Αναγνώριση Προτύπων

1ο Σύνολο Αναλυτικών Ασκήσεων

9ο Εξάμηνο - Χειμερινό εξάμηνο 2019-20 - Ροή Σ

Αντωνιάδης Παναγιώτης (03115009 - `el15009@central.ntua.gr`)

"Just as electricity transformed almost everything 100 years ago, today I actually have a hard time thinking of an industry that I don't think AI will transform in the next several years."

– Andrew Ng, *Computer scientist and Statistician*

Άσκηση 1 (Probabilities)

Θεωρήστε δύο Γκαουσιανές μονοδιάστατες κατανομές $N(\mu_1, \sigma_1^2)$ και $N(\mu_2, \sigma_2^2)$. Από τις κατανομές αυτές επιλέγουμε δύο τυχαία δείγματα x_1 και x_2 , αντίστοιχα, και υπολογίζουμε το άθροισμά τους $x_3 = x_1 + x_2$. Η δειγματοληψία αυτή επαναλαμβάνεται διαρκώς.

1. **Να δείξετε ότι η προκύπτουσα κατανομή των τιμών της x_3 ακολουθεί επίσης κανονική κατανομή.**

   Γνωρίζουμε ότι τα δείγματα x_1 αντιστοιχούν σε μία κανονική τυχαία μεταβλητή $X_1 \sim N(\mu_1, \sigma_1^2)$ και τα δείγματα x_2 σε μία κανονική τυχαία μεταβλητή $X_2 \sim N(\mu_2, \sigma_2^2)$. Από την στιγμή που επιλέγουμε κάθε φορά ανεξάρτητα τα δύο τυχαία δείγματα x_1 και x_2, αυτό που θέλουμε ουσιαστικά να δείξουμε είναι ότι το άθροισμα δύο ανεξάρτητων μονοδιάστατων κανονικών τυχαίων μεταβλητών αποτελεί κανονική τυχαία μεταβλητή. Υπάρχουν διάφοροι τρόποι στην βιβλιογραφία που να αποδεκνύουν το ζητούμενο αυτό. Ένας κομψός και σύντομος τρόπος βασίζεται στις χαρακτηριστικές συναρτήσεις των κατανομών. Η χαρακτηριστική συνάρτηση μιας μοναδιάστατης κανονικής κατανομής με μέση τιμή μ και διακύμανση σ^2 είναι:

   $$\phi(t) = \exp\left(it\mu - \frac{\sigma^2 t^2}{2}\right)$$

   Άρα για τις X_1 και X_2 έχουμε:

   $$\phi_{X_1}(t) = \exp\left(it\mu_1 - \frac{\sigma_1^2 t^2}{2}\right) \quad \phi_{X_2}(t) = \exp\left(it\mu_2 - \frac{\sigma_2^2 t^2}{2}\right)$$

   Γνωρίζουμε ότι η χαρακτηριστική συνάρτηση του αθροίσματος δύο ανεξάρτητων τυχαίων μεταβλητών ισούται με το γινόμενο των επιμέρους χαρακηριστικών συναρτήσεων. Συνεπώς:

   $$\begin{aligned}\phi_{X_1+X_2}(t) &= \exp\left(it\mu_1 - \frac{\sigma_1^2 t^2}{2}\right)\exp\left(it\mu_2 - \frac{\sigma_2^2 t^2}{2}\right)\\ &= \exp\left(it(\mu_1+\mu_2) - \frac{(\sigma_1^2+\sigma_2^2) t^2}{2}\right)\end{aligned}$$

   Βλέπουμε, λοιπόν, η $\phi_{X_1+X_2}(t)$ αποτελεί την χαρακτηριστική συνάρτηση μιας κανονικής τυχαίας μεταβλητής. Άρα η η προκύπτουσα κατανομή των τιμών x_3 αποτελεί επίσης κανονική κατανομή.

2. **Ποια είναι η μέση τιμή μ_3 αυτής της κατανομής;**

   Σύμφωνα με την παραπάνω σχέση, παρατηρούμε ότι $\mu_3 = \mu_1 + \mu_2$.

3. **Ποια είναι η διασπορά σ_3^2;**

   Αντίστοιχα, έχουμε ότι $\sigma_3^2 = \sigma_1^2 + \sigma_2^2$.

4. **Επαναλάβετε τα παραπάνω βήματα για δύο πολυδιάστατες κατανομές** $N(\mu_1, \Sigma_1)$ **και** $N(\mu_2, \Sigma_2)$

   Στην περίπτωση αυτή, η χαρακτηριστικές συναρτήσεις είναι:

$$\phi_{X_1}(t) = \exp\left(it^T\mu_1 - \frac{1}{2}t^T\Sigma_1 t\right) \quad \phi_{X_2}(t) = \exp\left(it^T\mu_2 - \frac{1}{2}t^T\Sigma_2 t\right)$$

   Αντίστοιχα, αν πάρουμε την χαρακτηριστική συνάρτηση του αθροίσματος:

$$\begin{aligned}\phi_{X_1+X_2}(t) &= \exp\left(it^T\mu_1 - \frac{1}{2}t^T\Sigma_1 t\right)\exp\left(it^T\mu_2 - \frac{1}{2}t^T\Sigma_2 t\right) \\ &= \exp\left(it^T(\mu_1+\mu_2) - \frac{1}{2}t^T(\Sigma_1+\Sigma_2)t\right)\end{aligned}$$

   Άρα, το άθροισμα ακολουθεί κανονική κατανομή με μέση τιμή $\mu_3 = \mu_1 + \mu_2$ και πίνακα διακύμανσης $\Sigma_3 = \Sigma_1 + \Sigma_2$.

Άσκηση 2 (Bayes Decision Theory)

Θεωρήστε ότι οι υπό συνθήκη κατανομές για ένα μονοδιάστατο πρόβλημα δύο κατηγοριών ω_1 και ω_2 δίνεται από την ακόλουθη κατανομή Cauchy:

$$p(x|\omega_i) = \frac{1}{\pi b}\cdot\frac{1}{1+\left(\frac{x-\alpha_i}{b}\right)^2},\ i = 1, 2$$

1. **Να ελέγξετε ότι η παραπάνω κατανομή είναι κανονικοποιημένη σωστά.**

   Έχουμε ότι:

$$\begin{aligned}\int_{-\infty}^{+\infty} p(x|\omega_i)dx &= \int_{-\infty}^{+\infty}\frac{1}{\pi b}\cdot\frac{1}{1+\left(\frac{x-\alpha_i}{b}\right)^2}dx \\ &= \frac{1}{\pi b}\int_{-\infty}^{+\infty}\frac{1}{1+\left(\frac{x-\alpha_i}{b}\right)^2}dx\end{aligned}$$

   Θέτω $u = \frac{x-\alpha_i}{b}$ με $du = \frac{dx}{b} \implies dx = b\ du$ και η η παραπάνω σχέση γίνεται:

$$\begin{aligned}\int_{-\infty}^{+\infty} p(x|\omega_i)dx &= \frac{1}{\pi b}\int_{-\infty}^{+\infty}\frac{1}{1+u^2}b\ du \\ &= \frac{1}{\pi}\left(tan^{-1}(+\infty) - tan^{-1}(-\infty)\right) \\ &= \frac{1}{\pi}\pi \\ &= 1\end{aligned}$$

Συνεπώς, η παραπάνω κατανομή είναι κανονικοποιημένη σωστά.

