVecListsFunctions.R.txt	workingWithNumpyArrays.py
Δημιουργία και χρήση συναρτήσεων	R-IntroVarVecListsFunctions.R.txt	workingWithNumpyArrays.py
Χειρισμός δεδομένων με χρήση data frame	subsetting-data.R	workingWithPandas.rar και python-examples.rar (αρχείο subsetting-data.py)
Γραφικές παραστάσεις		plottingWithMatplotlib.rar
Μέθοδο OLS	R-LinearRegression.rar	Python-LinearRegression.rar
Μέθοδο Στοχαστικής Καθόδου Δέσμης		R-GradientDescent.rar
Διασταυρωτική επικύρωση κ-πτυχών	Python-k-foldCrossValidation.rar	R-k-FoldCrossValidation.rar

Για την υποστήριξη της θεωρητικής σας μελέτης των ζητουμένων της εργασίας, πέραν των σημειώσεων και αρχείων .R και .py που υπάρχουν στην σελίδα του μαθήματος στο eclass, μπορείτε να παρακολουθήσετε (και συστήνετε!) και τη διάλεξη του Andrew Ng από το Πανεπιστήμιο του Stanford στο youtube: https://www.youtube.com/watch?v=4b4MUYve_U8 . Η διάλεξη καλύπτει την ίδια ύλη με τις σημειώσεις που υπάρχουν στο eclass.

Θέμα 1

Απαντήστε, ως ομάδα, στις online ερωτήσεις της άσκησης που έχει δημοσιευτεί στην ιστοσελίδα του μαθήματος στο eclass. Η άσκηση έχει τίτλο *«Ερωτήσεις θέματος 1) της 2ης εργασίας ακ. έτους 2022-2023 (Επεξεργασία δεδομένων με χρήση της βιβλιοθήκης pandas της Python)»* και μπορεί να βρεθεί στην ενότητα Ασκήσεις στον ιστότοπο του μαθήματος στο eclass: https://eclass.upatras.gr/modules/exercise/?course=ECON1332

Θέμα 2

Κατεβάστε από τον ιστότοπο "UCI Machine Learning Repository" και ειδικότερα από την ιστοσελίδα https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime το σύνολο δεδομένων που περιέχει παρατηρήσεις σχετικά με την εγκληματικότητα ανά 100000 κατοίκους σε περιοχές των ΗΠΑ (μεταβλητή ViolentCrimesPerPop) μαζί με κοινωνικοοικονομικά στοιχεία για την κάθε περιοχή (Communities and Crime Data Set). Το σύνολο δεδομένων βρίσκεται στο αρχείο communities.data, που θα το βρείτε εάν στην παραπάνω σελίδα ακολουθήσετε τον σύνδεσμο "Data Folder". Η ιστοσελίδα παρέχει και πληροφορίες για τις μεταβλητές του αρχείου δεδομένων. Τις ίδιες πληροφορίες μπορείτε να τις βρείτε και στο αρχείο "Data Set Description" από την παραπάνω ιστοσελίδα για να δείτε σε ποια σειρά και τί συλλαμβάνει κάθε μεταβλητή του συνόλου δεδομένων.
Αφού εξοικειωθείτε με το σύνολο δεδομένων, τα γνωρίσματα και τη σημασία τους, εκτιμήστε τους συντελεστές του παρακάτω μοντέλου παλινδρόμησης

$$\begin{aligned} ViolentCrimesPerPop \\ = \beta_1 medIncome + \beta_2 whitePerCap + \beta_3 blackPerCap \\ + \beta_4 HispPerCap + \beta_5 NumUnderPov + \beta_6 PctUnemployed \\ + \beta_7 HousVacant + \beta_8 MedRent + \beta_9 NumStreet + \beta_0 \end{aligned}$$

με τους τρόπους που ζητούνται παρακάτω:

i. Συγγράψτε πρόγραμμα στην **R και στην Python** το οποίο εκτιμά τους συντελεστές του παραπάνω γραμμικού μοντέλου παλινδρόμησης με τη μέθοδο των ελαχίστων τετραγώνων (OLS) χρησιμοποιώντας το σύνολο δεδομένων Communities and Crime Data Set. Τα προγράμματά σας θα πρέπει να εμφανίζουν στην οθόνη τις τιμές των συντελεστών που έχουν εκτιμηθεί. Τα προγράμματά σας θα πρέπει επίσης να αφαιρούν όσες παρατηρήσεις έχουν τουλάχιστον μία τιμή (σε οποιαδήποτε μεταβλητή) που λείπει (missing value) κατά τη διαδικασία προεπεξεργασίας. Ακολουθήστε τέτοια προεπεξεργασία των δεδομένων για όλα τα θέματα της εργασίας αυτής.
ii. Συγγράψτε πρόγραμμα **μόνο σε Python** που εκτιμά τους συντελεστές του παραπάνω γραμμικού μοντέλου παλινδρόμησης με τη μέθοδο της Σταδιακής Καθόδου Δέσμης

