ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

# Αναγνώριση Προτύπων

## 1η ΕΡΓΑΣΙΑ

Θέμα: "Παραδείγματα εφαρμογής της Bayesian θεωρίας αποφάσεων"

### Α' Μέρος

Θεωρήστε δύο κατηγορίες μαθητών: οι καλοί μαθητές (κατηγορία $\omega_1$) και οι μέτριοι μαθητές (κατηγορία $\omega_2$). Η πρότερη εμπειρία έχει δείξει ότι το 30% των μαθητών είναι καλοί και το 70% μέτριοι. Επιπλέον, οι καλοί μαθητές γράφουν βαθμούς που έχουν μια κανονική κατανομή με μέση τιμή $\mu = 8$ και τυπική απόκλιση $\sigma = 1$ ενώ η αντίστοιχη κατανομή βαθμολογίας για τους μέτριους μαθητές έχει μέση τιμή $\mu = 4$ και τυπική απόκλιση $\sigma = 2$.

Ορίζονται δύο ενέργειες $a_1$ και $a_2$ που αφορούν την ταξινόμηση κάποιου μαθητή ως καλού ή μέτριου ανάλογα με την βαθμολογία $x$ που έχει γράψει, με $0 \leq x \leq 10$. Οι σωστές ταξινομήσεις έχουν κόστος 0. Αντίθετα, το κόστος της ταξινόμησης ενός μέτριου μαθητή ως καλού είναι $\lambda_{gm} = \lambda(good|moderate) = 3$. Από την άλλη, το κόστος της ταξινόμησης ενός καλού μαθητή ως μέτριου είναι $\lambda_{mg} = \lambda(moderate|good) = 1$.

Να υπολογιστούν και να σχεδιαστούν οι καμπύλες που φαίνονται στην Εικόνα 1. Συγκεκριμένα, στα υπογραφήματα εμφανίζονται κατά σειρά:

1) Οι δύο συναρτήσεις πιθανοφάνειας $p(x|\omega_i)$ για τις δύο κατηγορίες μαθητών. [Λυμένο στον κώδικα που δίνεται]
2) Οι δύο συναρτήσεις διάκρισης $p(x|\omega_i)P(\omega_i)$. [Λυμένο στον κώδικα που δίνεται]
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.

<u>Ερωτήσεις</u>

Α1) Για ποιο εύρος τιμών βαθμολογίας έχουμε μεγαλύτερη πιθανοφάνεια $p(x|\omega_i)$ να είναι καλός ο μαθητής; [Λυμένο στον κώδικα που δίνεται]

A2) Για ποιο εύρος τιμών βαθμολογίας έχουμε μικρότερο ρίσκο να ταξινομηθεί ο μαθητής ως καλός;

A3) Γιατί διαφέρουν τα εύρη τιμών στα ερωτήματα Α1 και Α2; Προτείνετε μια αλλαγή στις τιμές κόστους ώστε τα εύρη αυτά να συμπίπτουν. Ξανατρέξτε το πρόγραμμα με τις τροποποιημένες παραμέτρους και κάντε αντιγραφή - επικόλληση το γράφημα που προκύπτει στο Word.

A4) Επαναφέρετε τις παραμέτρους στις αρχικές τιμές κόστους (πριν το ερώτημα Α3). Για ποια βαθμολογία ($x$) έχουμε το μικρότερο ρίσκο να ταξινομηθεί κάποιος μαθητής ως καλός και πόσο είναι το ρίσκο αυτό;

A5) Δοκιμάστε να βρείτε (μέσω δοκιμών) κάποιον συνδυασμό για τις εκ των προτέρων πιθανότητες (prior) ώστε για όλες τις βαθμολογίες $x \geq 6$ ο μαθητής να ταξινομείται με την μέγιστη εκ των υστέρων πιθανοφάνεια στην κατηγορία καλός, όπως φαίνεται στο υπογράφημα 4 της Εικόνας 2.

A6) Υπολογίστε αναλυτικά τα priors $P(\omega_1)$ και $P(\omega_2)$ για τα οποία ισχύει το ερώτημα Α5.

