Μηχανική Μάθηση

Μιχάλης Τίτσιας

Διάλεξη 8ή

Ο αλγόριθμος ΕΜ γενικά, ομαδοποίηση δυαδικών δεδομένων και αριθμητική ευστάθεια στην υλοποίηση ΕΜ αλγορίθμων

Περιεχόμενα

Πιθανοτική ομαδοποίηση και μίξεις Gaussian κατανομών

Το πιθανοτικό μοντέλο ομαδοποίησης (για K ομάδες) υποθέτει ότι κάθε δεδομένο παράγεται ως εξής

- Καθορισμός της ομάδας από την οποία θα παραχθεί το δεδομένο $\Rightarrow$ καθορισμός της τιμής μιας τυχαίας μεταβλητής $\mathbf{z}$

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$$

όπου π_k είναι η εκ των προτέρων πιθανότητα της ομάδας k

- Παραγωγή του δεδομένου δοθέντος του $\mathbf{z}$

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} [p(\mathbf{x}|k)]^{z_k}$$

όπου $p(\mathbf{x}|k)$ είναι μια υπό συνθήκη κατανομή που περιγράφει το πως παράγονται τα δεδομένα της ομάδας k

Πιθανοτική ομαδοποίηση και μίξεις Gaussian κατανομών

- Το $\mathbf{z}$ είναι κρυμμένη μεταβλητή $\Rightarrow$ δεν παρατηρείται
- Μόνο το δεδομένο $\mathbf{x}$ παρατηρείται και είναι γνωστό
- Η ολική πιθανότητα/κατανομή του $\mathbf{x}$ υπολογίζεται εφαρμόζοντας τον κανόνα αθροίσματος και περιθωριοποιώντας το άγνωστο $\mathbf{z}$

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}|k)$$

Η κατανομή $p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}|k)$ ονομάζεται μίξη κατανομών

- $\Rightarrow$ με μια εξίσωση η μίξη κατανομών περιγράφει όλο το πιθανοτικό μοντέλο για ομαδοποίηση δεδομένων

Πιθανοτική ομαδοποίηση και μίξεις Gaussian κατανομών

- **Μίξη κατανομών**:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}|k), \quad 0 \leq \pi_k \leq 1, \quad \sum_{k=1}^{K} \pi_k = 1$$

- Η υπό συνθήκη κατανομή $p(\mathbf{x}|k)$ ονομάζεται συνιστώσα κατανομή (mixture component)
- Η εκ των προτέρων πιθανότητα της ομάδας π_k ονομάζεται συντελεστής μίξης (mixing coefficient)

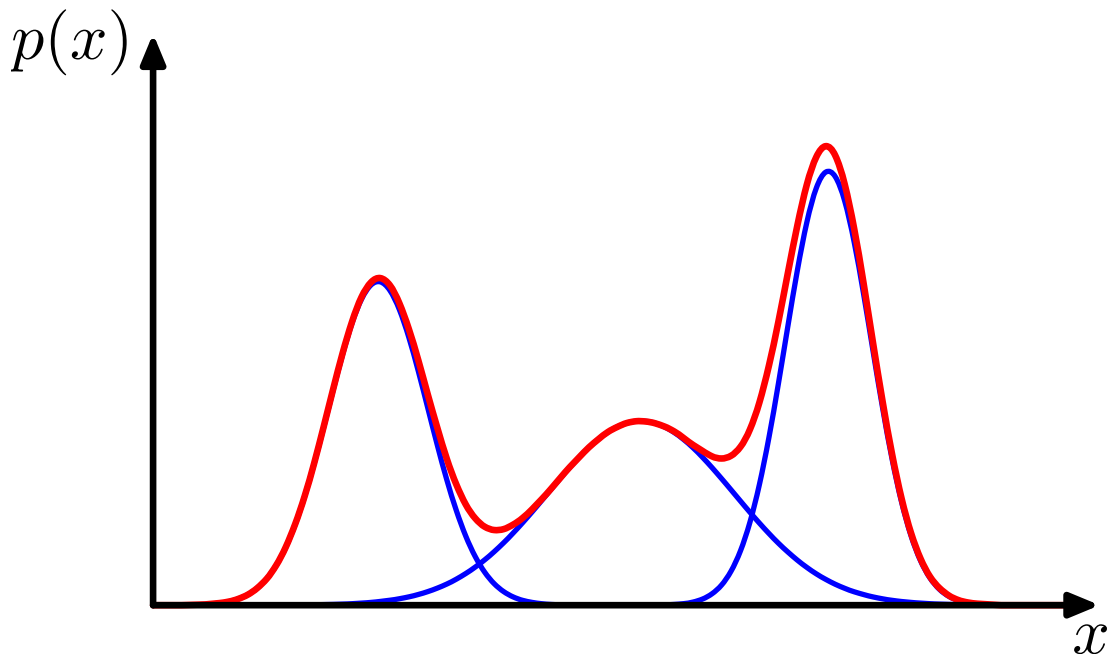
Συγκεκριμένες περιπτώσεις αυτών των μοντέλων προκύπτουν επιλέγοντας συγκεκριμένη συναρτησιακή μορφή για τις συνιστώσες κατανομές $p(\mathbf{x}|k)$

- Η συνηθέστερη είναι να επιλέξουμε για κάθε $p(\mathbf{x}|k)$ μια Gaussian

Πιθανοτική ομαδοποίηση και μίξεις Gaussian κατανομών

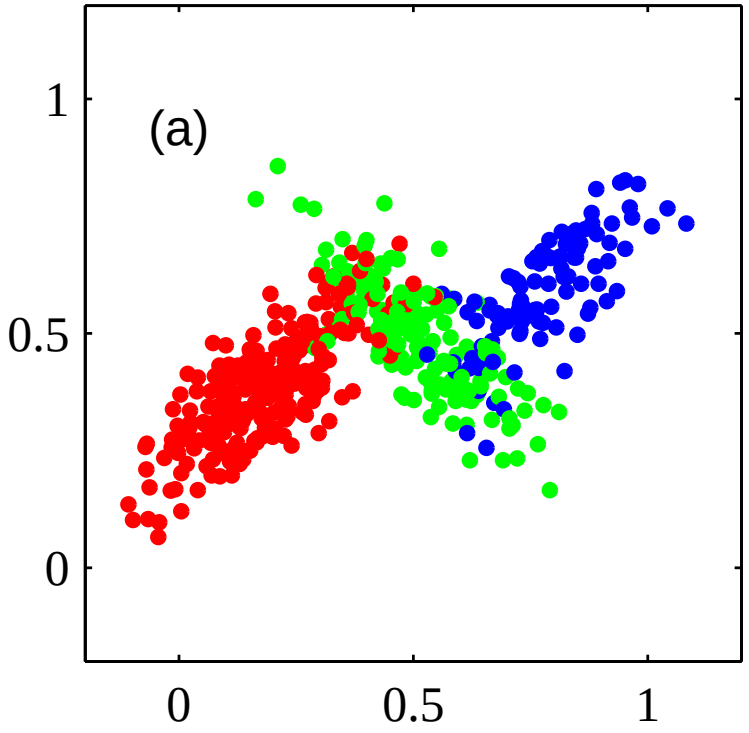
- **Μίξη Gaussian κατανομών**:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k), \quad 0 \leq \pi_k \leq 1, \quad \sum_{k=1}^{K} \pi_k = 1$$



Μίξη Gaussian κατανομών (κόκκινη γραμμή) για μονοδιάστατα δεδομένα. Οι μπλε γραμμές δείχνουν τους όρους $\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k)$

Πιθανοτική ομαδοποίηση και μίξεις Gaussian κατανομών



Μίξη από τρεις ($K = 3$) Gaussian θα μπορούσε να είχε παράγει τα δεδομένα του σχήματος τα οποία ανήκουν σε τρεις ομάδες

Η θέση και το ελλειψοειδές σχήμα της κάθε ομάδας k καθορίζεται πλήρως από τις τιμές των παραμέτρων $(\boldsymbol{\mu}_k, \Sigma_k)$

Η συχνότητα εμφάνισης δεδομένων μιας ομάδας εξαρτάται από την παράμετρο π_k

Πιθανοτική ομαδοποίηση και μίξεις Gaussian κατανομών

Μέγιστη πιθανοφάνεια: Εκτίμηση των παραμέτρων
$(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \{\pi_k, \boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^K$

- Δοθέντος ενός συνόλου δεδομένων εκπαίδευσης $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$, υποθέτουμε ότι το καθένα έχει παραχθεί ανεξάρτητα από τη μίξη $p(\mathbf{x}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k)$
- Από κοινού κατανομή

$$p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{n=1}^N \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \Sigma_k)$$

- Λογαριθμική πιθανοφάνεια

$$\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^N \log \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \Sigma_k)$$

- Ως συνήθως θέλουμε να την μεγιστοποιήσουμε ως προς τις παραμέτρους
  - $\Rightarrow$ δύσκολο λόγω του αθροίσματος που υπάρχει μέσα στον λογάριθμο

EM αλγόριθμος για μίξεις Gaussian κατανομών

Ο αλγόριθμος **EM (expectation-maximization)** για μίξη Gaussian κατανομών

1. **Ε βήμα:**

$$\gamma(z_{nk}) = \frac{\pi_k^{(t)} \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k^{(t)}, \Sigma_k^{(t)})}{\sum_{j=1}^{K} \pi_j^{(t)} \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j^{(t)}, \Sigma_j^{(t)})}, \quad n = 1, \ldots, N$$

2. **Μ βήμα:**

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n}{\sum_{n=1} \gamma(z_{nk})}$$

$$\Sigma_k^{(t+1)} = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}_n - \boldsymbol{\mu}_k^{(t+1)})(\mathbf{x}_n - \boldsymbol{\mu}_k^{(t+1)})^T}{\sum_{n=1}^{N} \gamma(z_{nk})}$$

$$\pi_k^{(t+1)} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})}{N}$$

Μπορεί να δειχθεί ότι επαναληπτική εφαρμογή των παραπάνω βημάτων συγκλίνει σε ένα τοπικό μέγιστο της λογαριθμικής πιθανοφάνειας $\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \Sigma_k)$

