Αλεξάνδρειο ΤΕΙ Θεσσαλονίκης

Τμήμα Μηχανικών Πληροφορικής

Η μέθοδος Ταξινόμησης Naïve Bayes

Διαμαντάρας Κωνσταντίνος

Απρίλιος 2016

Βασικές αρχές πιθανοτήτων

Οι πιθανότητες αποτελούν ένα μέτρο της βεβαιότητας ότι θα συμβεί κάποιο ενδεχόμενο. Για παράδειγμα, η πιθανότητα ένα μη πειραγμένο ζάρι να φέρει "2" είναι ίση με 1/6, καθώς το ενδεχόμενο "2" είναι ένα από τα έξι ισοπίθανα ενδεχόμενα αποτελέσματα του ζαριού. Η τιμή που φέρνει το ζάρι καλείται **τυχαία μεταβλητή** (**random variable**) και συνήθως συμβολίζεται με ένα κεφαλαίο γράμμα, πχ X. Η πιθανότητα συμβολίζεται ως συνάρτηση $P()$ του ενδεχόμενου. Συνεπώς η πρόταση:

«Η πιθανότητα το ζάρι να φέρει "2" είναι 1/6»

γράφεται μαθηματικά ως εξής:

$$P(X = 2) = \frac{1}{6}$$

Το σύνολο όλων των ενδεχομένων μιας τυχαίας μεταβλητής (τ.μ.) συνήθως συμβολίζεται με Ω. Πολλές φορές είναι εύκολο να υπολογίσουμε τις πιθανότητες ενός ενδεχομένου αν υποθέσουμε ότι όλα τα ενδεχόμενα είναι ισοπίθανα. Στην περίπτωση αυτή δεν έχουμε παρά να διαιρέσουμε το πλήθος των περιπτώσεων στις οποίες ισχύει το ενδεχόμενο δια του συνολικού πλήθους των περιπτώσεων.

Παράδειγμα: Έστω ότι επιλέγουμε τυχαία έναν άνθρωπο X από όλους τους ανθρώπους του κόσμου και ενδιαφερόμαστε να βρούμε την πιθανότητα του ενδεχομένου «ο X να είναι Ασιάτης». Έχουμε

- Ω = το σύνολο όλων των ανθρώπων του κόσμου
- N = το πλήθος των ανθρώπων (= πλήθος στοιχείων του Ω)
- A = το σύνολο των Ασιατών
- N_A = το πλήθος των Ασιατών (= πλήθος στοιχείων του A)

Αν υποθέσουμε ότι όλοι οι άνθρωποι έχουν ίσες πιθανότητες να επιλεγούν τότε η πιθανότητα του ενδεχομένου «ο X είναι Ασιάτης» γράφεται ως ένα απλό κλάσμα:

$$P(X \in A) = \frac{N_A}{N}.$$

Η πιθανότητα $P(E)$ οποιουδήποτε ενδεχομένου E είναι πάντα μεγαλύτερη ή ίση από το μηδέν και επίσης μικρότερη ή ίση από το ένα:

$$0 \leq P(E) \leq 1.$$

Κατανομή πιθανότητας

Έστω ότι η τ.μ. μπορεί να πάρει K διακριτές τιμές $x_1, \cdots, x_K$ (συνηθίζεται οι τιμές των ενδεχόμενων να συμβολίζονται με μικρά γράμματα). Τότε λέμε ότι η μεταβλητή είναι διακριτή και έχουμε μια σειρά από πιθανότητες

