ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

Αναγνώριση Προτύπων

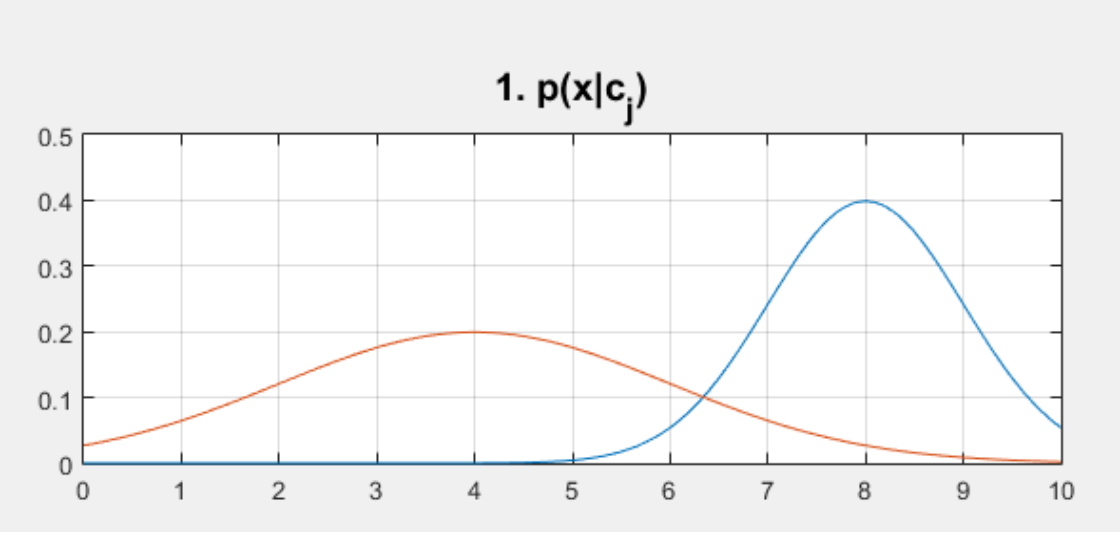
1η Εργασία
Εμμανουήλ Παπαδημητρίου
ΑΜ: mcse19021