ΕΜ αλγόριθμος για μίξεις Gaussian κατανομών

Ο αλγόριθμος **EM (expectation-maximization)** για μίξη Gaussian κατανομών

1. **Ε βήμα:**

$$\gamma(z_{nk}) = \frac{\pi_k^{(t)} \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k^{(t)}, \Sigma_k^{(t)})}{\sum_{j=1}^{K} \pi_j^{(t)} \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j^{(t)}, \Sigma_j^{(t)})}, \quad n = 1, \ldots, N$$

2. **Μ βήμα:**

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n}{\sum_{n=1} \gamma(z_{nk})}$$

$$\Sigma_k^{(t+1)} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k^{(t+1)})(\mathbf{x}_n - \boldsymbol{\mu}_k^{(t+1)})^T}{\sum_{n=1}^{N} \gamma(z_{nk})}$$

$$\pi_k^{(t+1)} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})}{N}$$

Παρατηρούμε ότι βασίζεται στη χρήση των κρυμμένων μεταβλητών $\{\mathbf{z}_n\}_{n=1}^{N}$. Ο ΕΜ εφαρμόζεται για μεγιστοποίηση λογαριθμικών πιθανοφάνειων σε πολλά άλλα πιθανοτικά μοντέλα και σε κάθε περίπτωση βασίζεται στη χρήση κρυμμένων μεταβλητών

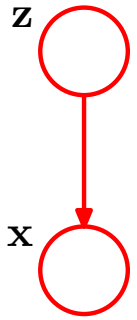
Κρυμμένες μεταβλητές

Κρυμμένες μεταβλητές εμφανίζονται με φυσικό τρόπο σε πολλά προβλήματα ... όχι απαραίτητα του τομέα της μάθησης μηχανικής!

- Παραδείγματα:
  - Παρατηρούμε τα κλινικά συμπτώματα ενός ασθενή, αλλά θέλουμε να εξακριβώσουμε την κρυμμένη ασθένεια
  - Παρατηρούμε τα αποτελέσματα των εξετάσεων κάποιων φοιτητών, αλλά αυτό που θα μας ενδιέφερε επίσης είναι η κρυμμένη νοητική τους ικανότητα
  - Παρατηρούμε τα clicks ενός χρήστη στο διαδίκτυο αλλά θα θέλαμε να προσδιορίσουμε τα κρυμμένα ενδιαφέροντα του
    - προκειμένου, π.χ. να του προτείνουμε προιόντα που ίσως να αγοράσει
- Διαισθητικά οι κρυμμένες μεταβλητές αναπαριστούν αιτίες που διέπουν την παραγωγή των δεδομένων
  - Ωστόσο ενώ παρατηρούμε τα δεδομένα, τις αιτίες δεν τις παρατηρούμε!

Τα περισσότερα πιθανοτικά μοντέλα για μάθηση χωρίς επίβλεψη έχουν να κάνουν με κρυμμένες μεταβλητές

Κρυμμένες μεταβλητές



- Τα μοντέλα με κρυμμένες μεταβλητές στην απλούστερή τους μορφή υποθέτουν για κάθε παρατηρούμενο δεδομένο $\mathbf{x}$ μια αντίστοιχη κρυμμένη μεταβλητή $\mathbf{z}$. Η από κοινού τους κατανομή αναπαρίστατα από το γράφο του Σχήματος και έχει την μορφή

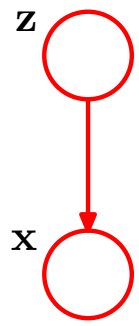
$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) = p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})$$

  όπου $\boldsymbol{\theta}$ είναι παραμέτροι

- Αφού το $\mathbf{z}$ είναι κρυμμένη (δηλ. μη παρατηρήσιμη) μεταβλητή, η κατανομή του δεδομένου $\mathbf{x}$ προκύπτει από το κανόνα αθροίσματος

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})$$

Κρυμμένες μεταβλητές



$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) = p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})$$

- Η μίξη Gaussian κατανομών είναι ένα μοντέλο της παραπάνω μορφής, αλλά υπάρχουν πολλά άλλα
- Ο αλγόριθμος ΕΜ γενικεύεται για οποιοδήποτε μοντέλο με κρυμμένες μεταβλητές
- Στην συνέχεια θα παρουσιάσουμε τον ΕΜ στη γενική του μορφή

Σημειώση: Η ιστορίες του ΕΜ και των πιθανοτικών μοντέλων με κρυμμένες μεταβλητές είναι μεταξύ τους αλληλένδετες. Πολλά μοντέλα ενδεχομένως επινοήθηκαν λόγω της υπάρξης του ΕΜ που μπορούσε να τα εκπαιδεύσει εύκολα!

Ο αλγόριθμος EM

- Έστω ότι έχουμε δεδομένα $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ τα οποία έχουν παραχθεί ανεξάρτητα μεταξύ τους μέσω κρυμμένων μεταβλητών $\{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$
- Η από κοινού κατανομή (η πιθανοφάνεια) είναι

$$p(X|\boldsymbol{\theta}) = \prod_{n=1}^{N} p(\mathbf{x}_n) = \prod_{n=1}^{N} \sum_{\mathbf{z}_n} p(\mathbf{x}_n, \mathbf{z}_n|\boldsymbol{\theta})$$

- Η λογαριθμική πιθανοφάνεια είναι

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log p(\mathbf{x}_n) = \sum_{n=1}^{N} \log \sum_{\mathbf{z}_n} p(\mathbf{x}_n, \mathbf{z}_n|\boldsymbol{\theta})$$

- Αυτού του είδους η λογαριθμική πιθανοφάνεια είναι δύσκολο να μεγιστοποιηθεί λόγω του αθροίσματος που υπάρχει μέσα στο λογάριθμο
  - η $\mathcal{L}(\boldsymbol{\theta})$ ενδέχεται να έχει πολλά τοπικά μέγιστα

Ο αλγόριθμος EM

- Αν γνωρίζαμε την τιμή της $\mathbf{z}_n$ (που εξήγει το πως παράχθηκε το $\mathbf{x}_n$) τότε το δεδομένο μας δεν θα ήταν το $\mathbf{x}_n$ αλλά το ζεύγος $(\mathbf{x}_n, \mathbf{z}_n)$ και η λογαριθμική πιθανοφάνεια δεν θα ήταν η

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log p(\mathbf{x}_n|\boldsymbol{\theta})$$

αλλά η

$$\mathcal{L}_c(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log p(\mathbf{x}_n, \mathbf{z}_n|\boldsymbol{\theta})$$

- Υπό αυτό το πρίσμα η $\mathcal{L}(\boldsymbol{\theta})$ ονομάζεται incomplete λογαριθμική πιθανοφάνεια, ενώ η $\mathcal{L}_c(\boldsymbol{\theta})$ complete
- Η μεγιστοποίηση της complete λογαριθμικής πιθανοφάνειας είναι θεωρητικά εύκολη (ωστόσο στην πράξη αδύνατη αφού δεν γνωρίζουμε τα $\{\mathbf{z}_n\}_{n=1}^{N}$!)

Ο αλγόριθμος EM

(Παρένθεση: Πάραδειγμα μεγιστοποίησης complete λογαριθμικής πιθανοφάνειας για μίξεις από Gaussian)

- Η complete λογαριθμική πιθανοφάνεια γενικά είναι $\mathcal{L}_c(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log p(\mathbf{x}_n, \mathbf{z}_n|\boldsymbol{\theta})$
- Στην περίπτωση της μίξης από Gaussian, όπου $p(\mathbf{x}_n, \mathbf{z}_n|\boldsymbol{\theta}) = \prod_{k=1}^{K} [\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \Sigma_k)]^{z_{nk}}$, έχουμε

$$\begin{aligned}
\mathcal{L}_c(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \sum_{n=1}^{N} \log \prod_{k=1}^{K} [\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \Sigma_k)]^{z_{nk}} \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \log [\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \Sigma_k)] \\
&= \sum_{k=1}^{K} \left(\sum_{n=1}^{N} z_{nk} \right) \log \pi_k + \sum_{k=1}^{K} \left(\sum_{n=1}^{N} z_{nk} \log \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \Sigma_k) \right) \\
&= \sum_{k=1}^{K} \left(\sum_{n=1}^{N} z_{nk} \right) \log \pi_k + \sum_{k=1}^{K} \mathcal{L}_k(\boldsymbol{\mu}_k, \Sigma_k)
\end{aligned}$$

Ο αλγόριθμος EM

(Παρένθεση: Πάραδειγμα μεγιστοποίησης complete λογαριθμικής πιθανοφάνειας για μίξεις από Gaussian)

$$\mathcal{L}_c(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{k=1}^{K} \left(\sum_{n=1}^{N} z_{nk} \right) \log \pi_k + \sum_{k=1}^{K} \mathcal{L}_k(\boldsymbol{\mu}_k, \Sigma_k)$$

όπου $\mathcal{L}_k(\boldsymbol{\mu}_k, \Sigma_k) = \sum_{n=1}^{N} z_{nk} \log \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \Sigma_k)$

Η μεγιστοποίηση γίνεται αναλυτικά (δες Διάλεξη 5ή, διαφάνειες 37-38 - η λύση είναι ακριβώς η ίδια!) και δίνει

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} z_{nk} \mathbf{x}_n}{\sum_{n=1}^{N} z_{nk}}$$

(μέση τιμή των δεδομένων της ομάδας k)

$$\Sigma_k = \frac{\sum_{n=1}^{N} z_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T}{\sum_{n=1}^{N} z_{nk}}$$

(μέση διακύμανση των δεδομένων της ομάδας k)

$$\pi_k = \frac{\sum_{n=1}^{N} z_{nk}}{N}$$

(πλήθος των δεδομένων της ομάδας k διά του συνολικού αριθμού)

Ο αλγόριθμος EM

Ιδέα πίσω από τον EM

- Αν γνωρίζαμε την τιμή της $\mathbf{z}_n$ (που εξήγει το πως παράχθηκε το $\mathbf{x}_n$) τότε το δεδομένο μας δεν θα ήταν το $\mathbf{x}_n$ αλλά το ζεύγος $(\mathbf{x}_n, \mathbf{z}_n)$ και η λογαριθμική πιθανοφάνεια θα ήταν η

$$\mathcal{L}_c(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log p(\mathbf{x}_n, \mathbf{z}_n|\boldsymbol{\theta})$$

  την οποία θα μεγιστοποιούσαμε εύκολα

- Αντιθέτως αν γνωρίζαμε την βέλτιστη τιμή για το $\boldsymbol{\theta}$ θα μπορούσαμε να εκτίμησουμε τις άγνωστες τιμές των κρυμμένων μεταβλήτων

- Ωστόσο το πρόβλημα μας είναι ότι ούτε τα $\mathbf{z}_n$ γνωρίζουμε αλλά ούτε και το βέλτιστο $\boldsymbol{\theta}$!