2. **Εάν υποθέσουμε ότι $P(\omega_1) = P(\omega_2)$, να δείξετε ότι $P(\omega_1|x) = P(\omega_2|x)$ αν $x = \frac{a_1+a_2}{2}$. Με άλλα λόγια, ότι η διαχωριστική γραμμή απόφασης που οδηγεί στο ελάχιστο σφάλμα είναι το σημείο στη μέση των θέσεων μεγίστου των δύο κατανομών, ανεξαρτήτως του b.**

$$\begin{aligned}
P(\omega_1|x) = P(\omega_2|x) &\implies p(x|\omega_1)P(\omega_1) = P(x|\omega_2)P(\omega_2) \\
&\implies p(x|\omega_1) = P(x|\omega_2) \\
&\implies \frac{1}{\pi b} \cdot \frac{1}{1+\left(\frac{x-\alpha_1}{b}\right)^2} = \frac{1}{\pi b} \cdot \frac{1}{1+\left(\frac{x-\alpha_2}{b}\right)^2} \\
&\implies \frac{1}{1+\left(\frac{x-\alpha_1}{b}\right)^2} = \frac{1}{1+\left(\frac{x-\alpha_2}{b}\right)^2} \\
&\implies 1+\left(\frac{x-\alpha_2}{b}\right)^2 = 1+\left(\frac{x-\alpha_1}{b}\right)^2 \\
&\implies \left(\frac{x-\alpha_2}{b}\right)^2 = \left(\frac{x-\alpha_1}{b}\right)^2 \\
&\implies \begin{cases} x-\alpha_2 = x-\alpha_1 \implies \alpha_2 = \alpha_1 \\ x-\alpha_2 = -x+\alpha_1 \implies x = \frac{\alpha_1+\alpha_2}{2} \end{cases}
\end{aligned}$$

3. **Να δείξετε ότι η ελάχιστη πιθανότητα σφάλματος δίνεται από τη σχέση:**

$$P(error) = \frac{1}{2} - \frac{1}{\pi} tan^{-1} \left| \frac{\alpha_1 - \alpha_2}{2b} \right|$$

Έχουμε ότι:

$$\begin{aligned}
P(error) &= \int_{-\infty}^{+\infty} P(error|x)p(x)dx \\
&= \int_{-\infty}^{+\infty} \min[P(\omega_1|x), P(\omega_2|x)]p(x)dx \\
&= \int_{-\infty}^{+\infty} \min[P(x|\omega_1)P(\omega_1), P(x|\omega_2)P(\omega_2)]dx
\end{aligned}$$

Χρησιμοποιώντας το παραπάνω ερώτημα, γνωρίζουμε ότι έχουμε ελάχιστη πιθανότητα σφάλματος όταν το σύνορο απόφασης είναι $x = \frac{a_1+a_2}{2}$. Άρα:

$$
\begin{aligned}
P(error) &= \int_{-\infty}^{\frac{a_1+a_2}{2}} P(x|\omega_2)P(\omega_2)dx + \int_{\frac{a_1+a_2}{2}}^{+\infty} P(x|\omega_1)P(\omega_1)dx \\
&= \int_{-\infty}^{\frac{a_1+a_2}{2}} \frac{1}{\pi b}\cdot\frac{1}{1+\left(\frac{x-\alpha_2}{b}\right)^2}P(\omega_2)dx + \int_{\frac{a_1+a_2}{2}}^{+\infty} \frac{1}{\pi b}\cdot\frac{1}{1+\left(\frac{x-\alpha_1}{b}\right)^2}P(\omega_1)dx \\
&= \frac{1}{\pi b}\left(\int_{-\infty}^{\frac{a_1+a_2}{2}} \frac{1}{1+\left(\frac{x-\alpha_2}{b}\right)^2}P(\omega_2)dx + \int_{\frac{a_1+a_2}{2}}^{+\infty} \frac{1}{1+\left(\frac{x-\alpha_1}{b}\right)^2}P(\omega_1)dx\right)
\end{aligned}
$$

Θέτουμε στο 1ο ολοκλήρωμα $u = \frac{x-\alpha_1}{b}$ και στο 2ο ολοκλήρωμα $z = \frac{x-\alpha_2}{b}$ και έχουμε:

$$
\begin{aligned}
P(error) &= \frac{1}{\pi b}\left(\int_{-\infty}^{\frac{a_2-a_1}{2b}} \frac{1}{1+(u)^2}P(\omega_2)bdu + \int_{\frac{a_1-a_2}{2b}}^{+\infty} \frac{1}{1+(z)^2}P(\omega_1)bdz\right) \\
&= \frac{1}{\pi}\left(\left(\tan^{-1}\left(\frac{a_2-a_1}{2b}\right)+\frac{\pi}{2}\right)P(\omega_2) + \left(\frac{\pi}{2}-\tan^{-1}\left(\frac{a_1-a_2}{2b}\right)\right)P(\omega_1)\right) \\
&= \frac{1}{\pi}\left(\left(-\tan^{-1}\left|\frac{a_1-a_2}{2b}\right|+\frac{\pi}{2}\right)P(\omega_2) + \left(\frac{\pi}{2}-\tan^{-1}\left|\frac{a_1-a_2}{2b}\right|\right)P(\omega_1)\right) \\
&= \frac{1}{\pi}\left(\left(-\tan^{-1}\left|\frac{a_1-a_2}{2b}\right|+\frac{\pi}{2}\right)(1-P(\omega_1)) + \left(\frac{\pi}{2}-\tan^{-1}\left|\frac{a_1-a_2}{2b}\right|\right)P(\omega_1)\right) \\
&= \frac{1}{\pi}\left(-\tan^{-1}\left|\frac{a_1-a_2}{2b}\right|+\frac{\pi}{2}\right) \\
&= \frac{1}{2}-\frac{1}{\pi}\tan^{-1}\left|\frac{a_1-a_2}{2b}\right|
\end{aligned}
$$