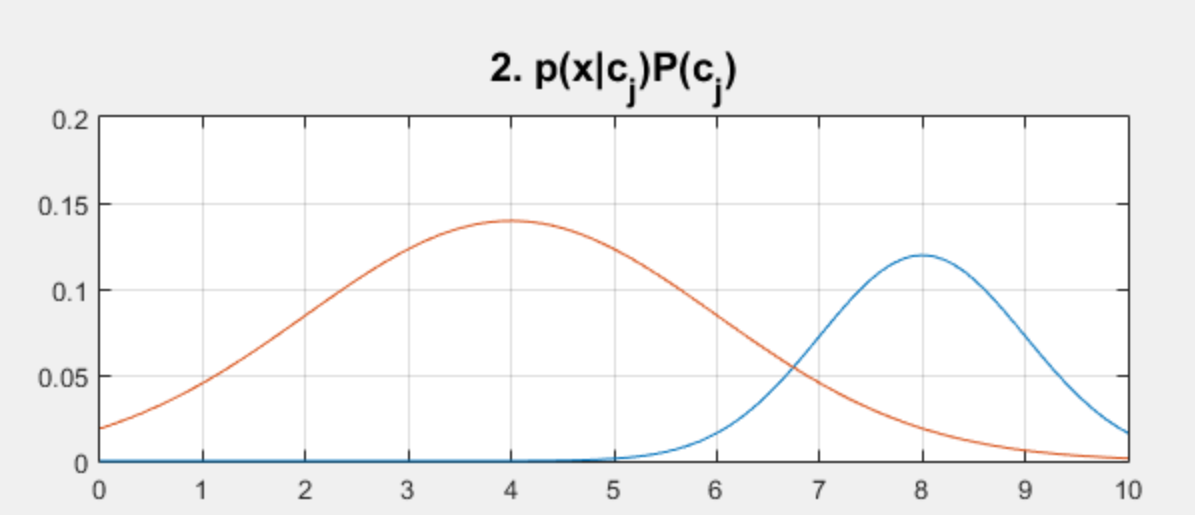
Α' Μέρος

Γραφήματα

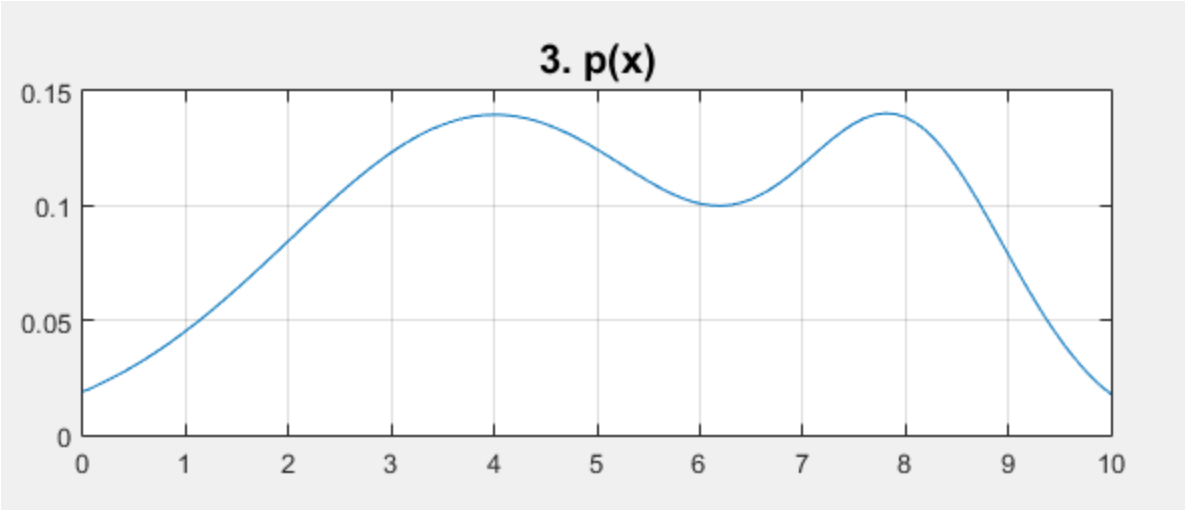
1) Οι δύο συναρτήσεις πιθανοφάνειας $p(x|\omega_i)$ για τις δύο κατηγορίες μαθητών.

2) Οι δύο συναρτήσεις διάκρισης $p(x|\omega_i)P(\omega_i)$.