Ο EM προσπαθεί το λύσει αυτό το chicken-and-egg πρόβλημα επαναληπτικά «γεμίζοντας» τις τιμές των κρυμμένων μεταβλητών και μαθαίνοντας το $\boldsymbol{\theta}$

Ο αλγόριθμος EM

Ιδέα πίσω από τον EM

Αφού δεν μπορούμε να υπολογίσουμε την complete λογαριθμική πιθανοφάνεια

$$\mathcal{L}_c(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log p(\mathbf{x}_n, \mathbf{z}_n|\boldsymbol{\theta})$$

(λόγω ότι δεν γνωρίζουμε το κάθε $\mathbf{z}_n$), ο EM υπολογίζει τη μέση τιμή της δόθεντος της τρέχουσας τιμής των παραμέτρων $\boldsymbol{\theta}^{(t)}$

Η μέση τιμή της κάθε ποσότητας $\log p(\mathbf{x}_n, \mathbf{z}_n|\boldsymbol{\theta})$ ως προς την εκ των υστέρων πιθανότητα $p(\mathbf{z}_n|\mathbf{x}_n, \boldsymbol{\theta}^{(t)})$ είναι

$$\sum_{\mathbf{z}_n} p(\mathbf{z}_n|\mathbf{x}_n, \boldsymbol{\theta}^{(t)}) \log p(\mathbf{x}_n, \mathbf{z}_n|\boldsymbol{\theta})$$

- $\Rightarrow$ κατά κάποιο τρόπο αυτό προσπαθεί να «γεμίσει» τις κρυμμένες μεταβλητές βάσει της τρέχουσας τιμής των παραμέτρων $\boldsymbol{\theta}^{(t)}$

Ιδέα πίσω από τον EM

Οπότε από την complete λογαριθμική πιθανοφάνεια

$$\mathcal{L}_c(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log p(\mathbf{x}_n, \mathbf{z}_n|\boldsymbol{\theta})$$

και παίρνωντας μέση τιμή προκύπτει η expected complete λογαριθμική πιθανοφάνεια

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = \sum_{n=1}^{N} \sum_{\mathbf{z}_n} p(\mathbf{z}_n|\mathbf{x}_n, \boldsymbol{\theta}^{(t)}) \log p(\mathbf{x}_n, \mathbf{z}_n|\boldsymbol{\theta})$$

- $\Rightarrow$ που είναι υπολογίσιμη και μπορεί να μεγιστοποιηθεί ως προς $\boldsymbol{\theta}$ και να μας δώσει μια νέα τιμή $\boldsymbol{\theta}^{(t+1)}$

Ο αλγόριθμος EM

Αλγορίθμος EM

1. Αρχικοποίησε $\boldsymbol{\theta}^{(0)}$, $t = 0$
2. Ε βήμα: (Υπολόγισε τις εκ των υστέρων πιθανοτήτες δοθέντος των παραμέτρων $\boldsymbol{\theta}^{(t)}$)

$$p(\mathbf{z}_n|\mathbf{x}_n, \boldsymbol{\theta}^{(t)}) = \frac{p(\mathbf{x}, \mathbf{z}_n|\boldsymbol{\theta}^{(t)})}{\sum_{\mathbf{z}_n} p(\mathbf{x}, \mathbf{z}_n|\boldsymbol{\theta}^{(t)})}, \quad n = 1, \ldots, N$$

3. Μ βήμα: (Ενημέρωσε τις παραμέτρους)

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$$

όπου $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = \sum_{n=1}^{N} \sum_{\mathbf{z}_n} p(\mathbf{z}_n|\mathbf{x}_n, \boldsymbol{\theta}^{(t)}) \log p(\mathbf{x}_n, \mathbf{z}_n|\boldsymbol{\theta})$

4. Θέσε $t = t + 1$ και πήγαινε στο βήμα 2 ή τερμάτισε

Ο αλγόριθμος EM

Π.χ. για μίξη από Gaussian όπου η complete λογ. πιθανοφάνεια είναι
$\mathcal{L}_c(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{k=1}^{K} \left(\sum_{n=1}^{N} z_{nk} \right) \log \pi_k + \sum_{k=1}^{K} \left(\sum_{n=1}^{N} z_{nk} \log \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \Sigma_k) \right)$

Η expected complete λογαριθμική πιθανοφάνεια προκύπτει απλά αντικαθιστώντας το κάθε z_{nk} με την εκ των υστέρων πιθανότητα
$\gamma(z_{nk}) = p(\mathbf{z}_n | \mathbf{x}_n, \boldsymbol{\theta}^{(t)})$

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = \sum_{k=1}^{K} \left(\sum_{n=1}^{N} \gamma(z_{nk}) \right) \log \pi_k + \sum_{k=1}^{K} \left(\sum_{n=1}^{N} \gamma(z_{nk}) \log \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \Sigma_k) \right)$$

Μεγιστοποίηση ως προς $\boldsymbol{\theta} = \{\pi_k, \boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^{K}$ δίνει τις (γνώστες μας) εξισώσεις του M βήματος

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n}{\sum_{n=1} \gamma(z_{nk})}$$

$$\Sigma_k^{(t+1)} = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}_n - \boldsymbol{\mu}_k^{(t+1)}) (\mathbf{x}_n - \boldsymbol{\mu}_k^{(t+1)})^T}{\sum_{n=1}^{N} \gamma(z_{nk})}$$

$$\pi_k^{(t+1)} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})}{N}$$

Η ιδιότητα σύγκλισης του ΕΜ

Ο ΕΜ μεγιστοποιεί την λογαριθμική πιθανοφάνεια

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log p(\mathbf{x}_n|\boldsymbol{\theta}) = \sum_{n=1}^{N} \log \sum_{\mathbf{z}_n} p(\mathbf{x}_n, \mathbf{z}_n|\boldsymbol{\theta})$$

δηλ. βρίσκει ένα τοπικό μέγιστο της ποσότητας αυτής

Αυτό συμβαίνει διότι έχει την ιδιότητα της μονότονης αύξησης της $\mathcal{L}(\boldsymbol{\theta})$ σε κάθε βήμα, δηλ. ισχύει

$$\mathcal{L}(\boldsymbol{\theta}^{(t+1)}) \geq \mathcal{L}(\boldsymbol{\theta}^{(t)})$$

- Η μαθηματική απόδειξη της ιδιότητας αυτής βασίζεται στην ανισότητα Jensen και περιγράφεται στο Παράρτημα

Η ιδιότητα σύγκλισης του EM

$$\mathcal{L}(\boldsymbol{\theta}^{(t+1)}) \geq \mathcal{L}(\boldsymbol{\theta}^{(t)})$$

Η ιδιότητα αυτή είναι **φοβερά χρήσιμη για debugging**

- Στο κώδικα σου μετά το τέλος κάθε επανάληψης του EM θα πρέπει να έχεις την συνθήκη:
  If $\mathcal{L}(\boldsymbol{\theta}^{(t+1)}) - \mathcal{L}(\boldsymbol{\theta}^{(t)}) < 0$ then error found, stop the code!

Εφαρμογή του EM σε μίξη Bernoulli κατανομών

Όταν βγάζουμε τις εξισώσεις ενός EM αλγορίθμου συνήθως ακολουθούμε τα παρακάτω βήματα

- Αναγνωρίζουμε ποια είναι η κρυμμένη μεταβλητή στο μοντέλο μας και γράφουμε την complete πιθανοφάνεια και έπειτα την complete λογαριθμική πιθανοφάνεια
- Γράφουμε την expected complete λογαριθμική πιθανοφάνεια $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$
- Παραγωγίζουμε την expected complete λογαριθμική πιθανοφάνεια ως προς τις παραμέτρους $\boldsymbol{\theta}$ και θέτουμε στο μηδέν προκειμένου να εξάγουμε τις εξισώσεις του M βήματος

Ας δούμε ένα παράδειγμα σε μίξη Bernoulli κατανομών

Εφαρμογή του EM σε μίξη Bernoulli κατανομών

Έχουμε δεδομένα $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ όπου κάθε $\mathbf{x}_n$ είναι ένα δυαδικό διάνυσμα διάστασης D. Υποθέτουμε την ακόλουθη μίξη Bernoulli κατανομών

$$p(\mathbf{x}|M, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k \prod_{d=1}^{D} \mu_{kd}^{x_{nd}} (1 - \mu_{kd})^{1-x_{nd}}$$

και θέλουμε να μεγιστοποιήσουμε την λογαριθμική πιθανοφάνεια $\mathcal{L}(M, \boldsymbol{\pi}) = \sum_{n=1}^{N} \log p(\mathbf{x}_n|M, \boldsymbol{\pi})$ ως προς $M = \{\mu_{kd}\}_{k=1,d=1}^{K,D}$ και $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^{K}$

Complete πιθανοφάνεια

$$p(X, Z|M, \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \left[\pi_k \prod_{d=1}^{D} \mu_{kd}^{x_{nd}} (1 - \mu_{kd})^{1-x_{nd}} \right]^{z_{nk}}$$