(Batch Gradient Descent) και χρησιμοποιώντας το σύνολο δεδομένων Communities and Crime Data Set.
Προς τούτο, στο πρόγραμμά σας Python, δημιουργείστε μία συνάρτηση με όνομα batchGradientDescent που δέχεται τις ακόλουθες παραμέτρους και η οποία θα υπολογίζει τους συντελεστές ενός πολλαπλού μοντέλου γραμμικής παλινδρόμησης με μέθοδο της Σταδιακής Καθόδου Δέσμης:

def batchGradientDescent(independentVars, dependentVar, thetas, alpha=0.01, numIters=100):

όπου *independentVars* η μήτρα των τιμών των ανεξάρτητων μεταβλητών, *dependentVar* η μήτρα των τιμών της εξαρτημένης μεταβλητής, *thetas* ένα διάνυσμα με τις αρχικές τιμές των συντελεστών θ, *alpha* η τιμή της παραμέτρου μάθησης α και *numIters* το πλήθος των επαναλήψεων που θα πρέπει να κάνει η μέθοδος. Η υλοποίηση της μεθόδου της Σταδιακής Καθόδου Δέσμης (Batch Gradient Descent) θα πρέπει να γίνει από την αρχή ("from scratch") χρησιμοποιώντας τον τύπο ενημέρωσης των συντελεστών για την εκτίμηση των συντελεστών γραμμικών μοντέλων παλινδρόμησης. Η συνάρτηση δεν θα πρέπει να κάνει χρήση υπάρχουσας βιβλιοθήκης Python, που να παρέχει έτοιμη τη μέθοδο της Σταδιακής Καθόδου Δέσμης. Το κριτήριο τερματισμού της μεθόδου Σταδιακής Καθόδου Δέσμης, είναι το πλήθος των επαναλήψεων.
Η διαδικασία της εκτίμησης των συντελεστών θα πρέπει να τερματίζει αφού έχει εκτελεστεί το πλήθος των επαναλήψεων που προσδιορίζεται από το όρισμα numIters. Η συνάρτηση batchGradientDescent που θα δημιουργήσετε θα πρέπει να επιστρέφει τόσο ένα διάνυσμα με τους συντελεστές που εκτιμήθηκαν όσο και τις τιμές της συνάρτησης κόστους σε κάθε επανάληψη της μεθόδου της Σταδιακής Καθόδου Δέσμης. Για την διευκόλυνσή σας, σας δίνεται στην ιστοσελίδα του μαθήματος στο eclass (https://eclass.upatras.gr/modules/units/?course=ECON1332&id=9946) έναν σκελετό κώδικα σε Python (αρχείο Python-BatchGradientDescent-Template.rar), προκειμένου να ξεκινήσετε από εκεί την υλοποίηση της μεθόδου Σταδιακής Καθόδου Δέσμης συμπληρώνοντας όσα τμήματα λείπουν.

Αφού έχετε υλοποιήσει τη συνάρτηση batchGradientDescent, χρησιμοποιείστε την για να εκτιμήσετε τους συντελεστές του γραμμικού μοντέλου παλινδρόμησης που αναφέρεται παραπάνω. Θέστε τις κατάλληλες τιμές για την παράμετρο μάθησης α και το πλήθος επαναλήψεων numIters για την εκτίμηση των συντελεστών. Το πρόγραμμά σας θα πρέπει, αφού έχουν εκτιμηθεί οι συντελεστές, να εμφανίζει:

a) τις τιμές των συντελεστών που εκτιμήθηκαν και
b) να απεικονίζει με γραφική παράσταση τις τιμές της συνάρτησης κόστους ως συνάρτηση του πλήθους επαναλήψεων ώστε να τεκμηριωθεί ότι επιλέξατε σωστά τις τιμές για την παράμετρο μάθησης α και το πλήθος επαναλήψεων.