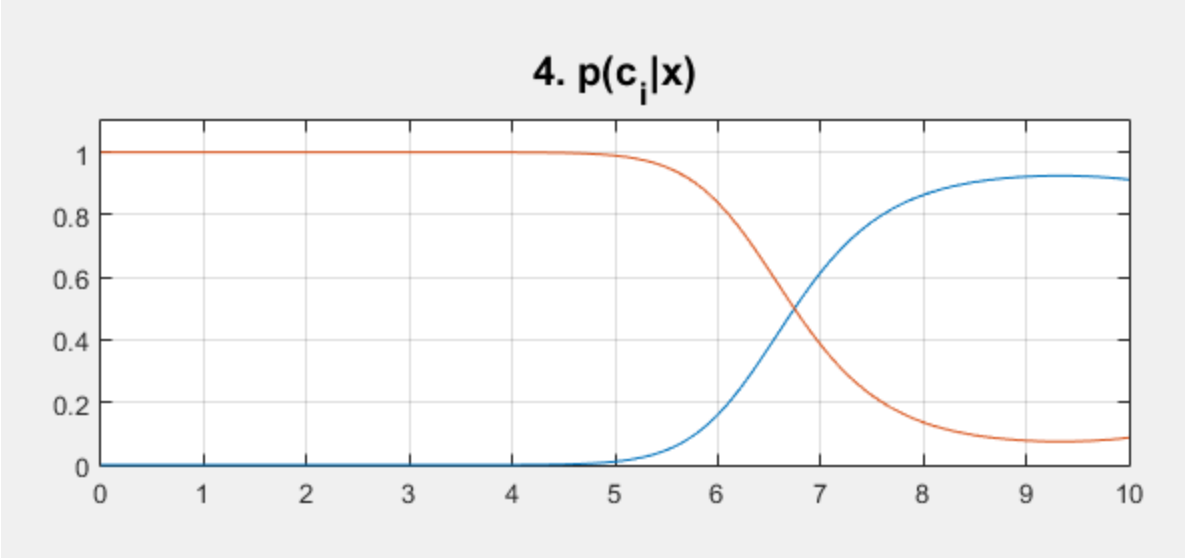




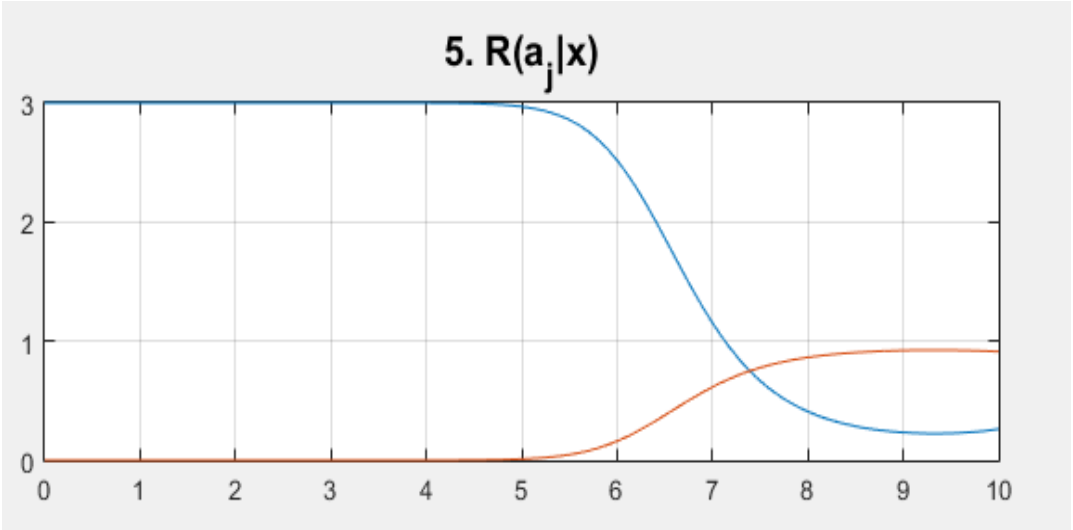
3) Η ολική πιθανότητα $p(x)$ για κάθε βαθμολογία.



4) Η εκ των υστέρων πιθανοφάνεια $p(\omega i|x)$ για κάθε κατηγορία.



5) Το ρίσκο για κάθε ενέργεια $R(a_i|x)$, δηλαδή, το ρίσκο να ταξινομηθεί ένας μαθητής ως καλός και ως μέτριος, αντίστοιχα.



Ερωτήσεις

Α1) Το εύρος τιμών βαθμολογίας στο οποίο έχουμε μεγαλύτερη πιθανοφάνεια να είναι καλός ο μαθητής είναι:

6.4 έως 10

Α2)

Πρώτη λύση

Επειδή έχουμε ότι ισχύει

$$R(a_1|x) < R(a_2|x)\ ,$$

και στην συνέχεια με το θέωρημα του Bayes κάνουμε υπολογισμούς και χρησιμοποιόντας τα δοσμένα μ και σ, καταλήγουμε στο

$$\frac{e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}}{e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}} > \frac{\sigma_1[\lambda(\alpha_1|\omega_2) - \lambda(\alpha_2|\omega_2)]P(\omega_2)}{\sigma_2[\lambda(\alpha_2|\omega_1) - \lambda(\alpha_1|\omega_1)]P(\omega_1)}$$

Με αντικατάσταση των μ1,μ2 και σ1,σ2 έχουμε

$$\frac{e^{-\frac{(x-8)^2}{2*1^2}}}{e^{-\frac{(x-4)^2}{2*2^2}}} > \frac{1*[\lambda(\alpha1\,|\,\omega2)\ -\lambda(\alpha2|\omega2)]P(\omega2)}{2*[\lambda(\alpha2\,|\,\omega1)\ -\lambda(\alpha1|\omega1)]P(\omega1)}$$

Τα λ(α1|ω1) και λ(α2|ω2) έχουν τιμή 0 γιατί σύμφωνα με την εκφώνηση, είναι σωστές ταξινομήσεις.