Οδηγίες

- Συμβουλευτείτε το αρχείο `ergasia1a_start.m` που σας δίνεται. Σε αυτό σχεδιάζονται ορισμένες από τις ζητούμενες καμπύλες και επίσης δίνεται η απάντηση στο ερώτημα Α1.
  Μετονομάστε το σε `ergasia1a.m` και συνεχίστε με αυτό.
- Για το ερώτημα Α6 θα χρειαστεί να χρησιμοποιήσετε στο κείμενό σας τον Equation Editor του Word. Συμβουλευτείτε το Baysian risk παράδειγμα στο αρχείο **PR_02 Bayesian θεωρία λήψης αποφάσεων - παραδείγματα.pdf**

**Β' Μέρος**

Θεωρήστε το σύνολο δεδομένων που δίνονται στον παρακάτω πίνακα.

| Home Owner | Marital Status | Annual Income | Defaulted Borrower |
|---|---|---|---|
| Yes | Single | 125K | No |
| No | Married | 100K | No |
| No | Single | 70K | No |
| Yes | Married | 120K | No |
| No | Divorced | 95K | Yes |
| No | Married | 60K | No |
| Yes | Divorced | 220K | No |
| No | Single | 85K | Yes |
| No | Married | 75K | No |
| No | Single | 90K | Yes |

Έστω άγνωστο δείγμα $\boldsymbol{x} = \{Home\ Owner = Yes, Martial\ Status = Divorced, Annual\ Income = 100\}$. Χρησιμοποιώντας Naive Bayesian προσέγγιση απαντήστε στα παρακάτω ερωτήματα:

Ερωτήσεις

B1) Υπολογίστε και εμφανίστε την εκ των υστέρων πιθανότητα για τις δύο κλάσεις Defaulted Borrower (υπερήμερος οφειλέτης). [Λυμένο στον κώδικα που δίνεται]

B2) Υπολογίστε και εμφανίστε σε ποια κλάση ανήκει το άγνωστο δείγμα.

B3) Υπάρχει ανάγκη να εφαρμοστεί Laplacian εξομάλυνση; Αιτιολογείστε την απάντησή σας στο κείμενο.

B4) Υπολογίστε και εμφανίστε την εκ των υστέρων πιθανότητα των δύο κλάσεων, για Laplacian smoothing με $m = 1$. Αλλάζει το αποτέλεσμα της ταξινόμησης σε σχέση με το ερώτημα B2; Αιτιολογήστε την απάντησή σας.

B5) Υπολογίστε και εμφανίστε την εκ των υστέρων πιθανότητα των δύο κλάσεων, για Laplacian smoothing με $m = 100$. Αλλάζει το αποτέλεσμα της ταξινόμησης σε σχέση με το ερώτημα B4; Αιτιολογήστε την απάντησή σας.

B6) Δοκιμάζοντας διάφορες τιμές για το εισόδημα ($Income$), αναφέρετε ένα παράδειγμα όπου η αύξηση του συντελεστή εξομάλυνσης $m$ από 1 σε 100 μεταβάλει το αποτέλεσμα της ταξινόμησης.

B7) Θεωρήστε πολύ μεγάλη τιμή για το $m$. Βρείτε μέσω δοκιμών κάποια τιμή του $Income$ για την οποία ο λόγος των εκ των υστέρων πιθανοτήτων $\frac{p(No|\boldsymbol{x})}{p(Yes|\boldsymbol{x})}$ να τείνει στην μονάδα. Με άλλα λόγια, υπάρχει κάποια τιμή του $Income$ για την οποία η τράπεζα «να μην μπορεί να αποφασίσει» εάν το συγκεκριμένο δείγμα είναι υπερήμερος οφειλέτης ή όχι;

B8) Στο ερώτημα B7 η τιμή για το $Income$ που βρήκατε είναι μοναδική ή υπάρχουν περισσότερες από μία; Αιτιολογήστε αναλυτικά την απάντησή σας στο κείμενο και προσδιορίστε επακριβώς την/τις λύσεις.

Οδηγίες

- Συμβουλευτείτε το αρχείο `ergasia1b_start.m` που σας δίνεται. Περιλαμβάνει την λύση στο ερώτημα B1.
  Μετονομάστε το σε `ergasia1b.m` και συνεχίστε με αυτό.
- Για το ερώτημα B8 θα χρειαστεί να χρησιμοποιήσετε στο κείμενό σας τον Equation Editor του Word.
  - Η βασική παραδοχή είναι η μεγάλη τιμή του συντελεστή εξομάλυνσης $m$ και το πώς αυτή επηρεάζει τις εκ των υστέρων πιθανότητες $p(No|\boldsymbol{x})$ και $p(Yes|\boldsymbol{x})$.
  - Για να έχετε μια αρχική εκτίμηση, μπορείτε να σχεδιάσετε σε ένα διάγραμμα τις εκ των υστέρων πιθανότητες $p(No|\boldsymbol{x})$ και $p(Yes|\boldsymbol{x})$ (όπως στην Εικόνα 1-4 του Α' Μέρους) για ένα εύρος τιμών του $Income$ π.χ. από 0 έως 150.
  - Δουλέψτε με τον λόγο $\frac{p(No|\boldsymbol{x})}{p(Yes|\boldsymbol{x})}$. Σκοπός σας είναι η λύση ως προς τις τιμές $x$ του $Income$ που κάνουν το κλάσμα να ισούται με μονάδα.