Άσκηση 3 (Bayes Decision Theory)

Ας υποθέσουμε ότι $p(x|\omega_i) \sim N(\mu_i, \Sigma)$ για ένα πρόβλημα δύο κατηγοριών ω_1 και ω_2 και d διαστάσεων με τους ίδιους πίνακες συνδιασπορών, αλλά διαφορετικά διανύσματα για τις μέσες τιμές και διαφορετικές εκ των προτέρων πιθανότητες. Θεωρήστε την υψωμένη στο τετράγωνο απόσταση Mahalanobis:

$$r_i^2 = (x-\mu_i)^T\Sigma^{-1}(x-\mu_i)$$

1. **Να δείξετε ότι το gradient της r_i^2 δίνεται από τη σχέση:**

$$\nabla r_i^2 = 2\Sigma^{-1}(x-\mu_i)$$

Αν αναπτύξουμε την αρχική σχέση έχουμε:

$$(x-\mu_i)^T\Sigma^{-1}(x-\mu_i) = \begin{bmatrix} x_1-\mu_{i1} & x_2-\mu_{i2} & . & . & x_d-\mu_{id} \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} & . & . & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & . & . & \sigma_{2d} \\ . & & & & \\ . & & & & \\ \sigma_{d1} & \sigma_{d2} & . & . & \sigma_{dd} \end{bmatrix} \begin{bmatrix} x_1-\mu_{i1} \\ x_2-\mu_{i2} \\ . \\ . \\ x_d-\mu_{id} \end{bmatrix}$$

$$= \begin{bmatrix} (x_1-\mu_{i1})\sigma_{11}+...+(x_d-\mu_{id})\sigma_{d1} & ... & (x_1-\mu_{i1})\sigma_{1d}+...+(x_d-\mu_{id})\sigma_{dd} \end{bmatrix} \begin{bmatrix} x_1-\mu_{i1} \\ x_2-\mu_{i2} \\ . \\ . \\ x_d-\mu_{id} \end{bmatrix}$$

$$= (x_1-\mu_{i1})^2\sigma_{11}+...+(x_d-\mu_{id})(x_1-\mu_{i1})\sigma_{d1}+...+(x_1-\mu_{i1})(x_d-\mu_{id})\sigma_{1d}+...+(x_d-\mu_{id})^2\sigma_{dd}$$

Αν πάρουμε τις μερικές παραγώγους έχουμε:

$$\begin{aligned} \frac{\partial r_i^2}{\partial x_1} &= 2(x_1-\mu_{i1})\sigma_{11}+...+(x_d-\mu_{id})\sigma_{d1}+...+(x_d-\mu_{id})\sigma_{1d} \\ &= 2(x_1-\mu_{i1})\sigma_{11}+...+(x_d-\mu_{id})\sigma_{1d}+...+(x_d-\mu_{id})\sigma_{1d} \\ &= 2(x_1-\mu_{i1})\sigma_{11}+...+2(x_d-\mu_{id})\sigma_{1d} \\ &= 2\left((x_1-\mu_{i1})\sigma_{11}+...+(x_d-\mu_{id})\sigma_{1d}\right) \\ &= 2\begin{bmatrix} \sigma_{11} & ... & \sigma_{1d} \end{bmatrix} \begin{bmatrix} x_1-\mu_{i1} \\ x_2-\mu_{i2} \\ . \\ . \\ x_d-\mu_{id} \end{bmatrix} \end{aligned}$$

όπου στο τέλος χρησιμοποιήσαμε το γεγονός ότι αφού Σ συμμετρικός τότε και Σ^{-1} συμμετρικός άρα $\sigma_{ij} = \sigma ji \quad \forall i,j \in [1,d]$.

Επαναλαμβάνοντας την ίδια διαδικασία για τις υπόλοιπες μερικές παραγώγους καταλήγουμε στο ζητούμενο:

$$\nabla r_i^2 = \begin{bmatrix} \frac{\partial r_i^2}{\partial x_1} \\ \frac{\partial r_i^2}{\partial x_2} \\ . \\ . \\ \frac{\partial r_i^2}{\partial x_d} \end{bmatrix} = 2\begin{bmatrix} \sigma_{11} & \sigma_{12} & . & . & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & . & . & \sigma_{2d} \\ . & & & & \\ . & & & & \\ \sigma_{d1} & \sigma_{d2} & . & . & \sigma_{dd} \end{bmatrix} \begin{bmatrix} x_1-\mu_{i1} \\ x_2-\mu_{i2} \\ . \\ . \\ x_d-\mu_{id} \end{bmatrix} = 2\Sigma^{-1}(x-\mu_i)$$

2. **Να δείξετε ότι σε οποιοδήποτε σημείο μιας δοσμένης ευθείας που περνάει από το μ_i, το gradient ∇r_i^2 δείχνει πάντα στην ίδια διεύθυνση. Πρέπει αυτή η διεύθυνση να είναι παράλληλη με τη δοσμένη ευθεία;**

Αν r η ευθεία που περνάει από το μ_i, τότε $r = \mu_i + \lambda(b-\mu_i)$ όπου b ένα άλλο σημείο της ευθείας. Αντικαθιστώντας το r στην εξίσωσή που βρήκαμε για το gradient:

$$\begin{aligned}\nabla r_i^2 &= 2\Sigma^{-1}(r-\mu_i)\\ &= 2\Sigma^{-1}(\mu_i+\lambda(b-\mu_i)-\mu_i)\\ &= 2\Sigma^{-1}(\lambda(b-\mu_i))\\ &= 2\lambda\Sigma^{-1}(b-\mu_i)\end{aligned}$$

Συνεπώς, όπως φαίνεται στην παραπάνω εξίσωση, το gradient ∇r_i^2 έχει πάντα την ίδια διεύθυνση για κάθε σημείο της ευθείας r ενώ το μέτρο του μεταβάλλεται ανάλογα με τον πραγματικό αριθμό λ.