Επιπλέον, συγκρίνετε τους συντελεστές που εκτιμήθηκαν με τη μέθοδο της Σταδιακής Καθόδου Δέσμης με τους συντελεστές που εκτιμήθηκαν από τη μέθοδο των ελαχίστων τετραγώνων στο υποερώτημα i). Τί παρατηρήσεις/σχόλια μπορείτε να κάνετε;

Θέμα 3

Από το αρχείο με όνομα «Κεφάλαιο-06-Ασκήσεις.pdf» που μπορείτε να το βρείτε στην ενότητα "Lecture 3: Regression analysis" στον ιστότοπο του μαθήματος (https://eclass.upatras.gr/modules/units/?course=ECON1332&id=9946), απαντήστε στα **ζητούμενα της άσκησης 19**. Απαντήστε **μόνο στα υποερωτήματα i) και ii) της άσκησης 19**.

ΣΗΜΕΙΩΣΗ: Η απάντηση στο υποερώτημα iii) είναι προαιρετική.

Θέμα 4

Από το αρχείο με όνομα «Κεφάλαιο-06-Ασκήσεις.pdf» που μπορείτε να το βρείτε στην ενότητα "Lecture 3: Regression analysis" στον ιστότοπο του μαθήματος (https://eclass.upatras.gr/modules/units/?course=ECON1332&id=9946), απαντήστε στα **ζητούμενα της άσκησης 29**.

Θέμα 5

Συγγράψτε πρόγραμμα **μόνο σε R** που εκτιμά ***και με τη μέθοδο των Ελαχίστων Τετραγώνων (OLS) και με τη μέθοδο της Σταδιακής Καθόδου Δέσμης (Batch Gradient Descent)*** τους συντελεστές του παρακάτω πολλαπλού γραμμικού μοντέλου παλινδρόμησης, χρησιμοποιώντας το σύνολο δεδομένων εκπαίδευσης HouseholdData.csv που δίνεται μαζί με την εκφώνηση

$$Κατανάλωση\ τροφίμων = \beta_1 Εισόδημα + \beta_2 Αριθμός\ ατόμων\ νοικοκυριού + \beta_0$$

Για την εκτίμηση των συντελεστών με τη μέθοδο της Σταδιακής Καθόδου Δέσμης (Batch Gradient Descent) μπορείτε να κάνετε χρήση της υλοποίησης της μεθόδου που υπάρχει στο eclass στο αρχείο R-GradientDescent.rar (https://eclass.upatras.gr/modules/units/?course=ECON1332&id=9946), αφού τροποποιήσετε τον κώδικα ώστε να λαμβάνει υπόψιν τα δεδομένα του αρχείου HouseholdData.csv. Φροντίσετε επίσης να προσδιορίσετε την κατάλληλη τιμή της παραμέτρου μάθησης α καθώς και το πλήθος επαναλήψεων.
Μπορείτε να βρείτε επίσης έτοιμο τον κώδικα για την εκτίμηση των συντελεστών με τη μέθοδο των Ελαχίστων Τετραγώνων στην R για το παραπάνω πολλαπλό μοντέλο γραμμικής παλινδρόμησης και για το σύνολο δεδομένων του αρχείου HouseholdData.csv στον χώρο του μαθήματος στο eclass στο αρχείο R-LinearRegression.rar.
Ζητούνται τα εξής:

i. Αφού εκτελέσετε για το ίδιο μοντέλο γραμμικής παλινδρόμησης που αναφέρεται παραπάνω και το ίδιο σύνολο εκπαίδευσης HouseholdData.csv τις δύο μεθόδους εκτίμησης συντελεστών, συγκρίνετε τους συντελεστές που προέκυψαν από τις δύο μεθόδους αυτές. Τί συμπεράσματα μπορείτε να κάνετε σχετικά με το πόσο αποκλίνουν οι αντίστοιχοι συντελεστές;
ii. Αν διαπιστώσετε απόκλιση των εκτιμημένων συντελεστών, που πιστεύετε ότι οφείλεται η απόκλιση που παρατηρείται στις τιμές ορισμένων συντελεστών που εκτιμώνται με τη μέθοδο της Σταδιακής Καθόδου; Αναφέρετε, δίχως να υλοποιήσετε/προγραμματίσετε, τρόπο με τον οποίο μπορεί να αντιμετωπιστεί η απόκλιση αυτή. Για να απαντήσετε στο ερώτημα αυτό, διαβάστε από το αρχείο 06Regression.pdf που υπάρχει διαθέσιμο στο χώρο του μαθήματος στο eclass τις

ενότητες από: *«6.6.3 Εκτίμηση συντελεστών γραμμικών μοντέλων παλινδρόμησης: Η μέθοδος της Σταδιακής Καθόδου (Gradient Descent)»* έως και την ενότητα *«6.6.4 Σύγκριση μεθόδων Ελαχίστων Τετραγώνων και Σταδιακής Καθόδου»*.