Παράδοση εργασίας

Το κείμενο **.docx** της εργασίας θα πρέπει αναφέρει τα στοιχεία σας και να περιέχει μια αναλυτική παρουσίαση των παραπάνω ερωτημάτων.

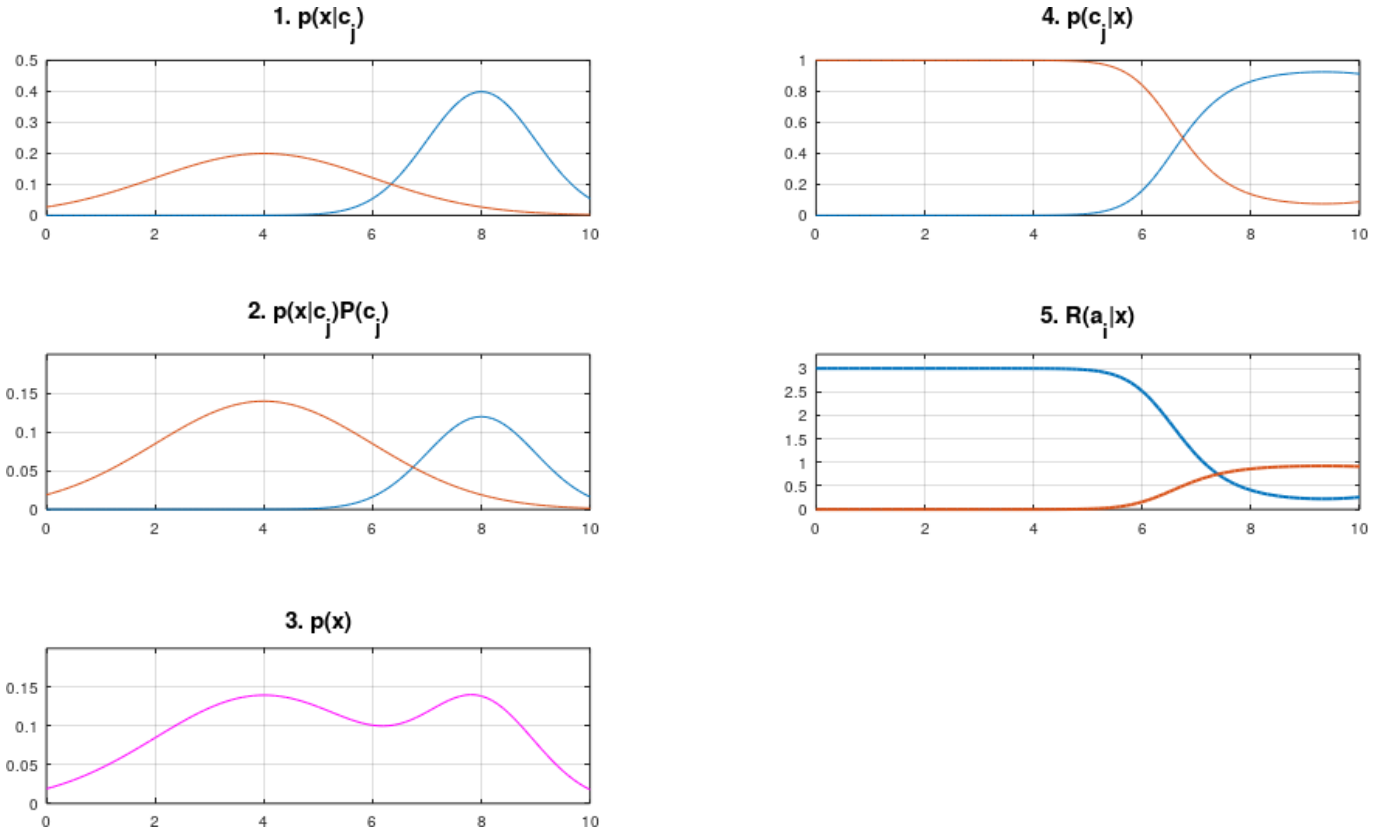
Παραδοτέο: ένα συμπιεσμένο αρχείο **ergasia1.zip** (ή **ergasia1.rar**) που να περιέχει τα αρχεία