3. **Να δείξετε ότι τα ∇r_1^2 και ∇r_2^2 δείχνουν σε αντίθετες κατευθύνσεις κατά μήκος της γραμμής που συνδέει το μ_1 με το μ_2.**

Η γραμμή που συνδέει το μ_1 με το μ_2 έχει εξίσωση $r=\mu_1+\lambda(\mu_2-\mu_1)$. Συνεπώς κατα μήκος αυτής της γραμμής έχουμε:

$$\begin{aligned}\nabla r_1^2 &= 2\Sigma^{-1}(r-\mu_1)\\ &= 2\Sigma^{-1}(\mu_1+\lambda(\mu_2-\mu_1)-\mu_1)\\ &= 2\lambda\Sigma^{-1}(\mu_2-\mu_1)\end{aligned}$$

$$\begin{aligned}\nabla r_2^2 &= 2\Sigma^{-1}(r-\mu_2)\\ &= 2\Sigma^{-1}(\mu_1+\lambda(\mu_2-\mu_1)-\mu_2)\\ &= 2\Sigma^{-1}((\lambda-1)\mu_2-(\lambda-1)\mu_1)\\ &= 2(\lambda-1)\Sigma^{-1}(\mu_2-\mu_1)\end{aligned}$$

4. **Να δείξετε ότι το βέλτιστο υπερεπίπεδο διαχωρισμού είναι εφαπτόμενο στα υπερελλειψοειδή σταθερής πυκνότητας πιθανότητας, στο σημείο όπου το υπερεπίπεδο διαχωρισμού τέμνει την ευθεία που συνδέει το μ_1 με το μ_2.**

Το βέλτιστο επίπεδο διαχωρισμού προκύπτει από την εξίσωση $P(\omega_1|x)=P(\omega_2|x)$, η οποία αναλύεται ως εξής:

$$\begin{aligned}&P(\omega_1|x)=P(\omega_2|x) \implies P(x|\omega_1)P(\omega_1)=P(x|\omega_2)P(\omega_2) \implies\\ &\frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma(x-\mu_1)\right)P(\omega_1)=\frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu_2)^T\Sigma(x-\mu_1)\right)P(\omega_2) \implies\\ &\exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma(x-\mu_1)\right)P(\omega_1)=\exp\left(-\frac{1}{2}(x-\mu_2)^T\Sigma(x-\mu_1)\right)P(\omega_2)\end{aligned}$$

Αντικαθιστούμε όπου x την εξίσωση της ευθείας που συνδέει τα μ_1 και μ_2, δηλαδή $x=\mu_1+\lambda(\mu_2-\mu_1)$:

$$\exp\left(-\frac{1}{2}(\mu_1+\lambda(\mu_2-\mu_1)-\mu_1)^T\Sigma(\mu_1+\lambda(\mu_2-\mu_1)-\mu_1)\right)P(\omega_1)=$$
$$\exp\left(-\frac{1}{2}(\mu_1+\lambda(\mu_2-\mu_1)-\mu_2)^T\Sigma(\mu_1+\lambda(\mu_2-\mu_1)-\mu_1)\right)P(\omega_2) \implies$$

$$\exp\left(-\frac{1}{2}\lambda(\mu_2-\mu_1)^T\Sigma\lambda(\mu_2-\mu_1)\right)P(\omega_1)=\exp\left(-\frac{1}{2}(\lambda-1)(\mu_2-\mu_1)^T\Sigma(\lambda-1)(\mu_2-\mu_1)\right)P(\omega_2) \implies$$
$$\exp\left(-\frac{1}{2}\lambda^2(\mu_2-\mu_1)^T\Sigma(\mu_2-\mu_1)\right)P(\omega_1)=\exp\left(-\frac{1}{2}(\lambda-1)^2(\mu_2-\mu_1)^T\Sigma(\mu_2-\mu_1)\right)P(\omega_2) \implies$$
$$\frac{\exp\left(-\frac{1}{2}\lambda^2(\mu_2-\mu_1)^T\Sigma(\mu_2-\mu_1)\right)}{\exp\left(-\frac{1}{2}(\lambda-1)^2(\mu_2-\mu_1)^T\Sigma(\mu_2-\mu_1)\right)}=\frac{P(\omega_2)}{P(\omega_1)} \implies$$
$$\exp\left(-\frac{1}{2}\lambda^2(\mu_2-\mu_1)^T\Sigma(\mu_2-\mu_1)+\frac{1}{2}(\lambda-1)^2(\mu_2-\mu_1)^T\Sigma(\mu_2-\mu_1)\right)=\frac{P(\omega_2)}{P(\omega_1)} \implies$$
$$-\frac{1}{2}\lambda^2(\mu_2-\mu_1)^T\Sigma(\mu_2-\mu_1)+\frac{1}{2}(\lambda-1)^2(\mu_2-\mu_1)^T\Sigma(\mu_2-\mu_1)=\ln\frac{P(\omega_2)}{P(\omega_1)} \implies$$
$$(\mu_2-\mu_1)^T\Sigma(\mu_2-\mu_1)=\frac{2\ln\frac{P(\omega_2)}{P(\omega_1)}}{(\lambda-1)^2-\lambda^2}=\frac{2\ln\frac{P(\omega_2)}{P(\omega_1)}}{1-2\lambda}$$

Η τελευταία εξίσωση αναπαριστάνει ένα υπερελλειψοειδές.

5. **Σωστό ή Λάθος:** Για ένα πρόβλημα δύο κατηγοριών με Γκαουσιανές κατανομές που έχουν διαφορετικές μέσες τιμές και πίνακες συνδιασπορών, και με $P(\omega_1)=P(\omega_2)=\frac{1}{2}$, το διαχωριστικό κριτήριο απόφασης Bayes αποτελείται από το σύνολο των σημείων ίσης απόστασης Mahalanobis από τους αντίστοιχους διανυσματικούς μέσους. Να δικαιολογήσετε την απάντησή σας αναλυτικά.

   Λάθος, γιατί:

$$P(\omega_1|x)=P(\omega_2|x) \implies P(x|\omega_1)P(\omega_1)=P(x|\omega_2)P(\omega_2) \implies$$
$$\frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma_1|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma_1(x-\mu_1)\right)P(\omega_1)=\frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma_2|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu_2)^T\Sigma_2(x-\mu_1)\right)P(\omega_2) \implies$$
$$\frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma_1|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma_1(x-\mu_1)\right)=\frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma_2|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu_2)^T\Sigma_2(x-\mu_1)\right) \implies$$
$$\frac{\exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma_1(x-\mu_1)\right)}{\exp\left(-\frac{1}{2}(x-\mu_2)^T\Sigma_2(x-\mu_1)\right)}=\frac{|\Sigma_1|^{\frac{1}{2}}}{|\Sigma_2|^{\frac{1}{2}}} \implies$$
$$\exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma_1(x-\mu_1)+\frac{1}{2}(x-\mu_2)^T\Sigma_2(x-\mu_1)\right)=\frac{|\Sigma_1|^{\frac{1}{2}}}{|\Sigma_2|^{\frac{1}{2}}} \implies$$
$$-\frac{1}{2}(x-\mu_1)^T\Sigma_1(x-\mu_1)+\frac{1}{2}(x-\mu_2)^T\Sigma_2(x-\mu_1)=\ln\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right)^{\frac{1}{2}} \implies$$
$$-(x-\mu_1)^T\Sigma_1(x-\mu_1)+(x-\mu_2)^T\Sigma_2(x-\mu_1)=\ln\frac{|\Sigma_1|}{|\Sigma_2|}$$

Συνεπώς, βλέπουμε ότι το διαχωριστικό κριτήριο απόφασης αποτελείται από το σύνολο των σημείων που η διαφορά των αποστάσεων Mahalanobis από τους αντίστοιχους διανυσματικούς μέσους είναι ίση με $\ln \frac{|\Sigma_1|}{|\Sigma_2|}$.