Εφαρμογή του EM σε μίξη Bernoulli κατανομών

Complete λογαριθμική πιθανοφάνεια

$$\mathcal{L}_c(M, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left\{ \log \pi_k + \sum_{d=1}^{D} x_{nd} \left[\log \mu_{kd} + (1 - x_{nd}) \log(1 - \mu_{kd}) \right] \right\}$$

Expected complete λογαριθμική πιθανοφάνεια: Αφού το κάθε z_{nk} εμφανίζεται γραμμικά στην $\mathcal{L}_c$, η μέση τιμή της $\mathcal{L}_c$ προκύπτει αντικαθιστώντας τα z_{nk} με τη μέση τιμή τους, δηλ.

$$Q(M, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \left\{ \log \pi_k + \sum_{d=1}^{D} x_{nd} \left[\log \mu_{kd} + (1 - x_{nd}) \log(1 - \mu_{kd}) \right] \right\}$$

όπου $\gamma(z_{nk}) = \frac{\pi_k \prod_{d=1}^{D} \mu_{kd}^{x_{nd}} (1-\mu_{kd})^{1-x_{nd}}}{\sum_{j=1}^{K} \pi_j \prod_{d=1}^{D} \mu_{jd}^{x_{nd}} (1-\mu_{jd})^{1-x_{nd}}}$

Εφαρμογή του EM σε μίξη Bernoulli κατανομών

Expected complete λογαριθμική πιθανοφάνεια

$$Q(\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \left\{ \log \pi_k + \sum_{d=1}^{D} x_{nd} \left[\log \mu_{kd} + (1 - x_{nd}) \log(1 - \mu_{kd}) \right] \right\}$$

Για να βγάλουμε τις εξισώσεις του M βήματος, παραγωγίζουμε την παραπάνω συνάρτηση ως προς κάθε μ_{kd} και π_k θεωρώντας ότι τα $\gamma(z_{nk})$ είναι σταθερές (ως προς τα μ_{kd} και π_k) αφού έχουν υπολογιστεί στο E βήμα βάσει των προηγούμενων τιμών των παραμέτρων:

$$\frac{dQ}{d\mu_{kd}} = \sum_{n=1}^{N} \gamma(z_{nk}) \left[\frac{x_{nd}}{\mu_{kd}} - \frac{1 - x_{nd}}{1 - \mu_{kd}} \right] = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_{nd}}{\mu_{kd}} - \frac{\sum_{n=1}^{N} \gamma(z_{nk}) - \sum_{n=1}^{N} \gamma(z_{nk}) x_{n}}{1 - \mu_{kd}}$$

Οπότε $\frac{dQ}{d\mu_{kd}} = 0 \Rightarrow \mu_{kd} = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_{nd}}{\sum_{n=1}^{N} \gamma(z_{nk})}$. Ομοίως προκύπτει $\pi_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})}{N}$
(το οποίο είναι η εξίσωση του π_k για οποιοδήποτε μοντέλο μίξης!)

Εφαρμογή του EM σε μίξη Bernoulli κατανομών

1. Αρχικοποίηση παραμέτρων $\pi^{(0)}, M^{(0)}$
2. **Ε βήμα**:

$$\gamma(z_{nk}) = \frac{\pi_k^{(t)} \prod_{d=1}^{D} (\mu_{kd}^{(t)})^{x_{nd}} (1 - \mu_{kd}^{(t)})^{1-x_{nd}}}{\sum_{j=1}^{K} \pi_j^{(t)} \prod_{d=1}^{D} (\mu_{jd}^{(t)})^{x_{nd}} (1 - (\mu_{jd})^{(t)})^{1-x_{nd}}}, \quad n = 1, \ldots, N, \; k = 1, \ldots, K$$

3. **Μ βήμα**:

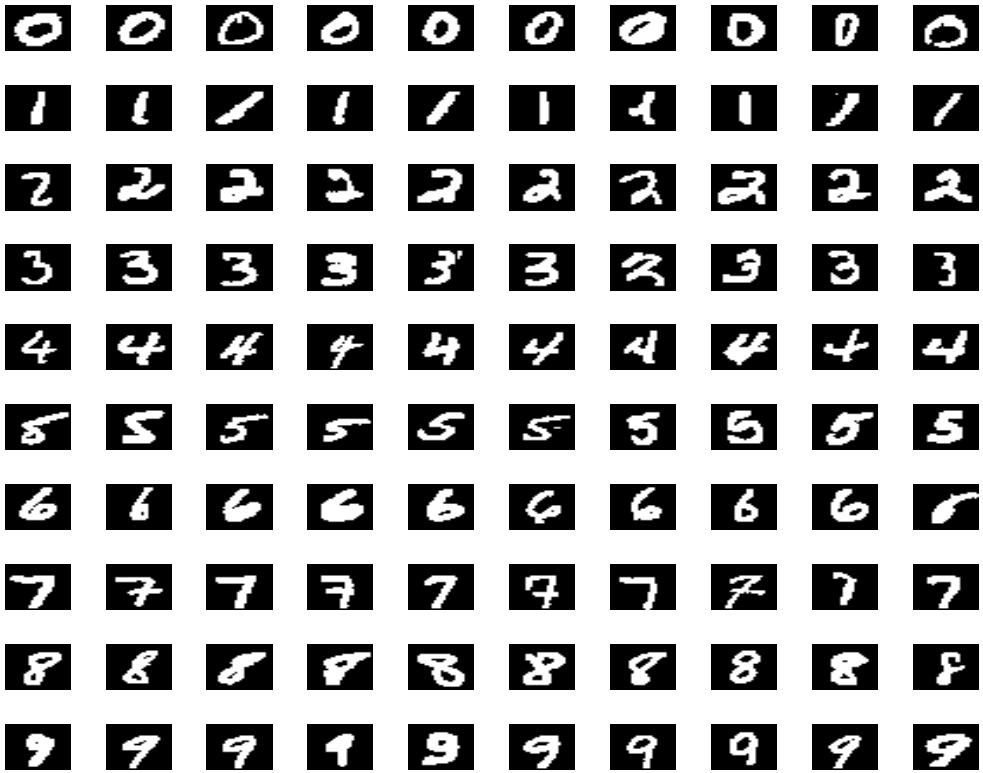
$$\mu_{kd}^{(t+1)} = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_{nd}}{\sum_{n=1} \gamma(z_{nk})}, \quad k = 1, \ldots, K, \; d = 1, \ldots, D$$

$$\pi_k^{(t+1)} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})}{N}, \quad k = 1, \ldots, K$$

4. Έλεγχος σύγκλισης (δηλ. τερματισμού) βάσει $|L(\boldsymbol{\theta}^{(t+1)}) - L(\boldsymbol{\theta}^{(t)})| < \epsilon$ όπου ϵ κάποια μικρή τιμή (tolerance)

Ομαδοποίηση με μίξη Bernoulli κατανομών

Παράδειγμα ομαδοποίησης χειρόγραφων χαρακτήρων

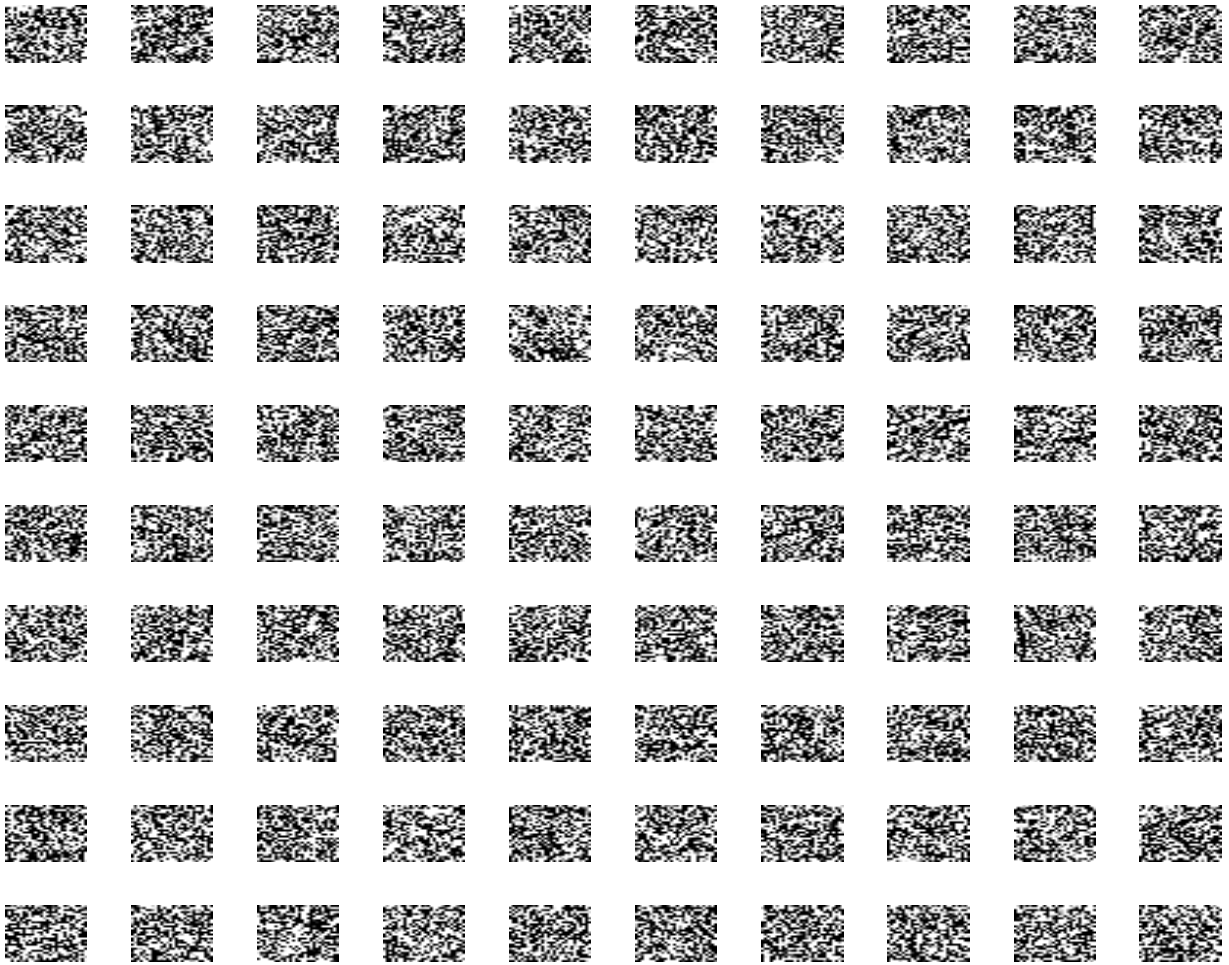


Σύνολο 60000 δυαδικών χειρόγραφων ψηφίων. Θα προσπαθήσουμε να ομαδοποιήσουμε όλους τους χαρακτήρες με μια μίξη από $K = 100$ Bernoulli κατανομές

$$p(\mathbf{x}) = \sum_{k=1}^{100} \pi_k \prod_{d=1}^{784} \mu_{kd}^{x_d} (1 - \mu_{kd})^{1 - x_{nd}}$$