- **ergasia1a.m**
- **ergasia1b.m**
- **ergasia1.docx**

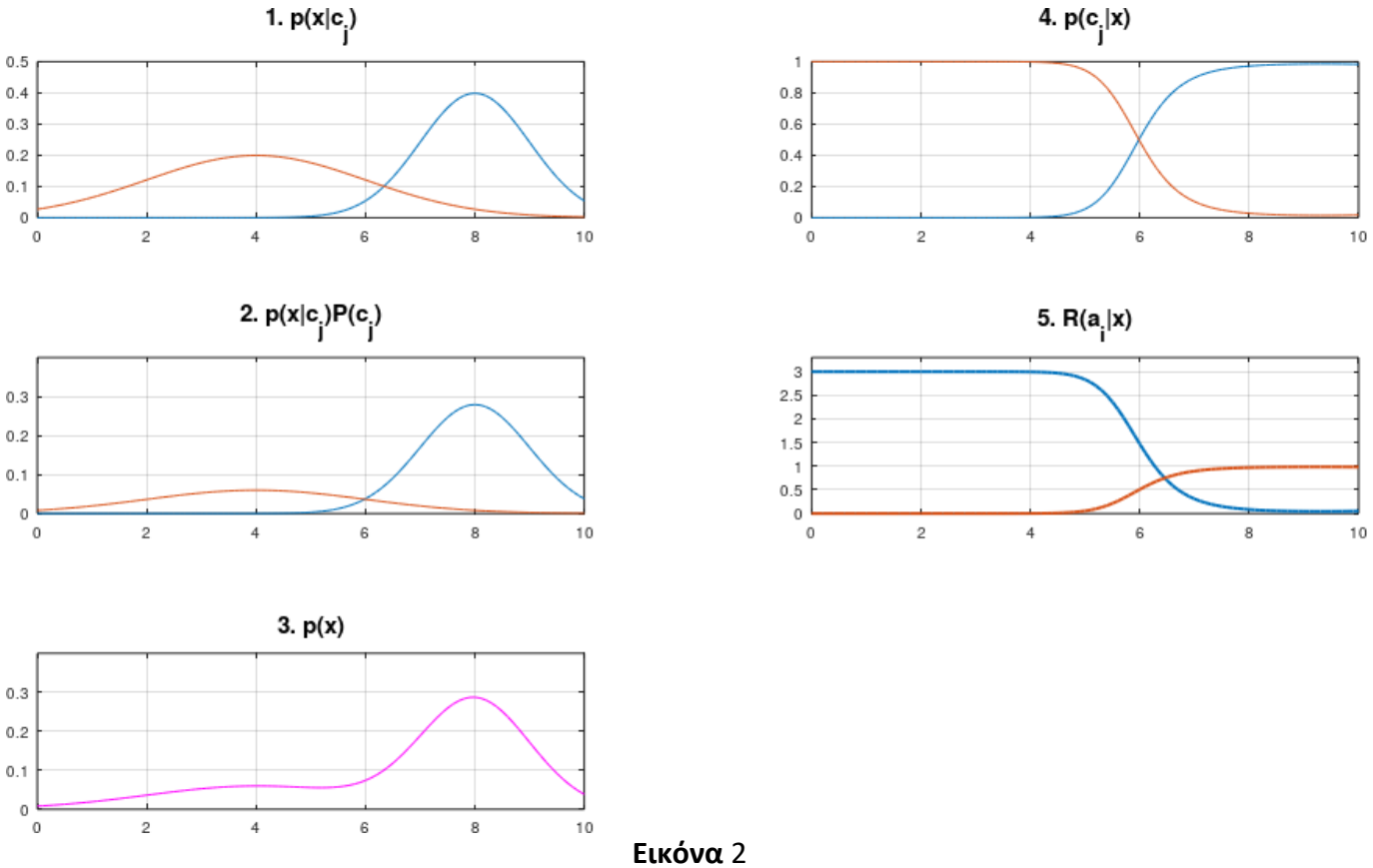
και αποστολή ΜΟΝΟ μέσω e-class στην ενότητα Εργασίες μέχρι και την **Κυριακή 24/11/2019** στις **12:00 το βράδυ**.

Για οποιαδήποτε διευκρίνιση μπορείτε να επικοινωνήσετε μαζί μου στο **akesidis@uniwa.gr**

Τάσος Κεσίδης




**Εικόνα** 1




**Εικόνα** 2