Η GAUSSIAN ΣΥΝΑΡΤΗΣΗ ΠΥΚΝΟΤΗΤΑΣ ΠΙΘΑΝΟΤΗΤΑΣ

- Η συνάρτηση πυκνότητας πιθανότητας του αθροίσματος *ενός πλήθους στατιστικώς ανεξάρτητων τυχαίων μεταβλητών, τείνει στη Gaussian συνάρτηση πυκνότητας πιθανότητας, όταν το πλήθος των όρων του αθροίσματος τείνει στο άπειρο.*

Πολυμεταβλητή Gaussian συνάρτηση πυκνότητας πιθανότητας

$$p(x) = \frac{1}{(2\pi)^{l/2}|S|^{1/2}} \exp\left(-\frac{1}{2}(x-m)^T S^{-1}(x-m)\right)$$

m=E[x] είναι το μέσο διάνυσμα

S είναι ο πίνακας συνδιασποράς

$$S = E[(x-m)(x-m)^T]$$

|S| είναι η ορίζουσα του S

Η Gaussian καλείται και *κανονική*

Μονοδιάστατη περίπτωση

- *N(m,S)*

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)$$

- σ^2 *είναι η διασπορά*

Παράδειγμα 1ο

- Υπολογίστε την τιμή της Gaussian συνάρτησης πυκνότητας πιθανότητας, *N(m,S), για*

$$x_1 = [0.2,\ 1.3]^T \quad x_2 = [2.2,\ -1.3]^T$$

$$m = [0,\ 1]^T, \quad S = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Παράδειγμα 2ο

- Θεωρήστε πρόβλημα ταξινόμησης δύο κλάσεων στον δισδιάστατο χώρο, όπου τα δεδομένα και στις δύο κλάσεις, ακολουθούν Gaussian κατανομές, *N(m1,S1), N(m2,S2).*

$$m_1 = [1,\ 1]^T, \quad m_2 = [3,\ 3]^T, \quad S_1 = S_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

- Υποθέτοντας ότι P(ω1)=P(ω2), ταξινομήστε το διάνυσμα σε μία εκ των δύο κλάσεων.

Παράδειγμα 3ο

- Δημιουργήστε *N=500* δισδιάστατα σημεία δεδομένων που υπακούουν στην Gaussian κατανομή *N(m,S)*

$$m = [0,\ 0]^T$$

$$S = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

$$\sigma_1^2 = \sigma_2^2 = 1,\ \sigma_{12} = 0$$
$$\sigma_1^2 = \sigma_2^2 = 0.2,\ \sigma_{12} = 0$$
$$\sigma_1^2 = \sigma_2^2 = 2,\ \sigma_{12} = 0$$
$$\sigma_1^2 = 0.2,\ \sigma_2^2 = 2,\ \sigma_{12} = 0$$
$$\sigma_1^2 = 2,\ \sigma_2^2 = 0.2,\ \sigma_{12} = 0$$
$$\sigma_1^2 = \sigma_2^2 = 1,\ \sigma_{12} = 0.5$$
$$\sigma_1^2 = 0.3,\ \sigma_2^2 = 2,\ \sigma_{12} = 0.5$$
$$\sigma_1^2 = 0.3,\ \sigma_2^2 = 2,\ \sigma_{12} = -0.5$$

- Όταν οι δύο συνιστώσες του *x* είναι ασυσχέτιστες και έχουν ίδια διασπορά, τα διανύσματα δεδομένων σχηματίζουν ομάδες «σφαιρικού σχήματος»
- Όταν οι δύο συνιστώσες του *x* είναι ασυσχέτιστες και με διαφορετική διασπορά, τα διανύσματα δεδομένων σχηματίζουν ελειψοειδείς ομάδες.
- Η συντεταγμένη με την μεγαλύτερη διασπορά αντιστοιχεί στον «πρωτεύοντα άξονα» της ελλειψοειδής ομάδας, ενώ η συντεταγμένη με τη μικρότερη διασπορά αντισοιχεί στο «δευτερέυοντα άξονα».
- Επιπλέον, ο «πρωτεύων» και ο «δευτερεύων» άξονας του ελλειψοειδούς είναι παράλληλοι με τους καρτεσιανούς άξονες