Ομαδοποίηση με μίξη Bernoulli κατανομών

Παράδειγμα ομαδοποίησης χειρόγραφων χαρακτήρων



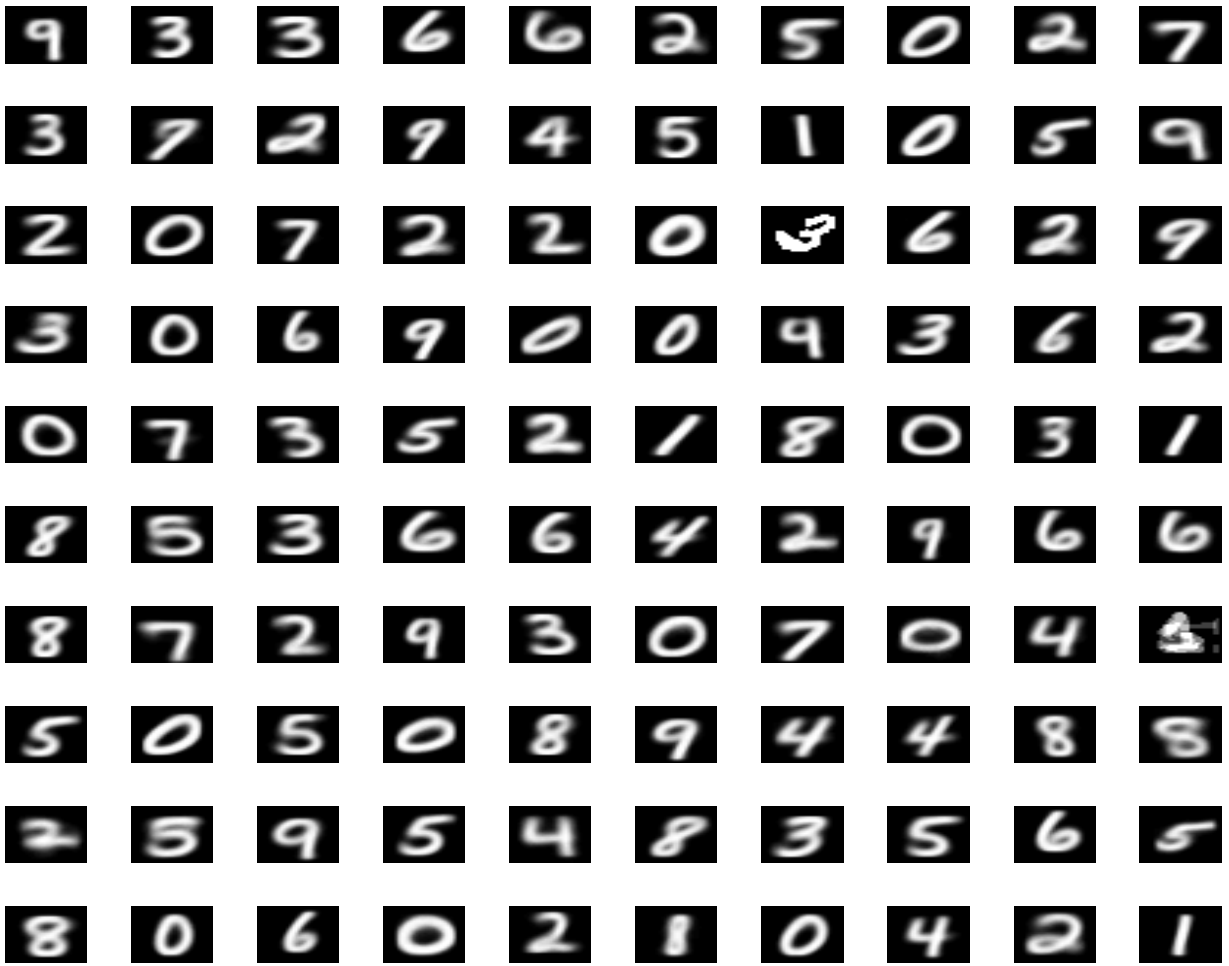
Αρχικοποίηση των παραμέτρων: $\pi_k = 1/K$, όπου $K = 100$ και

$$\mu_{kd} = \text{random in } [0.4, 0.6]$$

(η αρχικοποίηση του κάθε $\{\mu_{kd}\}_{d=1}^{784}$ φαίνεται στο Σχήμα)

Ομαδοποίηση με μίξη Bernoulli κατανομών

Παράδειγμα ομαδοποίησης χειρόγραφων χαρακτήρων



Στο Σχήμα φαίνονται οι τελικές τιμές του κάθε διανύσματος $\{\mu_{kd}\}_{d=1}^{784}$ μετά από την εφαρμογή του ΕΜ. Παρατηρούμε ότι το μοντέλο εκπαιδευμένο με ΕΜ ανακαλύπτει ως ομάδες ψηφία!

Πιθανοτική ομαδοποίηση για δεδομένα οποιασδήποτε μορφής

Ως τώρα έχουμε δει παραδείγματα ομαδοποίησης δεδομένων που παίρνουν συνεχείς τιμές (μίξεις από Gaussian) ή δυαδικές τιμές (μίξεις από Bernoulli)

Ωστόσο η ιδεά είναι πολύ γενική. Δηλ. μπορούμε με την ίδια μεθοδολογία να ομαδοποιούμε οπιασδήποτε μορφής δεδομένα

Αυτό που απαιτείται κάθε φορά είναι στο μοντέλο μίξης

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}|k)$$

να επιλέγουμε την κατάλληλη συνιστώσα κατανομή $p(\mathbf{x}|k)$. Π.χ.

- Αν το κάθε $\mathbf{x}$ είναι ένα γράφος θα πρέπει το $p(\mathbf{x}|k)$ να είναι μια κατανομή που να είναι κατάλληλη για γράφους
- Αν το κάθε $\mathbf{x}$ είναι ένα κείμενο θα πρέπει το $p(\mathbf{x}|k)$ να είναι μια κατανομή που να είναι κατάλληλη για κείμενα (η Bernoulli είναι κατάλληλη μέσω της δυαδικής αναπαράστασης κειμένων με παρουσία/απουσία λέξεων κλειδιών, ωστόσο υπάρχουν και αλλά μοντέλα βασιζόμενα σε άλλες μορφές αναπαράστασης - **δες στο Παράρτηρα στο τέλος των διαφανειών για ένα εναλλακτικό και αρκετά δημοφιλές μοντέλο**)

Κατηγοριοποίηση με μίξη κατανομών

Μπορούμε να χρησιμοποιήσουμε μίξεις για να κατασκευάσουμε περιγραφικά συστήματα κατηγοριοποίησης

Περιγραφικά συστήματα κατηγοριοποίησης (δες Διάλεξη 5ή) βασίζονται στο κανόνα του Bayes

$$p(c|\mathbf{x}) = \frac{p(\mathbf{x}|c)p(c)}{p(\mathbf{x})}$$

όπου

$$p(\mathbf{x}) = \sum_{c=1}^{C} p(\mathbf{x}|c)p(c)$$

Η $p(c|\mathbf{x})$ εκφράζει την εκ των υστέρων πιθανότητα (δηλ. αφού παρατηρηθεί ένα συγκεκριμένο $\mathbf{x}$) να ισχύει η κατηγορία c

Προκειμένου να μοντελοποιήσουμε την υπό συνθήκη κατανομή $p(\mathbf{x}|c)$ της κάθε κατηγορίας θα μπορούσαμε να χρησιμοποιήσουμε μια μίξη, δηλ.

$$p(\mathbf{x}|c) \text{ μοντελοποιείται από } \sum_{k=1}^{K_c} \pi_k^{(c)} p(\mathbf{x}|\boldsymbol{\theta}_k^c)$$

Κατηγοριοποίηση με μίξη κατανομών

Παράδειγμα εφαρμογής στους χειρόγραφους χαρακτήρες

Στα πρόβλημα αυτό έχουμε $C = 10$ κατηγορίες (ψηφία). Για κάθε κατηγορία εισάγουμε μια μίξη από Bernoulli

$$\sum_{k=1}^{K_c} \pi_k^{(c)} \prod_{d=1}^{784} (\mu_{kd}^{(c)})^{x_d} (1 - \mu_{kd}^{(c)})^{1-x_{nd}}$$

την οποία την εκπαιδεύουμε με τον αλγόριθμο EM χρησιμοποιώντας τα δεδομένα μόνο της κατηγορίας αυτής (του ψηφίου στην προκειμένη περίπτωση) που υπάρχουν στο σύνολο εκπαίδευσης

- $\Rightarrow$ που προκύπτει από την τεχνική της μέγιστης πιθανοφάνειας για κάθε περιγραφικό σύστημα κατηγοριοποίησης (δες Διάλεξη 5ή)

Έπειτα μπορούμε να κατηγοριοποιούμε άγνωστα δεδομένα υπολογίζοντας τις εκ των υστέρων πιθανότητες των κατηγορίων

Στην περίπτωση που χρησιμοποιούμε μια συνιστώσα κατανομή για κάθε κατηγορία, δηλ. $K_c = 1,\ c = 1, \ldots, C$ το παραπάνω σύστημα γίνεται ο naive Bayes!