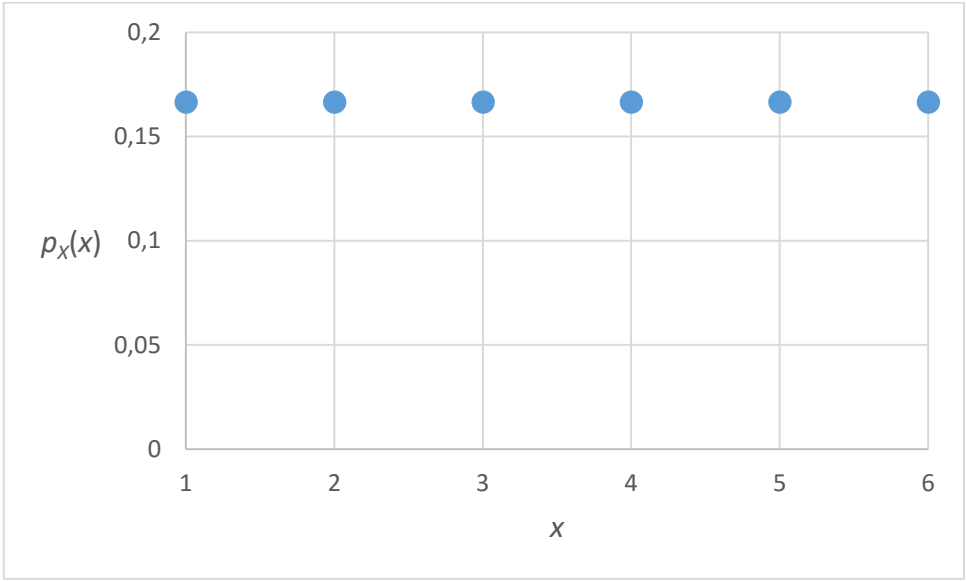
$$p_X(x_1) = P(X = x_1), \cdots, p_X(x_K) = P(X = x_K)$$

Η συνάρτηση $p_X(x)$ συμβολίζεται με μικρό γράμμα p σε αντίθεση με την πιθανότητα που συμβολίζεται με το κεφαλαίο γράμμα P. Η συνάρτηση $p_X()$ καλείται **κατανομή πιθανότητας** (**probability distribution**) της τ.μ. X.

Παράδειγμα: Στην περίπτωση του ζαριού υπάρχουν 6 ενδεχόμενα, οπότε η κατανομή είναι η παρακάτω σειρά 6 τιμών:

$$p_X(1) = P(X = 1) = \frac{1}{6}, \quad p_X(2) = P(X = 2) = \frac{1}{6}, \quad p_X(3) = P(X = 3) = \frac{1}{6},$$

$$p_X(4) = P(X = 4) = \frac{1}{6}, \quad p_X(5) = P(X = 5) = \frac{1}{6}, \quad p_X(6) = P(X = 6) = \frac{1}{6}$$



Αθροίζοντας τις πιθανότητες για όλα τα δυνατά ενδεχόμενα το αποτέλεσμα είναι πάντα ίσο με 1:

$$\sum_x p_X(x) = 1.$$

Πολλές φορές μια τ.μ. παίρνει συνεχείς τιμές. Για παράδειγμα, X: η θερμοκρασία ενός χημικού αντιδραστήρα, ή X: η τιμή της μετοχής μιας εταιρίας εισηγμένης στο χρηματιστήριο,

ή X: το βάρος ενός αυτοκινήτου, κλπ. Στην περίπτωση αυτή η κατανομή πιθανότητας $p(x)$ ορίζεται ως ο λόγος της πιθανότητα η τ.μ. να έχει τιμή μέσα στο διάστημα x έως $(x + \delta x)$ δια του εύρους του διαστήματος δx όταν αυτό είναι πολύ μικρό:

$$p_X(x) = \lim_{\delta x \to 0} \frac{P(x \leq X \leq x + \delta x)}{\delta x}$$