Άσκηση 4 (Maximum Likelihood estimation)

Εστω ότι η μεταβλητή x ακολουθεί μια ομοιόμορφη κατανομή:

$$p(x|\theta) \sim U(0,\theta) = \begin{cases} \frac{1}{\theta} & 0 \leq x \leq \theta \\ 0 & otherwise \end{cases}$$

1. **Υποθέστε ότι n δείγματα $D = \{x_1, x_2, ..., x_n\}$ επιλέγονται ανεξάρτητα μεταξύ τους σύμφωνα με την κατανομή $p(x|\theta)$. Να δείξετε ότι η εκτίμηση μέγιστης πιθανοφάνειας για τη θ είναι $max_i x_i$, δηλαδή η τιμή του μέγιστου στοιχείου του συνόλου D.**

   Θέλουμε να μεγιστοποιήσουμε την ποσότητα:

   $$p(D|\theta) = \prod_{k=1}^{n} p(x_k|\theta) = \begin{cases} \frac{1}{\theta^n}, & if \ \max_i x_i \leq \theta \\ 0, & otherwise \end{cases}$$

   Για $\max_i x_i < \theta$, αν πάρουμε τον λογάριθμο της παραπάνω ποσότητας:

   $$\ln p(D|\theta) = -n \ln \theta, \quad \max_i x_i \leq \theta$$

   Τέλος, παραγωγίζουμε ως προς την μεταβλητή θ:

   $$\frac{\partial \ln p(D|\theta)}{\partial \theta} = -\frac{n}{\theta}, \quad \max_i x_i \leq \theta$$

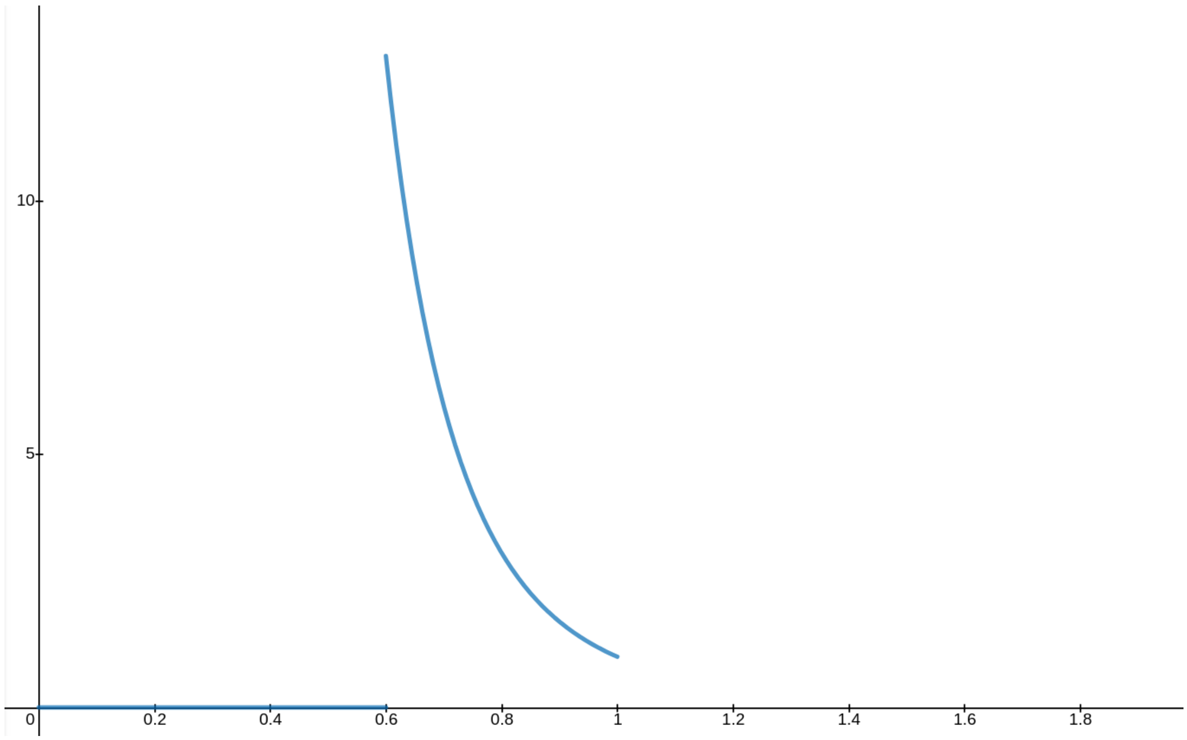
   Εδώ, δεν μπορούμε να πετύχουμε μηδενισμό της παραγώγου, οπότε εντοπίζουμε το μέγιστο από την μονοτονία στο πεδίο ορισμού. Βλέπουμε, δηλαδή, ότι καθώς το θ αυξάνεται η log-likelihood function μειώνεται. Συνεπώς, η μεγιστοποίησή της συμβαίνει για το ελάχιστο θ που μπορούμε να έχουμε, δηλαδή για $\theta = \max_i x_i$

2. **Υποθέστε ότι επιλέγονται n = 5 δείγματα από την κατανομή και ότι η μέγιστη τιμή αυτών ισούται με $max_i x_i = 0.6$. Να σχεδιάσετε την πιθανοφάνεια $p(D|\theta)$ στο εύρος τιμών $0 \leq \theta \leq 1$. Να εξηγήσετε γιατί δεν απαιτείται να γνωρίζουμε τις τιμές των υπόλοιπων τεσσάρων σημείων.**

   Για $n = 5$, η πιθανοφάνεια γίνεται:

   $$p(D|\theta) = \prod_{k=1}^{5} p(x_k|\theta) = \begin{cases} \frac{1}{\theta^5}, & 0.6 \leq \theta \leq 1 \\ 0, & 0 \leq \theta < 0.6 \end{cases}$$

Παρακάτω, βλέπουμε την γραφική παράσταστασηση της πιθανοφάνειας στο ζητούμενο εύρος τιμών:



Δεν απαιτείται να γνωρίζουμε τις τιμές των υπόλοιπων τεσσάρων σημείων, καθώς όπως βλέπουμε και στην γραφική παράσταση η μέγιστη τιμή της πιθανοφάνειας προκύπτει στο $max_i x_i = 0.6$.