Κατηγοριοποίηση με μίξη κατανομών

Παράδειγμα εφαρμογής στους χειρόγραφους χαρακτήρες

Κατασκευάσαμε τα ακόλουθα 3 σύστηματα κατηγοριοποίησης

1. 5 συνιστώσες για κάθε υπό συνθήκη κατανομή, δηλ. $K_c = 5, c = 1, \ldots, 10$. Η επίδοση στα δεδομένα ελέγχου ήταν

$$error = 8,6\%$$

2. 10 συνιστώσες για κάθε υπό συνθήκη κατανομή, δηλ. $K_c = 10, c = 1, \ldots, 10$. Η επίδοση στα δεδομένα ελέγχου ήταν

$$error = 7,37\%$$

3. 20 συνιστώσες για κάθε υπό συνθήκη κατανομή, δηλ. $K_c = 20, c = 1, \ldots, 10$. Η επίδοση στα δεδομένα ελέγχου ήταν

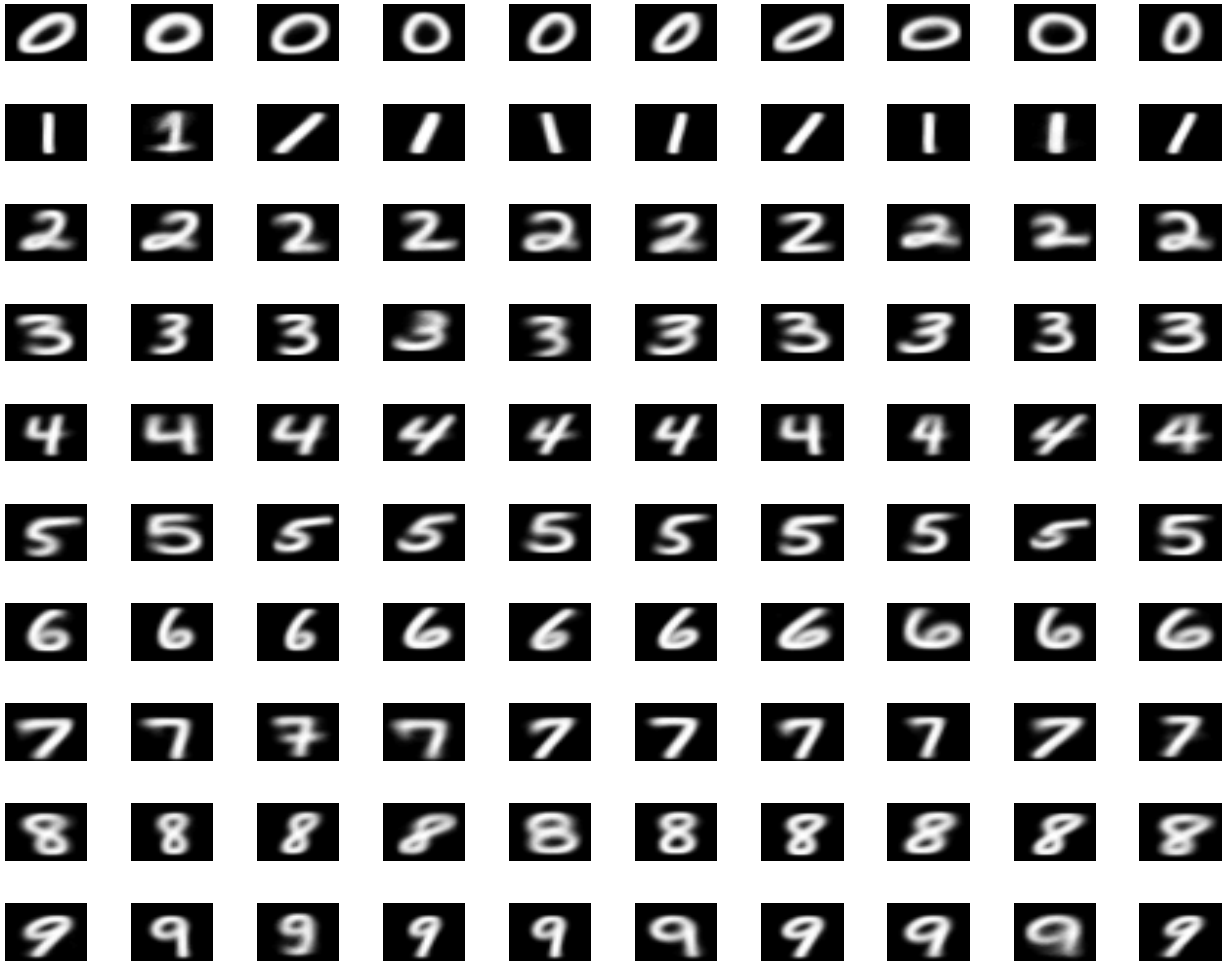
$$error = 6,49\%$$

Ο naive Bayes, που αντιστοιχεί στην περίπτωση $K_c = 1, c = 1, \ldots, 10$, είχε σφάλμα

$$error = 15.83\%$$

Κατηγοριοποίηση με μίξη κατανομών

Παράδειγμα εφαρμογής στους χειρόγραφους χαρακτήρες



Οι όμαδες που βρέθηκαν από την εφαρμογή των EM αλγορίθμων (για κάθε ψηφίο) στην περίπτωση $K_c = 10, c = 1, \ldots, 10$

Αριθμητική ευστάθεια

Έστω ότι έχουμε να υπολογίσουμε μια εκ των υστέρων πιθανότητα της μορφής

$$p = \frac{e^{-1000}}{e^{-1000} + e^{-1000} + e^{-1000}}$$

Προφανώς η απάντηση είναι $p = \frac{1}{3}$. Ωστόσο αν προγραμματίσουμε την παραπάνω έκφραση στον MATLAB (ή οποιαδήποτε άλλη γλώσσα), παίρνουμε

```
>> p = exp(-1000)/(exp(-1000) + exp(-1000) + exp(-1000))
p =
    NaN
```

Ομοίως αν είχαμε

```
>> p = exp(1000)/(exp(1000) + exp(1000) + exp(1000))
p =
    NaN
```

Τι φταίει στην πρώτη και τι στην δεύτερη περίπτωση;

Αριθμητική ευστάθεια

Έστω ότι έχουμε να υπολογίσουμε μια εκ των υστέρων πιθανότητα της μορφής

$$p = \frac{e^{-1000}}{e^{-1000} + e^{-1000} + e^{-1000}}$$

Προφανώς η απάντηση είναι $p = \frac{1}{3}$. Ωστόσο αν προγραμματίσουμε την παραπάνω έκφραση στον MATLAB (ή οποιαδήποτε άλλη γλώσσα), παίρνουμε

```
>> p = exp(-1000)/(exp(-1000) + exp(-1000) + exp(-1000)) = 0/0
p =
    NaN
```

Ομοίως αν είχαμε

```
>> p = exp(1000)/(exp(1000) + exp(1000) + exp(1000)) = Inf/Inf
p =
    NaN
```

Πώς θα μπορούσαμε να αποφύγουμε την διαίρεση με 0 ή *Inf*;

Αριθμητική ευστάθεια

Θα μπορούσαμε να αφαιρέσουμε (ή ισοδύναμα προσθέσουμε) μέσα στο όρισμα του κάθε εκθετικού (και έπειτα να πάρουμε το εκθετικό!) ένα κατάλληλο αριθμό m

$$p = \frac{e^{-1000-m}}{e^{-1000-m} + e^{-1000-m} + e^{-1000-m}}$$

αν επιλέγαμε $m = -1000$, θα είχαμε

$$p = \frac{e^{-1000+1000}}{e^{-1000+1000} + e^{-1000+1000} + e^{-1000+1000}} = \frac{e^{0}}{e^{0} + e^{0} + e^{0}}$$

το οποίο αν προγραμματίζαμε στο MATLAB θα έδινε

```
>> p = exp(0)/(exp(0) + exp(0) + exp(0))
p =
    0.3333
```

Αριθμητική ευστάθεια

Γενικότερα θα εργαζόμασταν ως εξής. Αν είχαμε να υπολογίσουμε (στον υπολογιστή!) την εκ των υστέρων πιθανότητα

$$p_k = \frac{e^{f_k}}{\sum_{j=1}^{K} e^{f_j}}$$

θα βρίσκαμε πρώτα

$$m = max(f_1, \ldots, f_K)$$

και έπειτα θα υπολογίζαμε

$$p_k = \frac{e^{f_k - m}}{\sum_{j=1}^{K} e^{f_j - m}}$$

Με αυτόν τον τρόπο δεν πρόκειται ποτέ να διαιρέσουμε με 0 ή Inf, διότι ο όρος στον παρανομαστή θα είναι πάντα της μορφής

$$\sum_{j=1}^{K} e^{f_j - m} = 1 + \sum_{j \neq j_*} e^{f_j - m}$$

όπου $e^{f_j - m} \leq 1, j \neq j_*$ και $m = f_{j_*}$

Αριθμητική ευστάθεια

Τέτοιοι υπολογισμοί εμφανίζονται συχνά στους EM αλγορίθμους (αλλά και γενικότερα). Π.χ. στον EM για την ομαδοποίηση των χειρόγραφων ψηφίων υπολογίζουμε στο E βήμα

$$\gamma(z_{nk}) = \frac{\pi_k \prod_{d=1}^{784} \mu_{kd}^{x_{nd}} (1-\mu_{kd})^{1-x_{nd}}}{\sum_{j=1}^{K} \pi_j \prod_{d=1}^{784} \mu_{jd}^{x_{nd}} (1-\mu_{jd})^{1-x_{nd}}}$$