Στην περίπτωση αυτή η συνάρτηση $p_X(x)$ λέγεται και **συνάρτηση πυκνότητας πιθανότητας** (**probability density function – pdf**). Το ολοκλήρωμα της πυκνότητας πιθανότητας είναι ίσο με 1:

$$\int_x p_X(x)dx = 1.$$

Δεσμευμένες πιθανότητες

Συχνά ενδιαφερόμαστε για την πιθανότητα ενός ενδεχομένου όταν δεν είναι όλα τα ενδεχόμενα δυνατά, αλλά γνωρίζουμε ότι μόνο ένα υποσύνολο των ενδεχομένων μπορεί να συμβεί. Για παράδειγμα, έστω ότι επιλέγουμε τυχαία έναν άνθρωπο X από το σύνολο των Βουδιστών. Γνωρίζουμε ότι δεν είναι δυνατό το ενδεχόμενο «ο Χ είναι Χριστιανός» ούτε το ενδεχόμενο «ο Χ είναι Μουσουλμάνος». Έχουμε περιορίσει τα ενδεχόμενα στο υποσύνολο των ανθρώπων που είναι Βουδιστές. Η πιθανότητα του ενδεχομένου «ο X να είναι Ασιάτης δεδομένου ότι είναι Βουδιστής» καλείται **δεσμευμένη πιθανότητα** (**conditional probability**). Οι δεσμευμένες πιθανότητες περιορίζονται από μια συνθήκη (στην συγκεκριμένη περίπτωση ο περιορισμός είναι ότι «ο Χ είναι Βουδιστής»). Αν Α είναι το σύνολο των Ασιατών και Β είναι το σύνολο των Βουδιστών, τότε μαθηματικά η δεσμευμένη πιθανότητα

«ο X είναι Ασιάτης δεδομένου ότι ο X είναι Βουδιστής»

συμβολίζεται ως εξής:

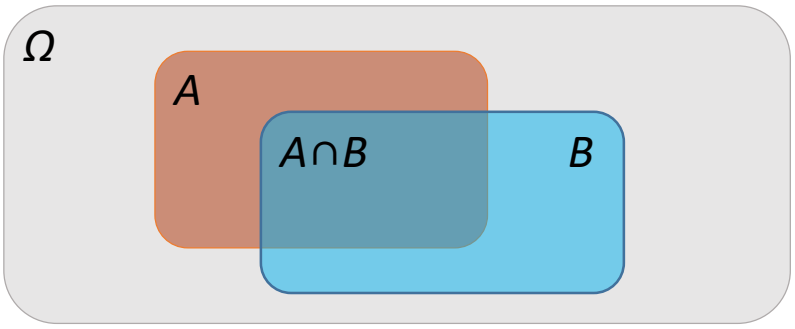
$$P(X \in A \mid X \in B)$$

Στο παρακάτω Σχήμα γίνεται αντιληπτή η διαφορά μεταξύ απλής πιθανότητας και δεσμευμένης πιθανότητας. Στο Σχήμα φαίνονται τα σύνολα A, B, Ω και η τομή $A \cap B$ μεταξύ των συνόλων A, B που περιέχει τους ανθρώπους που είναι και Ασιάτες και Βουδιστές ταυτόχρονα. Η απλή πιθανότητα υπολογίζεται ως

$$P(X \in A) = \frac{\text{Πλήθος ανθρώπων στο σύνολο } A}{\text{Πλήθος ανθρώπων στο σύνολο } \Omega}$$

ενώ η δεσμευμένη πιθανότητα υπολογίζεται ως:

$$P(X \in A \mid X \in B) = \frac{\text{Πλήθος ανθρώπων στο σύνολο } A \cap B}{\text{Πλήθος ανθρώπων στο σύνολο Β}}$$



Παράδειγμα: Με βάση το παραπάνω παράδειγμα έχουμε

- Ω = το σύνολο όλων των ανθρώπων του κόσμου
- N = το πλήθος των ανθρώπων (= πλήθος στοιχείων του Ω)
- A = το σύνολο των Ασιατών
- N_A = το πλήθος των Ασιατών (= πλήθος στοιχείων του A)
- B = το σύνολο των Βουδιστών
- N_B = το πλήθος των Βουδιστών (= πλήθος στοιχείων του B)
- $A \cap B$ = το σύνολο των ανθρώπων που είναι Ασιάτες και Βουδιστές
- $N_{A\cap B}$ = το πλήθος των Ασιατών Βουδιστών (= πλήθος στοιχείων του $A \cap B$)