Άσκηση 5 (k-Nearest Neighbors)

Θεωρήστε ότι το $D = x_1, x_2, ..., x_n$ είναι ένα σύνολο από n ανεξάρτητα και επισημειωμένα δείγματα και ότι το $D_k(x) = x_1^{'}, x_2^{'}, ..., x_k^{'}$ περιλαμβάνει τους k κοντινότερους γείτονες του x. Ο κανόνας για την ταξινόμηση του x σύμφωνα με τους k κοντινότερους γείτονες είναι να ταξινομηθεί το x στην κατηγορία που "εκπροσωπείται" περισσότερο στο $D_k(x)$. Θεωρήστε επίσης ότι μελετάμε ένα πρόβλημα δύο κατηγοριών με $P(\omega_1) = P(\omega_2) = \frac{1}{2}$ και ότι οι υπό συνθήκη πυκνότητες πιθανότητας $p(x|\omega_i)$ είναι ομοιόμορφες εντός μοναδιαίων υπερσφαιρών με απόσταση δέκα μονάδων μεταξύ τους.

1. **Να δείξετε ότι αν το k είναι περιττός αριθμός, τότε η μέση πιθανότητα λάθους δίνεται από τη σχέση:**

$$P_n(e) = \frac{1}{2^n} \sum_{j=0}^{\frac{k-1}{2}} \binom{n}{j}$$

2. **Να δείξετε ότι για αυτή την περίπτωση ο κανόνας του ενός (k = 1) κοντινότερου γείτονα παρουσιάζει μικρότερο ρυθμό σφαλμάτων σε σχέση με τον κανόνα των k κοντινότερων γειτόνων με $k > 1$.**

3. **Εάν το k επιτρέπεται να αυξάνει καθώς αυξάνεται και το n, αλλά περιορίζεται από τη σχέση $k < a\sqrt{n}$, να δείξετε ότι $P_n(e) \to 0$ καθώς το $n \to \infty$.**

Άσκηση 6 (Perceptrons)

Μας δίνονται 10 διανύσματα χαρακτηριστικών που προέρχονται από δύο κλάσεις ω_1 και ω_2:

$$\omega_1 : [-1, 4]^T, [1, 2]^T, [2, -2]^T, [1, -4]^T, [4, -1]^T$$
$$\omega_2 : [-4, 2]^T, [-2, 1]^T, [-2, -1]^T, [-1, -3]^T, [-1, -6]^T$$

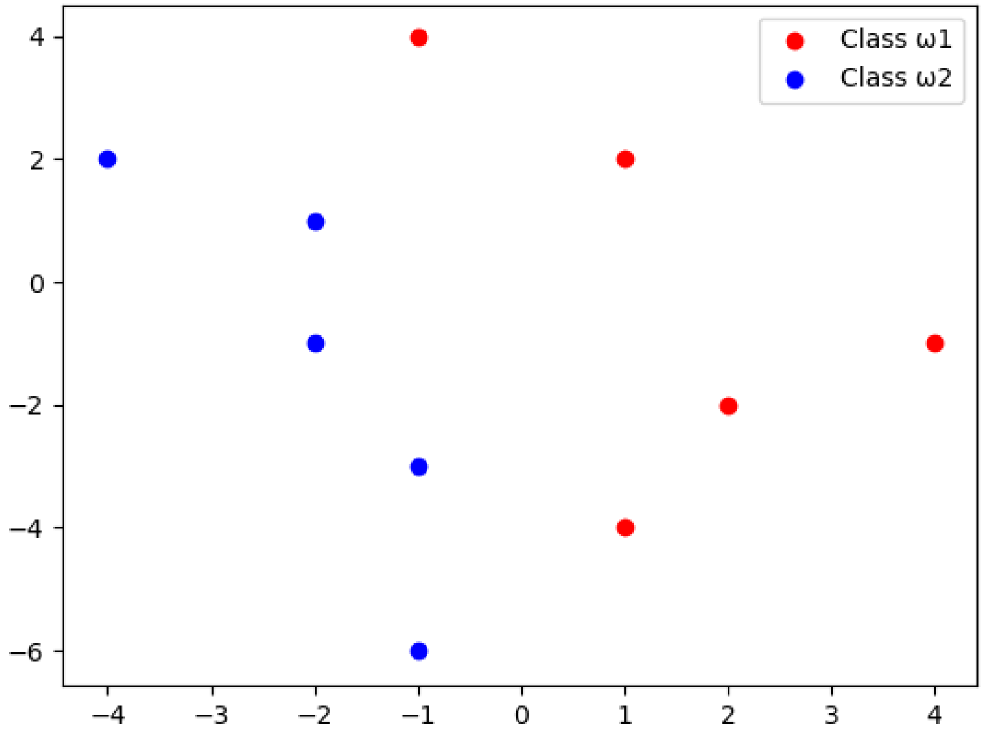
Αρχικά, ελέγξτε εάν οι δύο κλάσεις είναι γραμμικά διαχωρίσιμες, μέσω του σχεδιασμού των παραπάνω σημείων σε ένα γράφημα. Στη συνέχεια, χρησιμοποιήστε τον ακόλουθο αλγόριθμο εκπαίδευσης ενός perceptron με $\rho = 1$ και $w(0) = [0, 0]^T$, για να σχεδιάσετε μία ευθεία γραμμή που να διαχωρίζει τις δύο κλάσεις. Εάν το συγκεκριμένο διάνυσμα βαρών w δεν επαρκεί (να σχολιαστεί με μία πρόταση το γιατί), τότε να χρησιμοποιηθεί ως αρχικό διάνυσμα βαρών το $w(0) = [0, 0, 0]^T$, κάνοντας την κατάλληλη επαύξηση ταυτόχρονα και στα διανύσματα χαρακτηριστικών. Τέλος, να δοθεί με μορφή εξίσωσης η διαχωριστική καμπύλη που αντιστοιχεί στο υπολογισθέν διάνυσμα βαρών.

Αλγόριθμος Perceptron: Έστω $w(t)$ η εκτίμηση του διανύσματος βάρους και x_t το αντίστοιχο διάνυσμα χαρακτηριστικών στο t-οστό βήμα επανάληψης. Ο αλγόριθμος έχει ως εξής:

$$w(t+1) = w(t) + \rho x_t \quad \text{αν} \quad x_t \in \omega_1 \quad \text{και} \quad w(t)^T x_t \leq 0$$
$$w(t+1) = w(t) - \rho x_t \quad \text{αν} \quad x_t \in \omega_2 \quad \text{και} \quad w(t)^T x_t \geq 0$$

Ο ανωτέρω αλγόριθμος έχει την μορφή αλγορίθμων τύπου reward and punishment. Δηλαδή, αν το τωρινό δείγμα εκπαίδευσης ταξινομηθεί σωστά, τότε δεν γίνεται τίποτα (reward = no action). Αλλιώς, αν το δείγμα δεν ταξινομηθεί σωστά, η τιμή του διανύσματος βάρους μεταβάλλεται προσθέτοντας (αφαιρώντας) μία τιμή ανάλογη του x_t (punishment = correction cost).