Ο παρανομαστής εύκολα μπορεί στην ακρίβεια του υπολογιστή να γίνει 0. Ωστόσο η ποσότητα αυτή γράφεται ως

$$\gamma(z_{nk}) = \frac{e^{\log \pi_k + \sum_{d=1}^{784} x_{nd} \log \mu_{kd} + (1-x_{nd}) \log(1-\mu_{kd})}}{\sum_{j=1}^{K} e^{\log \pi_j + \sum_{d=1}^{784} x_{nd} \log \mu_{jd} + (1-x_{nd}) \log(1-\mu_{jd})}}$$

Οπότε αν θέσουμε (και αυσιαστικά είναι αυτό που θα προγραμματίσουμε/αποθηκεύσουμε!)

$$f_j = \log \pi_j + \sum_{d=1}^{784} x_{nd} \log \mu_{jd} + (1-x_{nd}) \log(1-\mu_{jd}), \quad j = 1, \ldots, K$$

τότε μπορούμε να εφαρμόσουμε το trick που περιγράφτηκε προηγουμένως

Αριθμητική ευστάθεια

logsumexp trick

Επίσης η λογαριθμική πιθανοφάνεια (την οποία θα πρέπει να υπολογίζουμε για debugging και έλεγχο σύγκλισης) στον EM έχει την μορφή

$$L(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k p(\mathbf{x}_n|k)$$

Λόγω ότι ο όρος $\sum_{k=1}^{K} \pi_k p(\mathbf{x}_n|k)$ εύκολα μπορεί να γίνει (στην ακρίβεια του υπολογιστή μιλάμε πάντα) 0 ή Inf, μια απλοική υλοποίηση θα οδηγούσε σε $\log 0 = -Inf$ και $\log Inf = NaN$. Για να αντιμετωπίσουμε αυτό το πρόβλημα εφαρμόζουμε ένα πολύ όμοιο trick. Παρατηρούμε ότι $\log \sum_{k=1}^{K} \pi_k p(\mathbf{x}_n|k)$ μπορεί να γραφεί στη μορφή

$$\log \sum_{k=1}^{K} e^{f_k}$$

Οπότε έπειτα θα βρίσκαμε το $m = max(f_1, \ldots, f_K)$, θα γράφαμε

$$\log \sum_{k=1}^{K} e^{f_k+m-m} = \log \sum_{k=1}^{K} e^m e^{f_k-m} = \log e^m \sum_{k=1}^{K} e^{f_k-m} = m + \log \sum_{k=1}^{K} e^{f_k-m}$$

και θα προγραμματίζαμε την τελευταία σχέση!

Επίλογος

- Διάβασμα για το σπίτι: Bishop: ενότητα 9.3, 9.3.1, 9.3.3, 9.4
- Επόμενο μάθημα: Μοντέλα μείωσης διάστασης δεδομένων (PCA, probabilistic PCA, factor analysis και οι αντίστοιχοι EM αλγόριθμοι)

Παράρτημα: Η ιδιότητα σύγκλισης του ΕΜ

Υπάρχει μια εναλλακτική ισοδύναμη ερμήνεια του ΕΜ ως μεγιστοποίηση ενός κάτω φράγματος της λογαριθμικής πιθανοφανείας

- Χρησιμοποιώντας την **ανισότητα Jensen** (δες παρακάτω στο Παράρτημα) έχουμε

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log \sum_{\mathbf{z}_n} p(\mathbf{x}_n, \mathbf{z}_n|\boldsymbol{\theta}) \geq \sum_{\mathbf{z}_n} q(\mathbf{z}_n) \log \frac{p(\mathbf{x}_n, \mathbf{z}_n|\boldsymbol{\theta})}{q(\mathbf{z}_n)} = \mathcal{L}(\mathbf{q}, \boldsymbol{\theta})$$

όπου η $\mathcal{L}(\mathbf{q}, \boldsymbol{\theta})$ είναι ένα κάτω φράγμα (lower bound) για οποιοδήποτε $\boldsymbol{\theta}$ και κατανομές $\mathbf{q} = \{q(\mathbf{z}_n)\}_{n=1}^{N}$. Τα βήματα Ε και Μ του αλγορίθμου ισοδύναμα εκφράζονται ως:

1. Ε βήμα: (μεγιστοποίηση ως προς $\mathbf{q}$ δοθέντος $\boldsymbol{\theta}^{(t)}$)

$$\mathbf{q}^* = \arg\max_{\mathbf{q}} \mathcal{L}(\mathbf{q}, \boldsymbol{\theta}^{(t)})$$

2. Μ βήμα: (μεγιστοποίηση ως προς $\boldsymbol{\theta}$ δοθέντος $\mathbf{q}^*$)

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{q}^*, \boldsymbol{\theta})$$

Παράρτημα: Η ιδιότητα σύγκλισης του EM

1. Ε βήμα: (μεγιστοποίηση ως προς $\mathbf{q}$ δοθέντος $\boldsymbol{\theta}^{(t)}$)

$$\mathbf{q}^* = \arg\max_{\mathbf{q}} \mathcal{L}(\mathbf{q}, \boldsymbol{\theta}^{(t)})$$

2. Μ βήμα: (μεγιστοποίηση ως προς $\boldsymbol{\theta}$ δοθέντος $\mathbf{q}^*$)

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{q}^*, \boldsymbol{\theta})$$

Η μεγιστοποίηση στο Ε βήμα αναγκάζει το κάτω φράγμα $\mathcal{L}(\mathbf{q}, \boldsymbol{\theta}^{(t)})$ να γίνει tight, δηλ. $\mathcal{L}(\mathbf{q}, \boldsymbol{\theta}^{(t)}) = \mathcal{L}(\boldsymbol{\theta}^{(t)})$ το οποίο επιτυγχάνεται για $q^*(\mathbf{z}_n) = p(\mathbf{z}_n|\mathbf{x}_n, \boldsymbol{\theta}^{(t)}),\ n = 1, \ldots, N$ (το Ε βήμα όπως είχε οριστεί αρχικά)

Η μεγιστοποίηση στο Μ βήμα οδηγεί στο

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) + \mathcal{H}(\boldsymbol{\theta}^{(t)})$$

όπου $\mathcal{H}(\boldsymbol{\theta}^{(t)}) = -\sum_{n=1}^{N} \sum_{\mathbf{z}_n} p(\mathbf{z}_n|\mathbf{x}_n, \boldsymbol{\theta}^{(t)}) \log p(\mathbf{z}_n|\mathbf{x}_n, \boldsymbol{\theta}^{(t)})$ είναι σταθέρα (δεν εξαρτάται από το $\boldsymbol{\theta}$). Οπότε το παράπανω ισοδυναμεί με την μεγιστοποίηση της $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$ (το Μ βήμα όπως είχε οριστεί αρχικά)

Παράρτημα: Η ιδιότητα σύγκλισης του ΕΜ

- Το Ε βήμα οδηγεί πάντα σε ένα κάτω φράγμα το οποίο είναι ακριβώς ίσο (tight) με την λογαριθμική πιθανοφάνεια για την τρέχουσα τιμή των παραμέτρων $\boldsymbol{\theta}^{(t)}$, δηλ.

$$\mathcal{L}(\mathbf{q}^*, \boldsymbol{\theta}^{(t)}) = \mathcal{L}(\boldsymbol{\theta}^{(t)})$$

- Έπειτα το Μ βήμα μεγιστοποιεί την $\mathcal{L}(\mathbf{q}^*, \boldsymbol{\theta})$ ως προς $\boldsymbol{\theta}$ ώστε προκύπτει ένα νέο $\boldsymbol{\theta}^{(t+1)}$ τέτοιο ώστε

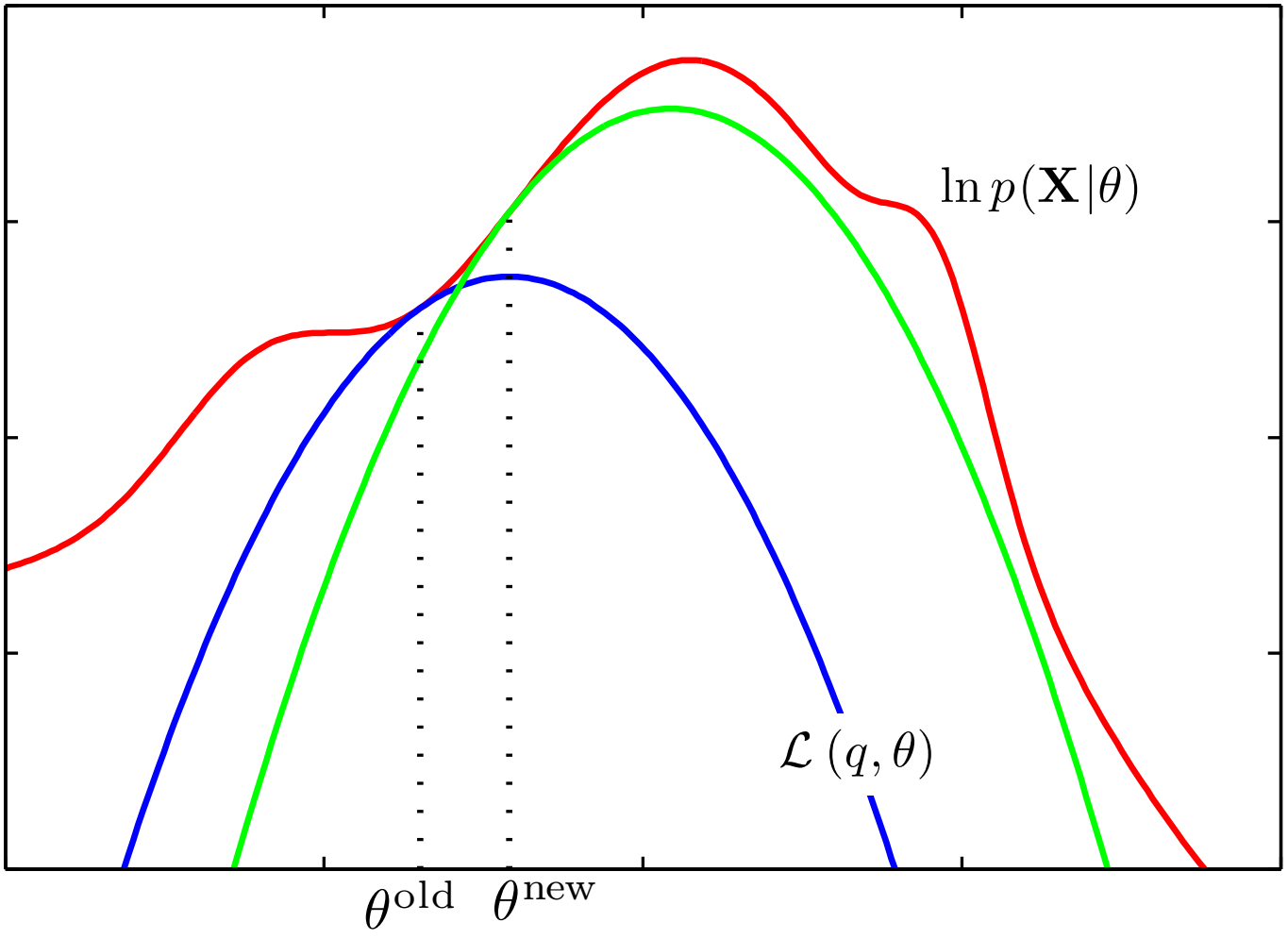
$$\mathcal{L}(\mathbf{q}^*, \boldsymbol{\theta}^{(t+1)}) \geq \mathcal{L}(\mathbf{q}^*, \boldsymbol{\theta}^{(t)}) = \mathcal{L}(\boldsymbol{\theta}^{(t)})$$