- Όταν οι δύο συνιστώσες του *x* παρουσιάζουν συσχέτιση, ο «κύριος» και ο «δευτερεύων» άξονας της ελλειψοειδής ομάδας δεν είναι πλεόν παράλληλοι με τους καρτεσιανούς άξονες.
- Ο βαθμός της περιστροφής ως προς τους καρτεσιανούς άξονες εξαρτάται από την τιμή της σ_{12}
- Η συσχέτιση επιδρά ανάλογα με το αν έχει θετική ή αρνητική τιμή
- Όταν σ_{12}<>0, τα δεδομένα σχηματίζουν ελλειψοειδείς ομάδες παρά το γεγονός ότι οι διασπορές ανά συνιστώσα είναι ίδιες.

ΤΑΞΙΝΟΜΗΤΕΣ ΕΛΑΧΙΣΤΗΣ ΑΠΟΣΤΑΣΗΣ

ΤΑΞΙΝΟΜΗΤΗΣ ΕΥΚΛΕΙΔΕΙΑΣ ΑΠΟΣΤΑΣΗΣ

Ο βέλτιστος Bayesian ταξινομητής απλοποιείται σημαντικά υιοθετώντας τις παρακάτω παραδοχές:

- Οι κλάσεις είναι ισοπίθανες.
- Τα δεδομένα σε *όλες* τις κλάσεις ακολουθούν Gaussian κατανομές.
- Το μητρώο συνδιασποράς είναι το *ίδιο* για όλες τις κλάσεις.
- Το μητρώο συνδιασποράς είναι διαγώνιο και *όλα* τα στοιχεία της διαγωνίου είναι μεταξύ τους *ίσα*. Δηλαδή, $S=\sigma^2 I$, όπου I ο ταυτοτικός πίνακας.

- Προκύπτει ότι ο βέλτιστος Bayesian ταξινομητής είναι ισοδύναμος με τον ταξινομητή ελάχιστης Ευκλείδειας απόστασης.
- Δηλαδή, το άγνωστο διάνυσμα *x*, ταξινομείται στην κλάση ω_i άν

$$||x - m_i|| \equiv \sqrt{(x - m_i)^T (x - m_i)} < ||x - m_j||, \quad \forall i \neq j$$

- Είναι αξιοσημείωτο ότι, εξαιτίας της απλότητάς του, ο ταξινομητής Ευκλείδειας απόστασης χρησιμοποιείται συχνά, ακόμα και όταν γνωρίζουμε ότι οι παραπάνω υποθέσεις δεν ισχύουν.
- Ταξινομεί ένα πρότυπο στην κλάση της οποίας η μέση τιμή είναι πλησιέστερα στο πρότυπο με βάση την Ευκλείδεια νόρμα.

ΤΑΞΙΝΟΜΗΤΗΣ MAHALANOBIS ΑΠΟΣΤΑΣΗΣ

- Αν αφαιρέσουμε την υπόθεση που απαιτεί να είναι διαγώνιο το μητρώο συνδιασποράς και τα στοιχεία της διαγωνίου να είναι ίσα μεταξύ τους, τότε ο βέλτιστος Bayesian ταξινομητής γίνεται ισοδύναμος με τον ταξινομητή ελάχιστης Mahalanobis απόστασης.

$$\sqrt{(x-m_i)^T S^{-1}(x-m_i)} < \sqrt{(x-m_j)^T S^{-1}(x-m_j)}, \quad \forall j \neq i$$

όπου S είναι το κοινό μητρώο συνδιασποράς

Παράδειγμα 4ο

- Έστω πρόβλημα ταξινόμησης σε δύο κλάσεις στον τρισδιάστατο χώρο
- Να ταξινομηθεί το *x* με Ευκλίδειο και Mahalanobis ταξινομητή.

$$m_1 = [0,\ 0,\ 0]^T \qquad m_2 = [0.5,\ 0.5,\ 0.5]^T$$

$$S = \begin{pmatrix} 0.8 & 0.01 & 0.01 \\ 0.01 & 0.2 & 0.01 \\ 0.01 & 0.01 & 0.2 \end{pmatrix}$$

$$x = [0.1,\ 0.5,\ 0.1]^T$$