Σύμφωνα με τα παραπάνω, η πιθανότητα

«ο X είναι Ασιάτης <u>δεδομένου</u> ότι ο X είναι Βουδιστής»

είναι:

$$P(X \in A \mid X \in B) = \frac{N_{A\cap B}}{N_B}$$

και η πιθανότητα

«ο X είναι Βουδιστής <u>δεδομένου</u> ότι ο X είναι Ασιάτης»

είναι:

$$P(X \in B \mid X \in A) = \frac{N_{A\cap B}}{N_A}$$

Με βάση τον ορισμό της δεσμευμένης πιθανότητας είναι εύκολο να δείξουμε ότι

$$P(A|B) = \frac{N_{A\cap B}}{N_B} = \frac{N_{A\cap B}}{N} \cdot \frac{N}{N_B}$$

Το πηλίκο $\frac{N_{A\cap B}}{N}$ είναι η **συνδυασμένη πιθανότητα** (**joint probability**)

«ο X είναι Βουδιστής <u>και</u> ο X είναι Ασιάτης»

$$P(A, B) = \frac{N_{A\cap B}}{N}$$

οπότε:

$$P(A|B) = \frac{P(A,B)}{P(B)}.$$

Στατιστική Ανεξαρτησία.

Λέμε ότι δύο τυχαίες μεταβλητές X, Y είναι **στατιστικά ανεξάρτητες** (**statistically independent**) ή απλά ανεξάρτητες, αν η κοινή τους κατανομή είναι το γινόμενο των επί μέρους κατανομών:

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$$

ή ισοδύναμα,

$$p_{X,Y}(x, y) = p_X(x) \cdot p_Y(y).$$

Αν οι μεταβλητές X, Y είναι ανεξάρτητες μεταξύ τους τότε από τους τύπους των δεσμευμένων πιθανοτήτων παίρνουμε

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{P(X = x) \cdot P(Y = y)}{P(Y = y)} = P(X = x)$$

Με άλλα λόγια η δεσμευμένη πιθανότητα $P(X = x|Y = y)$ δεν εξαρτάται από την τιμή του Y. Παρομοίως βρίσκουμε

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{P(X = x) \cdot P(Y = y)}{P(X = x)} = P(Y = y)$$

δηλ., η δεσμευμένη πιθανότητα $P(Y = y|X = x)$ δεν εξαρτάται από την τιμή του X. Για το λόγο αυτό λέμε ότι οι μεταβλητές αυτές είναι ανεξάρτητες.

Ο κανόνας του Bayes

Με βάση τα παραπάνω είναι εύκολο να δούμε ότι οι δεσμευμένες πιθανότητες $P(A|B)$ και $P(B|A)$ για οποιαδήποτε ενδεχόμενα A, B, σχετίζονται με ένα απλό τύπο. Ισχύει:

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

$$P(B|A) = \frac{P(A,B)}{P(A)}$$

οπότε διαιρώντας τις δύο παραπάνω εξισώσεις βρίσκουμε

$$P(A|B) = P(B|A)\frac{P(A)}{P(B)} \quad \textbf{(κανόνας του Bayes)}$$

Οι πιθανότητες στην αναγνώριση προτύπων

Η σημασία όλων των παραπάνω στην Μηχανική Μάθηση και συγκεκριμένα στην αναγνώριση προτύπων γίνεται κατανοητή στην περίπτωση που είναι εύκολο να υπολογίσουμε την πιθανότητα του διανύσματος χαρακτηριστικών $\mathbf{X}$ ενός αντικειμένου με δεδομένη την κλάση στην οποία ανήκει. Η πιθανότητα αυτή λέγεται **πιθανοφάνεια** (**likelihood**)

$$P(\mathbf{X}=\mathbf{x}|C_i) \quad (\text{πιθανοφάνεια})$$

Μια δεύτερη ποσότητα που χρειάζεται να εκτιμήσουμε είναι η πιθανότητα το αντικείμενο να ανήκει σε αυτή την κλάση χωρίς να ξέρουμε το διάνυσμα χαρακτηριστικών του. Η ποσότητα αυτή λέγεται **εκ των προτέρων πιθανότητα** (**a priori probability**)