1. **Έλεγχος αν οι κλάσεις είναι γραμμικά διαχωρίσιμες**



Παρατηρούμε ότι οι δύο κλάσεις ω_1 και ω_2 είναι γραμμικά διαχωρίσιμες.

2. **Εκπαίδευση perceptron σε 2 διαστάσεις**

   Στον παρακάτω πίνακα συνοψίζεται η εκτέλεση του αλγορίθμου perceptron πάνω στα δεδομένα μας με $\rho = 1$ και $w(0) = [0, 0]^T$.

w(t)	$\boldsymbol{x_t}$	$\boldsymbol{w(t)^T x_t}$	**w(t+1)**
0, 0	-1, 4	0	-1, 4
-1, 4	1, 2	7	-1, 4
-1, 4	2, -2	-10	1, 2
1, 2	1, -4	-7	2, -2
2, -2	4, -1	10	2, -2
2, -2	-4, 2	-12	2, -2
2, -2	-2, 1	-6	2, -2
2, -2	-2, -1	-2	2, -2
2, -2	-1, -3	4	3, 1
3, 1	-1, -6	-9	3, 1
3, 1	-1, 4	1	3, 1
3, 1	1, 2	5	3, 1
3, 1	2, -2	4	3, 1
3, 1	1, -4	-1	4, -3
4, -3	4, -1	19	4, -3
4, -3	-4, 2	-22	4, -3
4, -3	-2, 1	-11	4, -3
4, -3	-1, -3	5	5, 0
5, 0	-1, -6	-5	5, 0
5, 0	-1, 4	-5	4, 4
4, 4	1, 2	12	4, 4
4, 4	2, -2	0	6, 2
6, 2	1, -4	-2	7, -2
7, -2	4, -1	30	7, -2
7, -2	-4, 2	-32	7, -2
7, -2	-2, 1	-16	7, -2
7, -2	-2, -1	-12	7, -2
7, -2	-1, -3	-1	7, -2
7, -2	-1, -6	5	8, 4
8, 4	-1, 4	8	8, 4
8, 4	1, 2	16	8, 4
8, 4	2, -2	8	8, 4
8, 4	1, -4	-8	9, 0
9, 0	4, -1	36	9, 0
9, 0	-4, 2	-36	9, 0
9, 0	-2, 1	-18	9, 0
9, 0	-2, -1	-18	9, 0
9, 0	-1, -3	-9	9, 0
9, 0	-1, -6	-9	9, 0
9, 0	-1, 4	-9	8, 4

Παρατηρούμε ότι ο αλγόριθμος δεν θα τερματίσει ποτέ, καθώς εναλλάσονται συνεχώς οι ίδιες τιμές. Συνεπώς, θα αυξήσουμε την διάσταση του προβλήματος κατά 1, ώστε να μπορέσουμε να βρούμε ένα επίπεδο που να διαχωρίζει τα δεδομένα μας.

3. **Εκπαίδευση perceptron σε 3 διαστάσεις**

   Το αρχικό διάνυσμα βαρώς γίνεται $w(0) = [0, 0, 0]^T$ και η επαύξηση των διανυσμάτων των χαρακτηριστικών γίνεται ως εξής:

   Αν πριν το διάνυσμα ήταν $[x, y]$ τώρα θα είναι $[x, y, xy]$. Συνεπώς, τα νέα διανύσματα είναι:

$$\omega_1 : [-1, 4, -4]^T, [1, 2, 2]^T, [2, -2, -4]^T, [1, -4, -4]^T, [4, -1, -4]^T$$
$$\omega_2 : [-4, 2, -8]^T, [-2, 1, -2]^T, [-2, -1, 2]^T, [-1, -3, 3]^T, [-1, -6, 6]^T$$

$\mathbf{w(t)}$	x_t	$\mathbf{w(t+1)}$
0, 0, 0	-1, 4, -4	-1, 4, -4
-1, 4, -4	1, 2, 2	0, 6, -2
0, 6, -2	2, -2, -4	2, 4, -6
2, 4, -6	1, -4, -4	2, 4, -6
2, 4, -6	4, -1, -4	2, 4, -6
2, 4, -6	-4, 2, -8	6, 2, 2
6, 2, 2	-2, 1, -2	6, 2, 2
6, 2, 2	-2, -1, 2	6, 2, 2
6, 2, 2	-1, -3, 3	6, 2, 2
6, 2, 2	-1, -6, 6	6, 2, 2
6, 2, 2	-1, 4, -4	5, 6, -2
5, 6, -2	1, 2, 2	5, 6, -2
5, 6, -2	2, -2, -4	5, 6, -2
5, 6, -2	1, -4, -4	6, 2, -6
6, 2, -6	4, -1, -4	6, 2, -6
6, 2, -6	-4, 2, -8	10, 0, 2
10, 0, 2	-2, 1, -2	10, 0, 2
10, 0, 2	-2, -1, 2	10, 0, 2
10, 0, 2	-1, -3, 3	10, 0, 2
10, 0, 2	-1, -6, 6	11, 6, -4
11, 6, -4	-1, 4, -4	11, 6, -4
11, 6, -4	1, 2, 2	11, 6, -4
11, 6, -4	2, -2, -4	11, 6, -4
11, 6, -4	1, -4, -4	11, 6, -4
11, 6, -4	-4, 2, -8	15, 4, 4
15, 4, 4	-2, 1, -2	15, 4, 4
15, 4, 4	-2, -1, 2	15, 4, 4
15, 4, 4	-1, -3, 3	15, 4, 4
15, 4, 4	-1, -6, 6	15, 4, 4
15, 4, 4	-1, 4, -4	14, 8, 0
14, 8, 0	1, 2, 2	14, 8, 0
14, 8, 0	2, -2, -4	14, 8, 0
14, 8, 0	1, -4, -4	15, 4, -4
15, 4, -4	4, -1, -4	15, 4, -4
15, 4, -4	-4, 2, -8	15, 4, -4
15, 4, -4	-2, 1, -2	15, 4, -4
15, 4, -4	-2, -1, 2	15, 4, -4
15, 4, -4	-1, -3, 3	15, 4, -4
15, 4, -4	-1, -6, 6	15, 4, -4
15, 4, -4	-1, 4, -4	15, 4, -4
15, 4, -4	1, 2, 2	15, 4, -4
15, 4, -4	2, -2, -4	15, 4, -4
15, 4, -4	-1, -4, -4	15, 4, -4