Και από εκφώνηση έχουμε και τις ταξινομήσεις ενός μέτριου ως καλού μαθητή και ενός καλού ως μέτριου μαθητη.
$\lambda_{gm}=\lambda(good|moderate) = \lambda(\alpha1|\omega2) = 3$ και $\lambda_{mg}=\lambda(moderate|good)=\lambda(\alpha2|\omega1) = 1$.
Άρα, με όλες τις αντικαταστήσεις έχουμε

$$\frac{e^{-\frac{(x-8)^2}{2*1^2}}}{e^{-\frac{(x-4)^2}{2*2^2}}} > \frac{1*[3\ -0]*0.7}{2*[1\ -0]*0.3}$$

Και με την βοήθεια λογάριθμών έχουμε

$$-\frac{(x-8)^2}{2*1^2} + -\frac{(x-4)^2}{2*2^2} > 1.2528$$

Με πράξεις έχουμε την τελική μορφή της ανισότητας

$$-3x^2 + 56x - 250.0224 > 0$$

Με την εντολή roots(ανισότητα) έχουμε τις λύσεις και μας δίνει τις δύο παρακάτω:

11.275062230762405
7.391604435904262

Δεχόμαστε την 7.391604435904262 γιατί ανήκει στο διάσημα [0,10] και με στρογγυλοποίηση έχουμε 7.4
Άρα το εύρος είναι:

7.4 έως 10

Δεύτερη Λύση

Χτίστηκε αλγόριθμος που υλοποιηεί το θέωρημα του Bayes όπως φαίνεται εδώ

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{[\lambda(\alpha_1|\omega_2) - \lambda(\alpha_2|\omega_2)]P(\omega_2)}{[\lambda(\alpha_2|\omega_1) - \lambda(\alpha_1|\omega_1)]P(\omega_1)}$$

Για όλες τις τιμές του x, ελέγχουμε αν ισχύει η ανισότητα.
Μόλις βρούμε το πρώτο που την ικανοποιεί, σταματάμε την loop.
Έτσι έχουμε σαν αποτέλεσμα την θέση του 7.4 στον πίνακα του x.

Άρα, το εύρος είναι **7.4 έως 10** όπως και στην πρώτη λύση (μετά την απαραίτητη στρογγυλοποίηση).

A3)

Πρώτη Λύση

Στόχος μας είναι να κάνουμε δοκιμές με τα λ_{gm} και λ_{mg}.

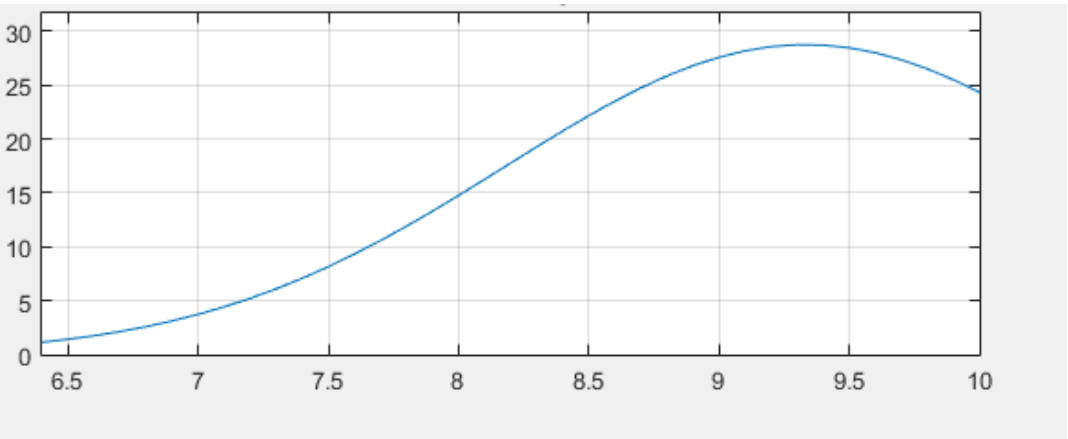
Οι τιμές αυτές επηρεάζουν το δεξί μέρος της ανισότητας, άρα στο αρχείο matlab, υπολογίζεται δυναμικά ο αριθμός σύμφωνα με τα λgm και λmg

Μετά από δοκιμές στα λ_{gm} και λ_{mg}.Οι τιμές που βγάζουν εύρος 6.4 έως 10 είναι: □□□ = **1 και** □□□ = **2.1 έως 2.5 (λόγω στρογγυλοποίησης)**

Δεύτερη Λύση

Στον πίνακα L, έγιναν δοκιμές και βρέθηκε ότι με τιμή **1 και 2.5** έχουμε εύρος 6.4. Έγινε χρήση του

αλγορίθμου στο A2 και βρέθηκε η τιμή **6.4**



A4)

Βασισμένοι στο Bayesian ρίσκο, χτίζουμε τα ρίσκα

R(good | x) = λmg * p(moderate | x)
R(moderate | x) = λgm * p(good | x)

Θα χρησιμοποιήσουμε το ρίσκο R(good | x) σε αυτό το ερώτημα.