$$P(C_i) \quad (\text{εκ των προτέρων πιθανότητα})$$

Αν οι δύο παραπάνω ποσότητες είναι γνωστές ή μπορούν να εύκολα να εκτιμηθούν τότε μπορούμε με τον κανόνα του Bayes να υπολογίσουμε την **εκ των υστέρων πιθανότητα** (**a posteriori probability**) να ανήκει το αντικείμενο στην κλάση C_i με δεδομένο το διάνυσμα χαρακτηριστικών του:

$$P(C_i|\mathbf{X}=\mathbf{x}) = \frac{P(\mathbf{X}=\mathbf{x}|C_i)}{P(\mathbf{X}=\mathbf{x})}P(C_i) \quad (\text{εκ των υστέρων πιθανότητα})$$

Αυτό που λείπει είναι ο παρονομαστής ο οποίος όμως υπολογίζεται εύκολα από τον τύπο

$$\begin{aligned} P(\mathbf{X}=\mathbf{x}) &= P(\mathbf{X}=\mathbf{x}, C_1) + \cdots + P(\mathbf{X}=\mathbf{x}, C_N) \\ &= P(\mathbf{X}=\mathbf{x}|C_1)P(C_1) + \cdots + P(\mathbf{X}=\mathbf{x}|C_N)P(C_N) \end{aligned}$$

όπου $C_1, \cdots, C_N$ είναι όλες οι πιθανές κλάσεις στις οποίες μπορεί να ανήκει το αντικείμενο.

Μετά από όλα αυτά, η απόφαση για την κλάση στην οποία θα ταξινομηθεί το αντικείμενο παίρνεται με βάση την κλάση που έχει την μεγαλύτερη εκ των υστέρων πιθανότητα. Με άλλα λόγια, το αντικείμενο ταξινομείται στην κλάση C_j εφόσον

$$P\big(C_j|\mathbf{X}=\mathbf{x}\big) > P(C_i|\mathbf{X}=\mathbf{x})$$

για όλες τις κλάσεις C_i $(i \neq j)$. Χρησιμοποιώντας τον κανόνα του Bayes έχουμε

$$\frac{P\big(\mathbf{X}=\mathbf{x}|C_j\big)}{P(\mathbf{X}=\mathbf{x})}P\big(C_j\big) > \frac{P(\mathbf{X}=\mathbf{x}|C_i)}{P(\mathbf{X}=\mathbf{x})}P(C_i)$$

κι επειδή το $P(\mathbf{X}=\mathbf{x})$ είναι κοινός και θετικός παρονομαστής, η παραπάνω συνθήκη απλοποιείται ως εξής:

$$P\big(\mathbf{X}=\mathbf{x}|C_j\big)P\big(C_j\big) > P(\mathbf{X}=\mathbf{x}|C_i)P(C_i)$$

Η μέθοδος ταξινόμησης Naïve Bayes

Η μέθοδος αυτή είναι ιδιαίτερα δημοφιλής λόγω της απλότητάς της αλλά και των γενικά καλών αποτελεσμάτων που δίνει χωρίς ωστόσο να είναι πάντα η μέθοδος με την καλύτερη

επίδοση. Είναι αρκετά ελκυστική σε περίπτωση προβλημάτων με πρότυπα πολύ μεγάλων διαστάσεων. Η βασική ιδέα είναι απλή:

Ας υποθέσουμε ότι τα πρότυπα εισόδου είναι διανύσματα της μορφής $\mathbf{X} = [X_1, \cdots, X_n]^T$ όπου τα χαρακτηριστικά X_i είναι τυχαίες μεταβλητές. Σύμφωνα με τη μέθοδο αυτή, υποθέτουμε ότι οι μεταβλητές αυτές είναι στατιστικά ανεξάρτητες και μπορούμε να γράψουμε κατά προσέγγιση ότι η δεσμευμένη κατανομή πιθανότητας του διανύσματος $\mathbf{X}$ αν γνωρίζουμε ότι προέρχεται από την κλάση C_j μπορεί να γραφεί ως γινόμενο των δεσμευμένων κατανομών των μεταβλητών:

$$P(\mathbf{X}|C_j) = P(X_1|C_j) \cdot \ldots \cdot P(X_n|C_j) = \prod_{i=1}^{n} P(X_i|C_j)$$

Αν και η παραπάνω υπόθεση στατιστικής ανεξαρτησίας δεν είναι γενικά ορθή, συνήθως δίνει μια καλή προσέγγιση της κατανομής ενώ ταυτόχρονα απλοποιεί πάρα πολύ τους υπολογισμούς. Συγκεκριμένα δεν έχουμε παρά να εκτιμήσουμε τις κατανομές $P(X_i|C_j)$ για κάθε ζευγάρι X_i, C_j. Επειδή οι κατανομές είναι μονοδιάστατες, το πρόβλημα είναι σχετικά απλό:

- Αν το χαρακτηριστικό X_i παίρνει διακριτές τιμές (για παράδειγμα, 1, 2, 3, 4) τότε η κατανομή μπορεί να εκτιμηθεί με ένα απλό ιστόγραμμα.
- Αν το χαρακτηριστικό X_i παίρνει συνεχείς τιμές τότε μπορεί να χρησιμοποιηθεί μια κλασική μονοδιάστατη μέθοδος εκτίμησης κατανομής

Με βάση τα παραπάνω και σύμφωνα με τον κανόνα του Bayes, η εκ των υστέρων πιθανότητα γράφεται:

$$P(C_j|\mathbf{X}) = \frac{P(\mathbf{X}|C_j)P(C_j)}{P(\mathbf{X})}$$

Αν υπάρχουν μόνο δύο κλάσεις C_0, C_1 τότε η απόφαση του ταξινομητή γίνεται με βάση τη σύγκριση των εκ των υστέρων κατανομών. Συγκεκριμένα συγκρίνουμε τον **λόγο των εκ των υστέρων πιθανοτήτων** με τη μονάδα:

$$L = \frac{P(C_1|\mathbf{X})}{P(C_0|\mathbf{X})} = \frac{P(\mathbf{X}|C_1)P(C_1)}{P(\mathbf{X}|C_0)P(C_0)}$$

Μετά από μερικές απλές πράξεις μπορούμε να βρούμε ότι

$$L = \frac{P(C_1)}{P(C_0)} \prod_{i=1}^{n} \frac{P(X_i|C_1)}{P(X_i|C_0)}$$

Η απόφαση του ταξινομητή παίρνεται ως εξής:

$$\begin{array}{ll} \text{Αν } L > 1 & \text{τότε το } \mathbf{X} \text{ ταξινομείται στην } C_1 \\ \text{Αν } L < 1 & \text{τότε το } \mathbf{X} \text{ ταξινομείται στην } C_0 \end{array}$$

Παράδειγμα:

Θέλουμε να ταξινομήσουμε ζώα σε μια από τις δύο κλάσεις C_0: άλογα, C_1: γαϊδούρια. Χρησιμοποιούμε τα χαρακτηριστικά X_1: «ύψος ζώου» και X_2: «μήκος αυτιού».

Έχουμε συλλέξει δεδομένα από 100 άλογα και 100 γαϊδούρια και βρήκαμε ότι

- η μέση τιμή και η διασπορά του ύψους του ζώου όταν πρόκειται για άλογο είναι:

$$\mu_0(X_1) = 2.2, \qquad \sigma_0(X_1) = 0.2$$

- η μέση τιμή και η διασπορά του ύψους του ζώου όταν πρόκειται για γαϊδούρι είναι:

$$\mu_1(X_1) = 1.7, \qquad \sigma_1(X_1) = 0.15$$

- η μέση τιμή και η διασπορά του μήκους των αυτιών του ζώου όταν πρόκειται για άλογο είναι:

$$\mu_0(X_2) = 0.15, \qquad \sigma_0(X_2) = 0.01$$

- η μέση τιμή και η διασπορά του μήκους των αυτιών του ζώου όταν πρόκειται για γαϊδούρι είναι:

$$\mu_1(X_2) = 0.25, \qquad \sigma_1(X_2) = 0.02$$

Υποθέτουμε ότι οι δεσμευμένες κατανομές πιθανοτήτων των χαρακτηριστικών $X_1,\ X_2$ είναι Γκαουσσιανές. Συνεπώς

$$P(X_1 = x_1|C_0) = \frac{1}{\sigma_0(X_1)\sqrt{2\pi}}\exp\left(-\frac{\left(x_1 - \mu_0(X_1)\right)^2}{2\sigma_0(X_1)^2}\right) = \frac{1}{0.2\sqrt{2\pi}}\exp\left(-\frac{(x_1 - 2.2)^2}{2\cdot(0.2)^2}\right)$$

$$P(X_1 = x_1|C_1) = \frac{1}{\sigma_1(X_1)\sqrt{2\pi}}\exp\left(-\frac{(x_1 - \mu_1(X_1))^2}{2\sigma_1(X_1)^2}\right) = \frac{1}{0.15\sqrt{2\pi}}\exp\left(-\frac{(x_1 - 1.7)^2}{2\cdot(0.15)^2}\right)$$

$$P(X_2 = x_2|C_0) = \frac{1}{\sigma_0(X_2)\sqrt{2\pi}}\exp\left(-\frac{(x_2 - \mu_0(X_2))^2}{2\sigma_0(X_2)^2}\right) = \frac{1}{0.01\sqrt{2\pi}}\exp\left(-\frac{(x_2 - 0.15)^2}{2\cdot(0.01)^2}\right)$$

$$P(X_2 = x_2|C_1) = \frac{1}{\sigma_1(X_2)\sqrt{2\pi}}\exp\left(-\frac{(x_2 - \mu_1(X_2))^2}{2\sigma_1(X_2)^2}\right) = \frac{1}{0.02\sqrt{2\pi}}\exp\left(-\frac{(x_2 - 0.25)^2}{2\cdot(0.02)^2}\right)$$

Σύμφωνα με τη μέθοδο Naïve Bayes η πιθανοφάνεια $P(\mathbf{X}|C_i)$ υπολογίζεται από το γινόμενο:

$$P(\mathbf{X} = \mathbf{x}|C_i) = P(X_1 = x_1|C_i)\cdot P(X_2 = x_2|C_i)$$

Χρειαζόμαστε επίσης τις εκ των προτέρων πιθανότητες $P(C_0), P(C_1)$. Αφού έχουμε 100 πρότυπα από κάθε κλάση κάνουμε την υπόθεση ότι οι δύο κλάσεις είναι εκ των προτέρων ισοπίθανες, δηλαδή:

$$P(C_0) = P(C_1) = 0.5$$

Έτσι, ο λόγος των εκ των υστέρων πιθανοτήτων είναι:

$$L = \frac{P(C_1)}{P(C_0)} \prod_{i=1}^{2} \frac{P(X_i|C_1)}{P(X_i|C_0)}$$
$$= \frac{P(X_1 = x_1|C_1) \cdot P(X_2 = x_2|C_1)}{P(X_1 = x_1|C_0) \cdot P(X_2 = x_2|C_0)}$$

Έτσι, για παράδειγμα, αν μας δοθεί ένα καινούργιο πρότυπο με χαρακτηριστικά:

$$X_1 = 2, \qquad X_2 = 0.18$$

τότε αντικαθιστώντας υπολογίζουμε τον λόγο L ως

$$L = \frac{0.3599 \cdot 0.0436}{1.2099 \cdot 0.4432} = 0.0293 < 1.$$

Συνεπώς αποφασίζουμε ότι το πρότυπο ανήκει στην κλάση C_0 (άλογο).