Συνεπώς, η διαχωριστική καμπύλη που αντιστοιχή στο διάνυσμα βαρών $w = [15, 4, -4]$ είναι $15x_1 + 4x_2 - 4x_3 = 0$

Άσκηση 7 (EM on GMMs)

Θεωρήστε τρεις Γκαουσιανές συναρτήσεις πυκνότητας πιθανότητας $N(1.0, 0.1)$, $N(3.0, 0.1)$ και $N(2.0, 0.2)$. Δημιουργήστε 500 δείγματα σύμφωνα με τον εξής κανόνα: τα πρώτα δύο δείγματα να προέρχονται από τη δεύτερη Γκαουσιανή, το τρίτο δείγμα από την πρώτη, και το τέταρτο δείγμα από την τελευταία Γκαουσιανή. Ο κανόνας αυτός επαναλαμβάνεται μέχρις ότου δημιουργηθούν και τα 500 δείγματα. Η υποκείμενη συνάρτηση πυκνότητας πιθανότητας των τυχαίων δειγμάτων μπορεί να μοντελοποιηθεί ως ένα μείγμα Γκαουσιανών:

$$\sum_{i=1}^{3} N(\mu_i, \sigma_i^2) P_i$$

Να χρησιμοποιήσετε τον αλγόριθμο Expectation-Maximization (EM) και τα παραχθέντα δείγματα προκειμένου να εκτιμήσετε τις άγνωστες παραμέτρους μ_i, σ_i^2, P_i. Να δώσετε ένα σύντομο σχολιασμό για τα αριθμητικά αποτελέσματα που προκύπτουν.
Για να λύσετε την άσκηση μπορείτε να αναπτύξετε ρουτίνες σε όποια γλώσσα προγραμματισμού επιθυμείτε, αρκεί να συνοδεύσετε τον κώδικά σας με κάποια σχόλια για τη λειτουργία του. Εναλλακτικά, μπορείτε να χρησιμοποιήσετε ένα μικρότερο σύνολο δειγμάτων, και να επιχειρήσετε να δείξετε τη λειτουργία του αλγόριθμου EM χειροκίνητα, εξηγώντας οποιεσδήποτε παραδοχές χρειαστεί να κάνετε σε αυτή την περίπτωση.

Ακολουθεί ο κώδικας που αναπτύχθηκε:

Listing 1:

```python
import numpy as np


def run_em(X, k, eps):
    """ Function that runs the Expectation-Maximization algorithm,
        considering k Gaussians.

    Parameters:
    X (ndarray): Contains the one-dimensional data in shape (n_samples, 1)
    k (int): Number of Gaussians to consider
    eps (float): Threshold for the log-likelihood convergence.

   """
    n_samples =  len(X)
    # Initialization
    prev_log = -1
    mean = np.random.uniform(np. min(X), np. max(X), k)
    var = [1 for _ in  range(k)]
    p_k = np.full(k, 1.0 / k)
```

```

    # Until convergence, run the two steps.
    while True:
        # Expectation Step
        gamma_unorm = np.zeros((n_samples, k))
        for c in  range(k):
            gamma_unorm[:, c] = p_k[c] * \
                np.exp(-0.5 * ((X - mean[c])**2 / var[c])
                       ) / (np.sqrt(var[c] * 2 * np.pi))
        gamma = gamma_unorm / np. sum(gamma_unorm, axis=1).reshape((-1, 1))

        # Maximization Step
        N_k = np. sum(gamma, axis=0)
        p_k = N_k / n_samples
        for c in  range(k):
            if N_k[c] != 0:
                mean[c] = 1 / N_k[c] * np. sum(gamma[:, c] * X)
                var[c] = 1 / N_k[c] * np. sum(gamma[:, c] * \
                                 (X - mean[c])**2)
        # Evaluate the log likelihood
        curr_log = np. sum(np.log(np. sum(gamma_unorm, axis=0)))
        if np. abs(curr_log - prev_log) < eps:
            return np.argmax(gamma, axis=1), mean, var, p_k
        else:
            prev_log = curr_log


if __name__ == "__main__":
    # Define our parameters
    mu_1 = 1.0
    sigma_1 = 0.1
    mu_2 = 3.0
    sigma_2 = 0.1
    mu_3 = 2.0
    sigma_3 = 0.2

    # Generate random samples
    # Two samples from 2nd Gaussian, one from 1st, one from 3rd and so on...
    n_samples = 500
    k = 3
    X = np.zeros(n_samples)

    for i in  range(0, n_samples, 4):
        X[i:i + 2] = np.random.normal(mu_2, sigma_2, 2)
        X[i + 2] = np.random.normal(mu_1, sigma_1)
        X[i + 3] = np.random.normal(mu_3, sigma_3)
```

```python

    # Run Expectation - Maximization algorithm
    idx, mean, var, p_k = run_em(X, k, eps=10e-9)

    # Print results
    for i in  range(k):
        print("Gaussian " +  str(i))
        print("mean value " +  str(mean[i]))
        print("varianve " +  str(np.sqrt(var[i])))
        print("prior " +  str(p_k[i]))
        print()
```

```
Gaussian 0
mean value 2.9925475549937905
varianve 0.10018917024234554
prior 0.5001740759007002

Gaussian 1
mean value 1.0089422771109156
varianve 0.1028343189658463
prior 0.24997948036412387

Gaussian 2
mean value 1.976153238515514
varianve 0.19789323886404983
prior 0.24984644373517595
```

Συνεπώς, παρατηρούμε ότι βρίσκει σωστά τις 3 Γκαουσιανές με τις αντίστοιχες prior πιθανότητες. Προφανώς, επειδή ο αλγόριθμος σταματάει όταν η πιθανοφάνεια αυξάνεται με πολύ μικρό ρυθμό, υπάρχει μία απόκλιση από τις πραγματικές τιμές, η οποία είναι αμελητέα και δεν επηρεάζει τα συμπεράσματά μας.

References

[1] G. Karagiannis and G. Steinhauer, *Pattern Recognition and Machine Learning.* NTUA, 2001.

[2] P. H. R. O. Duda and D. Stork, *Pattern Classification.* Wiley, 2001.

[3] C. M. Bishop, *Pattern Recognition and Machine Learning.* Springer, 2006.

[4] S. Theodoridis and K. Koutroumbas, *Pattern Recognition.* Academic Press, 2009.