Βλέποντας τις τιμές του πίνακα R(good | x), παρατηρούμε ότι είναι ταξινομημένος κατά φθίνουσα σειρά **ΜΕΧΡΙ** ένα σημείο και από το σημείο αυτό και μετά είναι ταξινομημένος κατά αύξουσα σειρά. Άρα χρειαζόμαστε αυτό το στοιχείο, και το βρίσκουμε αυτό το στοιχείο υπολογίζοντας σε ποιο σημείο του πίνακα R(good | x) συμβαίνει αυτό.

Βρίσκουμε ότι συμβαίνει στη θέση 94 με **βαθμολογία 9.3 και ρίσκο 0.225043.**

A5)

Ο πίνακας prios Pc = [0.3 0.7] έχει αυτή τη μορφή τώρα.

Μετά από δοκιμές βρέθηκε ότι με τις παρακάτω τιμές **Pc = [0.5 0.5]** έχουμε όλες τις βαθμολογίες μεγαλύτερες ή ίσες του 6.

A6)
Στόχος μας η εύρεση και των δύο priors για Pc = [0.5 0.5] με δεδομένο ότι βαθμολογία >=6

P(good) = P(ω1)
P(moderate) = P(ω2)

Β' Μέρος

B1)
Οι δύο κλάσεις είναι οι P(Class =No) και P(Class=Yes)

Οι πιθανότητες τους είναι:

- P(Class = No) = 0.7
- P(Class = Yes) = 0.3

B2)
Για την εύρεση της κλάσης που ανήκει το δείγμα X, πρέπει να υπολογιστεί η παρακάτω εξίσωση για τις δύο κλάσεις αντίστοιχα:

- p(Class = No|X) = p(X|Class=No) * p(Class = No) / p(X)
- p(Class = Yes|X) = p(X|Class=Yes) * p(Class = Yes) / p(X)

-Το p(X) το χρειαζόμαστε για κανονικοποίηση των δεδομένων:

p(X) = p(Owner=Yes) * p(Status=Divorced) * p(Income=100) = 0.0175

-Το p(X|Class=No) είναι ίσο με :

p(Owner = Yes|Class=No) * p(Status=Divorced|Class=No) * p(Income=100|Class=No)
= 0.00044034

-Το p(X|Class=Yes) είναι ίσο με :

p(Owner = Yes|Class=Yes) * p(Status=Divorced|Class=Yes) * p(Income=100|Class=Yes)
= 0

-p(Class = No) = 0.7
-p(Class = Yes) = 0.3

Τελικά Αποτελέσματα χωρίς κανονικοποίηση:

- **p(Class=No|X) = 0.0003**
- **p(Class=Yes|X) = 0.0000**

Με κανονικοποίηση:

- **p(Class=No|X) = 0.0176**
- **p(Class=Yes|X) = 0.0000**

Άρα το δείγμα X ανήκει στην κλάση **p(Class = No)**

B3)

Υπάρχει η ανάγκη να εφαρμοστεί Laplacian smoothing, καθώς όπως είδαμε παραπάνω το **p(Owner=Yes|Class=Yes) = 0.**

Είναι προβληματικό όταν μία δεσμευμένη πιθανότητα είναι 0, καθώς θα εξαλείψει όλα τα αποτελέσματα που συσχετίζονται με αυτό, όπως έγινε με τον υπολογισμό του **p(Class=Yes|X).**

B4)

Με Laplacian smoothing m = 1, έχουμε νέα αποτελέσματα

- **p(Class=No|X) = 0.0004**
- **p(Class=Yes|X) = 0.0002**

Οπότε και πάλι το δείγμα X ανήκει στην κλάση **p(Class=No).**

Έχουμε νέες τιμές καθώς με την τιμή του m έχουμε νέες δεσμευμένες πιθανότητες.