Θέμα 6

Συγγράψτε πρόγραμμα **σε R και Python** που εκτιμά τους συντελεστές ενός γραμμικού μοντέλου παλινδρόμησης με τη μέθοδο της **Στοχαστικής Σταδιακής Καθόδου (Stochastic Gradient Descent)**.

Χρησιμοποιείστε τα προγράμματα σε R και Python που έχετε συγγράψει, θέτοντας τις κατάλληλες τιμές για την παράμετρο μάθησης α και το πλήθος επαναλήψεων, και εκτιμήστε τους συντελεστές του γραμμικού μοντέλου παλινδρόμησης που αναφέρεται στο θέμα 2 της τρέχουσας εργασίας και για το ίδιο σύνολο δεδομένων (Communities and Crime Data Set). Τα προγράμματά σας σε R και Python θα πρέπει να εμφανίζουν στην οθόνη:

a) τους συντελεστές που έχουν εκτιμηθεί και
b) τη γραφική παράσταση των τιμών της συνάρτησης κόστους όπως αυτές έχουν προκύψει από την εκτέλεση της μεθόδου Στοχαστικής Σταδιακής Καθόδου που έχετε υλοποιήσει.

Τί παρατηρήσεις μπορείτε να κάνετε τόσο για τους συντελεστές όσο και για τη γραφική παράσταση των τιμών της συνάρτησης κόστους που έχουν προκύψει με τη μέθοδο της Στοχαστικής Σταδιακής Καθόδου εάν τους συγκρίνετε με τους συντελεστές και τις τιμές της συνάρτησης κόστους που προέκυψαν στο υποερώτημα ii) του θέματος 2 της τρέχουσας εργασίας;

Θέμα 7

Κατεβάστε από την ιστοσελίδα https://archive.ics.uci.edu/ml/datasets/Forest+Fires σύνολο δεδομένων για πυρκαγιές από περιοχές της Πορτογαλίας. Τα δεδομένα περιέχουν γεωγραφικά και μετεωρολογικά στοιχεία όταν εκδηλώθηκαν πυρκαγιές καθώς επίσης και την επιφάνεια που κάηκε που μετριέται σε εκτάρια[1] (hectars). Έχοντας ως στόχο την πρόβλεψη της επιφάνειας που θα καεί βάσει των μετεωρολογικών συνθηκών που επικρατούν, συγγράψτε κώδικα **σε R και Python** που εκτιμά τους συντελεστές του παρακάτω μοντέλου παλινδρόμησης και κάνει μια εκτίμηση της ακρίβειας πρόβλεψης του μοντέλου

$$area = \beta_1 temp + \beta_2 wind + \beta_3 rain + \beta_0$$

Ειδικότερα ζητούνται τα εξής:

i. Χρησιμοποιώντας όλες τις παρατηρήσεις στο αρχείο που έχετε κατεβάσει, κάντε χρήση διασταυρωτικής επικύρωση 10-πτυχών (10-Fold Cross Validation) κατά την οποία θα εκτιμώνται οι συντελεστές του παραπάνω μοντέλου παλινδρόμησης με τη μέθοδο των Ελαχίστων Τετραγώνων (OLS) και επιπλέον θα υπολογίζει το μέσο τετραγωνικό σφάλμα (Root Mean Squared Error – RMSE) της πρόβλεψης. Το πρόγραμμά σας θα πρέπει να εμφανίζει το μέσο τετραγωνικό σφάλμα (RMSE).

[1] 1 εκτάριο = 10 στρέμματα

ii. Εκτιμήστε πάλι τους συντελεστές του ίδιου μοντέλου παλινδρόμησης με τη μέθοδο των Ελαχίστων Τετραγώνων και διασταυρωτικής επικύρωσης 10-πτυχών (10-Fold Cross Validation), αλλά αυτή τη φορά χρησιμοποιείστε όχι ολόκληρο το σύνολο δεδομένων αλλά μόνο εκείνες τις παρατηρήσεις όπου η τιμή της εξαρτημένης μεταβλητής (μεταβλητή area) είναι μικρότερη από 3.2 εκτάρια (δηλαδή area < 3.2) και χαρακτηρίζει μικρές πυρκαγιές. Εμφανίστε το μέσο τετραγωνικό σφάλμα (Root Mean Squared Error – RMSE) της πρόβλεψης.