Στο p(Owner=Yes|Class=Yes) πριν το smoothing είχαμε 0, τώρα υπάρχει τιμή :

p(Owner=Yes|Class=Yes)
= ((φορές που Owner = Yes σε Κλάση Yes) + m) / (πλήθος Κλάσης Yes + πλήθος πιθανών απαντήσεων * m)

Όπως βλέπουμε, δεν υπάρχει περίπτωση η δεσμευμένη πιθανότητα να είναι 0, καθώς προστίθεται το smoothing κάθε φορά.

Ομοίως ο υπολογισμός και στις άλλες δεσμευμένες πιθανότητες.

B5)

Με Laplacian smoothing m = 100, έχουμε νέα αποτελέσματα

- **p(Class=No|X) = 0.0008**
- **p(Class=Yes|X) = 0.0005**

Βλέπουμε ξανά ότι το δείγμα X ανήκει στην κλάση **p(Class=No).**
Το οποίο είναι λογικό, καθώς αν υπήρχε περίπτωση το δείγμα X να ανήκει στην κλάση p(Class=Yes), θα το είχαμε δει όταν χρησιμοποιήσαμε αρχικά το laplacian smoothing. Εφόσον παρέμεινε η ίδια κλάση μετά το smoothing, δεν μπορεί υπάρξει άλλη τιμή που να κάνει το συγκεκριμένο δείγμα να ανήκει στην κλάση p(Class=Yes). Άρα για όλες τις τιμές του Laplacian smoothing m, το δείγμα X θα ανήκει στην κλάση **p(Class=No).**

B6)

Δημιουργήθηκε ένας αλγόριθμος που ελέγχει όλες τις τιμές του Laplacian smoothing από 1 έως 100 και για κάθε τιμή του m, και για Income από 1 έως 150.

Τα αποτελέσματα αποθηκεύθηκαν στα αρχεία με όνομα “b6_no.txt” και “b6_yes.txt”, για τις κλάσεις p(Class=No) και p(Class=Yes) αντίστοιχα.

Ο αλγόριθμος, επειδή κάνει πολλούς υπολογισμούς, είναι αρκετά αργός οπότε μπορείτε να τον κάνετε commented. Αλλιώς, κρατήστε τα σχόλια για να τον τρέξετε και εσείς. Τα αρχεία είναι ήδη συμπληρωμένα με τα αποτελέσματα.

Οπότε, για τιμές του m από 1 έως 100 και Income από 1 έως 150 βρέθηκαν ότι για όλα τα **Income από 82 έως 97 για m από 1 έως 100**, το δείγμα X ανήκει στην κλάση **p(Class = Yes).**

Όμως, παρατήρηθηκε ότι στις τιμές του Income 81 και 98:

- Για Income = 81, το δείγμα X ανήκει στην κλάση **p(Class = Yes)** για τιμές του **Laplacian Smoothing m>=3** και για **m=1,2** ανήκει στην κλάση **p(Class = No)**

- Για Income = 98, το δείγμα X ανήκει στην κλάση p(Class = Yes) για τιμές του **Laplacian Smoothing m>=2** και για **m=1** ανήκει στην κλάση **p(Class = No)**

Επομένως, οι τιμές που αναζητούμε του **Income είναι 81 και 98**.

B7)

Δημιουργήθηκε ένας αλγόριθμος για τον υπολογισμό του λόγου των υστέρων πιθανοτήτων p(Class = No|X) / p(Class=Yes) για Laplacian Smoothing m = 100000 και για τιμές του Income από 1 έως 150.

Το αποτέλεσμα είναι ότι **ΔΕΝ** βρέθηκαν τιμές που να κάνουν τον λόγο ακριβώς ίσον με 1, αλλά βρέθηκαν δύο τιμές του Income που τείνουν πολύ. Οι τιμές είναι:

- Income = 81, ο λόγος έχει αποτέλεσμα 0.93837
- Income = 99, ο λόγος έχει αποτέλεσμα 1.05908

Αν είναι να δεχθούμε κάποια τιμή, αυτή ίσως να είναι για **Income = 99** καθώς είναι το πιο κοντινό στο 1.

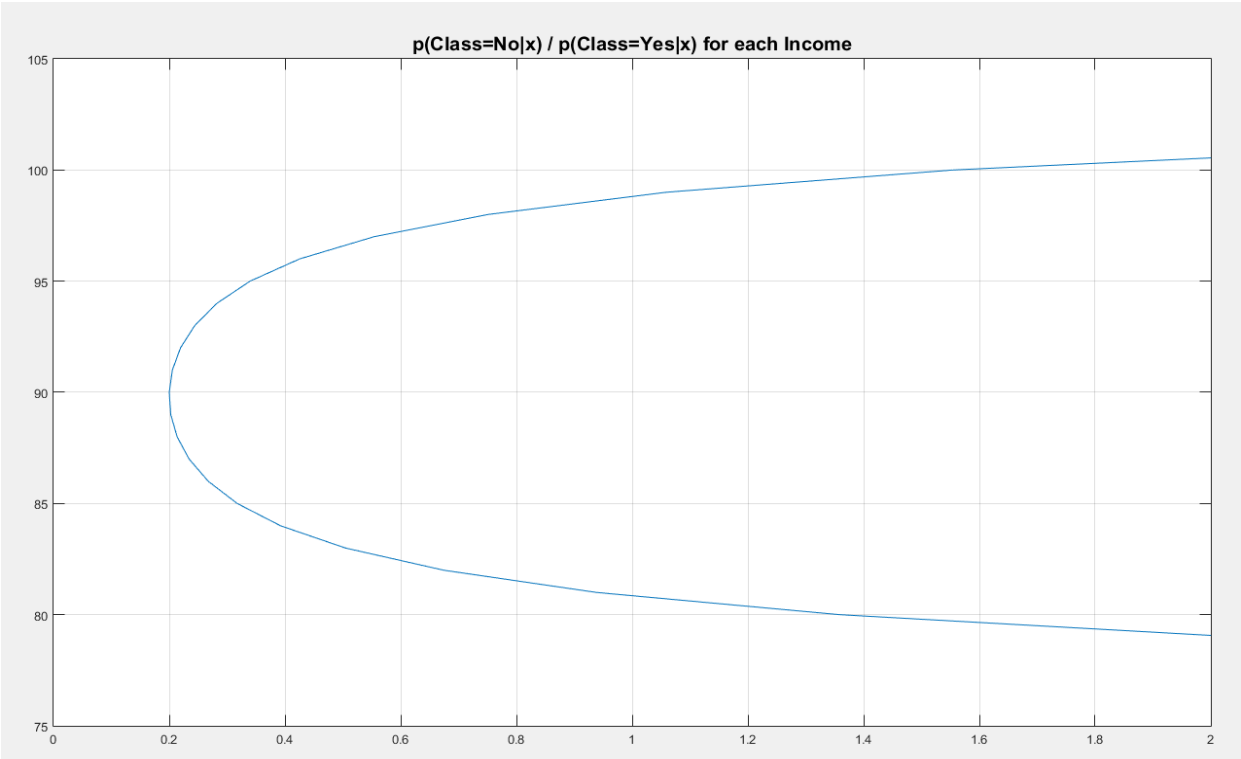
Άρα, δεν υπάρχει τιμή για την οποία η τράπεζα που να μην μπορεί να αποφασίσει εάν το συγκεκριμένο δείγμα είναι υπερήμερος οφειλέτης ή οχι

B8)

Όπως αναφέρθηκε προηγουμένως, οι τιμές για το Income που τείνουν στο 1 είναι 81 και 91 με την 91 να είναι η πιο κοντινή.

Ανάλυση αλγορίθμου:

1. Αρχικά, τέθηκε μια τιμή του m=100000 (ζητήθηκε να είναι αρκετά μεγάλη).
2. Στην συνέχεια, ξεκινάει μια loop για τιμές του Income από 1 έως 150.
3. Μέσα στην loop, υπολογίζονται νέες τιμές των conditional probabilities για κάθε κλάση No και Yes (δηλαδή των Owner, Status και Income).
4. Στην συνέχεια γίνεται ο υπολογισμός των υστέρων πιθανοτήτων p(Class=No|X) και p(Class=Yes|X).
5. Και στο τέλος ελέγχουμε αν ο λόγος των υστέρων πιθανοτήτων είναι >=0.9 και <=1.1
6. Μετά από πολλούς υπολογισμούς, βρέθηκαν το 81 και το 91 σαν τιμές που κάνουν τον λόγο των υστέρων πιθανοτήτων να τείνει στο 1.


p(Class=No|x) / p(Class=Yes|x) for each Income
105
100
95
90
85
80
75
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2