iii. Τί συμπέρασμα μπορείτε να βγάλετε σχετικά με την ακρίβεια της πρόβλεψης, αν συγκρίνετε τα μέσα τετραγωνικά σφάλματα που εκτιμήσατε στις περιπτώσεις i) και ii) παραπάνω;

Για την υλοποίηση της διασταυρωτικής επικύρωσης k-πτυχών στο περιβάλλον της R, μελετήστε από το κεφάλαιο 06Regression.pdf που υπάρχει στον χώρο του μαθήματος στο eclass, την ενότητα *«6.7.2 Γραμμικά μοντέλα γραμμικής παλινδρόμησης με στόχο την πρόβλεψη»*. Επιπλέον, ο κώδικας R που υλοποιεί διασταυρωτική επικύρωση k-πτυχών και αναφέρεται στην ενότητα 6.7.2 του κεφαλαίου 06Regression.pdf, υπάρχει διαθέσιμος και στον χώρο του μαθήματος στο eclass στο αρχείο R-k-FoldCrossValidation.rar στην ιστοσελίδα της ενότητας (https://eclass.upatras.gr/modules/units/?course=ECON1332&id=9946) . Στην ίδια ενότητα υπάρχει και κώδικας που υλοποιεί τη μέθοδο της διασταυρωτικής επικύρωσης k-πτυχών στην Python (αρχείο Python-k-foldCrossValidation.rar). Μπορείτε να χρησιμοποιήσετε τον κώδικα αυτόν για την δημιουργία των προγραμμάτων σας, αφού κάνετε (προφανώς) τις κατάλληλες αλλαγές.

Θέμα 8

Από το αρχείο με όνομα «Κεφάλαιο-06-Ασκήσεις.pdf» που μπορείτε να το βρείτε στην ενότητα "Lecture 3: Regression analysis" στον ιστότοπο του μαθήματος (https://eclass.upatras.gr/modules/units/?course=ECON1332&id=9946), απαντήστε στα **ζητούμενα της άσκησης 58**.

ΣΗΜ: Μπορείτε εσείς να καθορίσετε το πλήθος των μεταβλητών στα πλασματικά δεδομένα καθώς και το γραμμικό μοντέλο παλινδρόμησης. Για την εκτίμηση των συντελεστών του γραμμικού μοντέλου θα πρέπει να γίνεται χρήση των συνθετικών δεδομένων που έχετε παράξει. Για την εκτίμηση των συντελεστών, μπορείτε να κάνετε χρήση της μεθόδου OLS.

Ομάδες εργασίας

Η εργασία θα εκπονηθεί ομαδικά. Οι ίδιες ομάδες που εκπόνησαν την Εργασία 1 θα εκπονήσουν και την Εργασία 2

Παράδοση της εργασίας: Τί και πως;

Κάθε ομάδα θα πρέπει να παραδώσει μία αναφορά σε αρχείο μορφής .pdf, γραμμένη σε LaTeχ, η οποία θα περιέχει τις απαντήσεις στα ζητούμενα των θεμάτων και μέσα στην ίδια αναφορά τον κώδικα σε R και Python που υπολογίζει τα ζητούμενα της εκφώνησης. Επιπλέον, ο κώδικας σε R και Python που έχετε δημιουργήσει για όλα τα θέματα, θα πρέπει να σταλεί και σε ξεχωριστά αρχεία (ένα ή περισσότερα) με τη μορφή κειμένου, ώστε να μπορεί να εκτελείται. Το αρχείο της αναφοράς καθώς και τα αρχεία με τον κώδικα R και Python θα πρέπει να συμπιεστούν σε ένα αρχείο (μορφή .zip ή .rar) και να υποβληθούν στον κατάλληλο χώρο του eclass.

Αναλυτικότερες πληροφορίες για τον τρόπο παράδοσης θα δοθεί κατά τη διάρκεια των διαλέξεων.

ΠΡΟΣΟΧΗ! Καταληκτική ημερομηνία παράδοσης της 2ης εργασίας είναι: Πέμπτη 24 Νοεμβρίου 2022.

Ερωτήσεις/Απορίες

Για οποιαδήποτε ερώτηση ή απορία σχετικά με την εργασία μπορείτε να στείλετε email στη διεύθυνση tzagara@upatras.gr . Απορίες μπορούν επίσης (**και συστήνεται!**) να συζητηθούν κατά τη διάρκεια του μαθήματος.

Βαρύτητα της εργασίας

Η 2η εργασία είναι υποχρεωτική για τους φοιτητές και συνεισφέρει το 10% του τελικού τους βαθμού.

Καλή επιτυχία!