- Ωστόσο το $\mathcal{L}(\mathbf{q}^*, \boldsymbol{\theta}^{(t+1)})$ είναι πάντα κάτω φράγμα, δηλ.

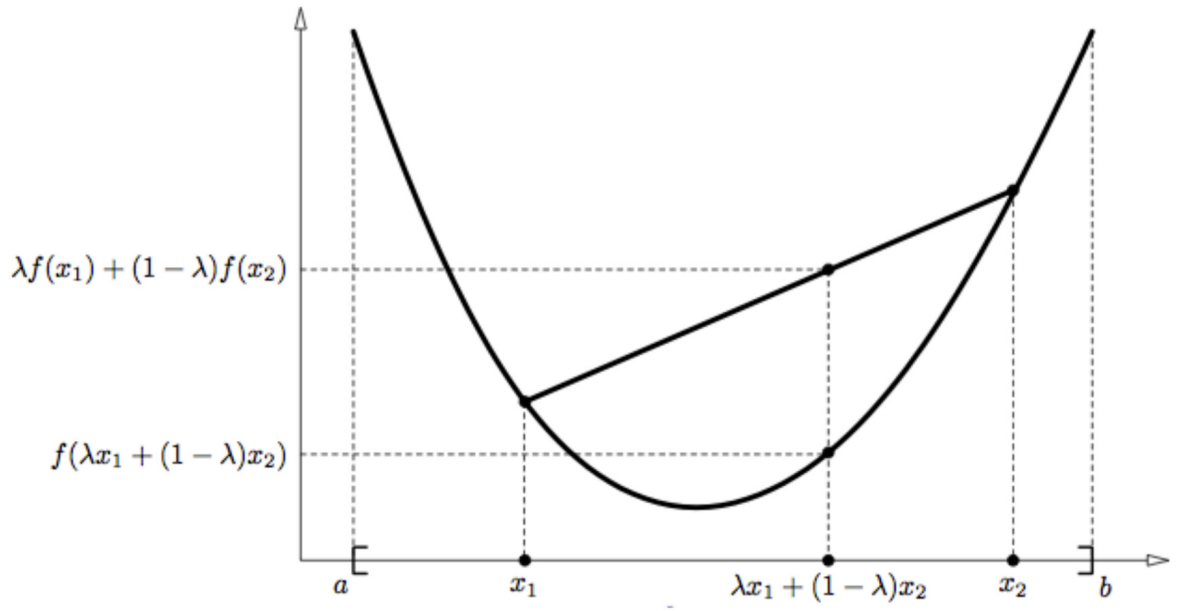
$$\mathcal{L}(\boldsymbol{\theta}^{(t+1)}) \geq \mathcal{L}(\mathbf{q}^*, \boldsymbol{\theta}^{(t+1)})$$

- Οπότε συμπεραίνουμε ότι $\mathcal{L}(\boldsymbol{\theta}^{(t+1)}) \geq \mathcal{L}(\boldsymbol{\theta}^{(t)})$

Παράρτημα: Η ιδιότητα σύγκλισης του EM



Παράρτημα: Η ανισότητα Jensen

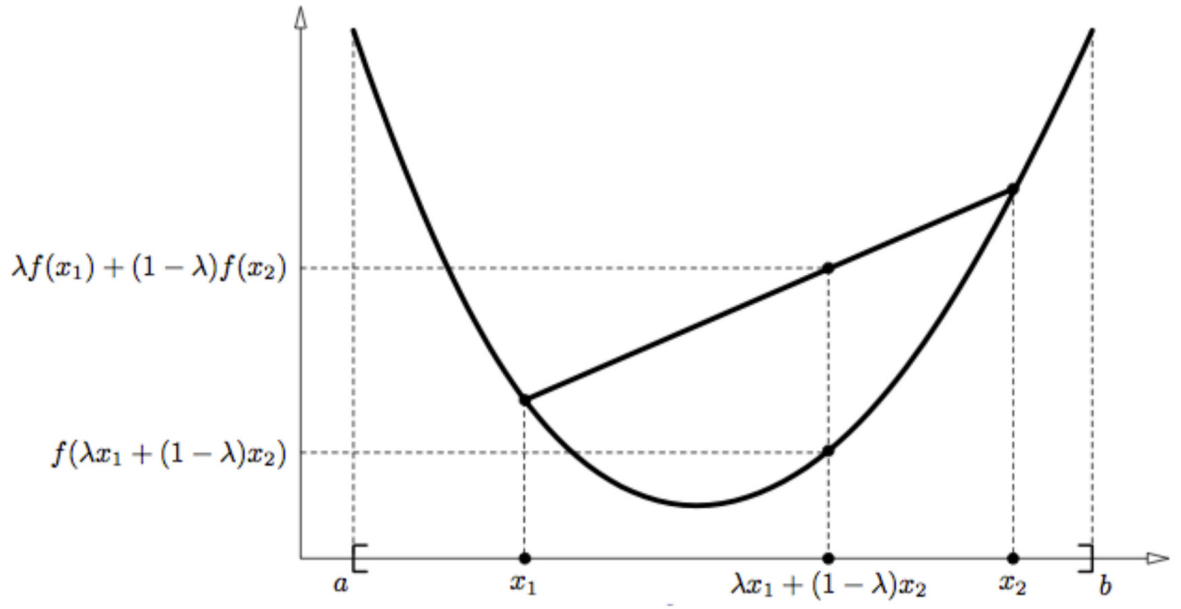


Αν μια συνάρτηση $f(x)$ είναι κυρτή (convex) τότε εξ ορισμού ισχύει

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

για κάθε x_1, x_2 (στο πεδίο ορισμού της συνάρτησης) και $\lambda \in [0, 1]$

Παράρτημα: Η ανισότητα Jensen

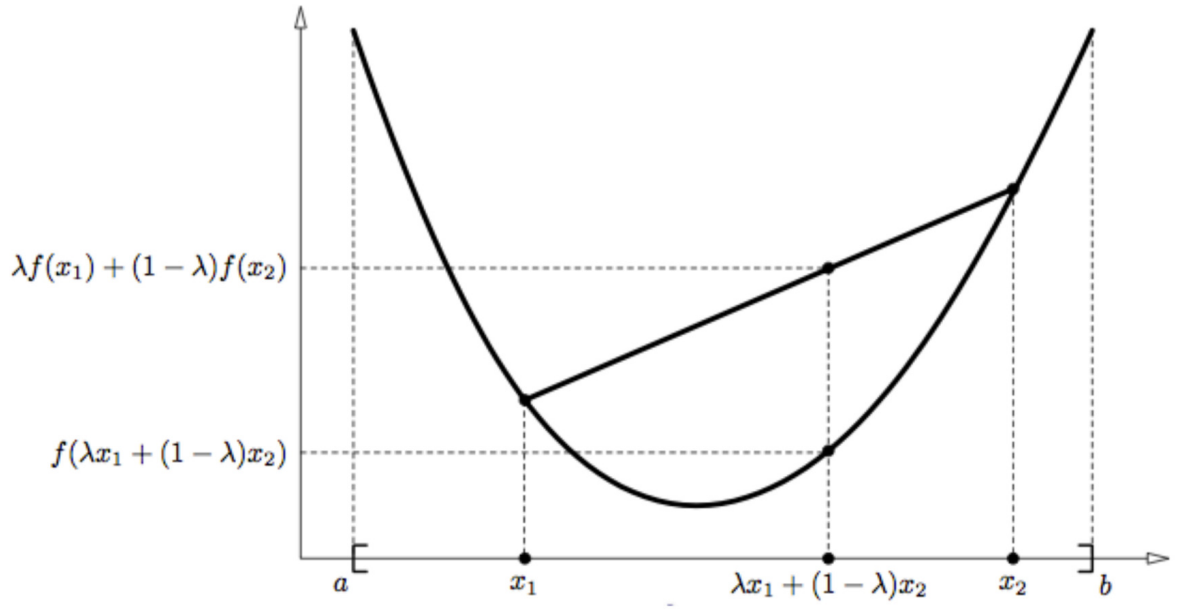


Με απλή απόδειξη (μέσω της μεθόδου της επαγωγής) η ανισότητα αυτή γενικεύεται ώστε

$$f\left(\sum_{k=1}^{K} \lambda_k x_k\right) \leq \sum_{k=1}^{K} \lambda_k f(x_k) \quad \Rightarrow f\left(\mathbb{E}[x]\right) \leq \mathbb{E}[f(x)]$$

όπου $\lambda_k \leq 0$ και $\sum_{k=1}^{K} \lambda_k = 1$

Παράρτημα: Η ανισότητα Jensen

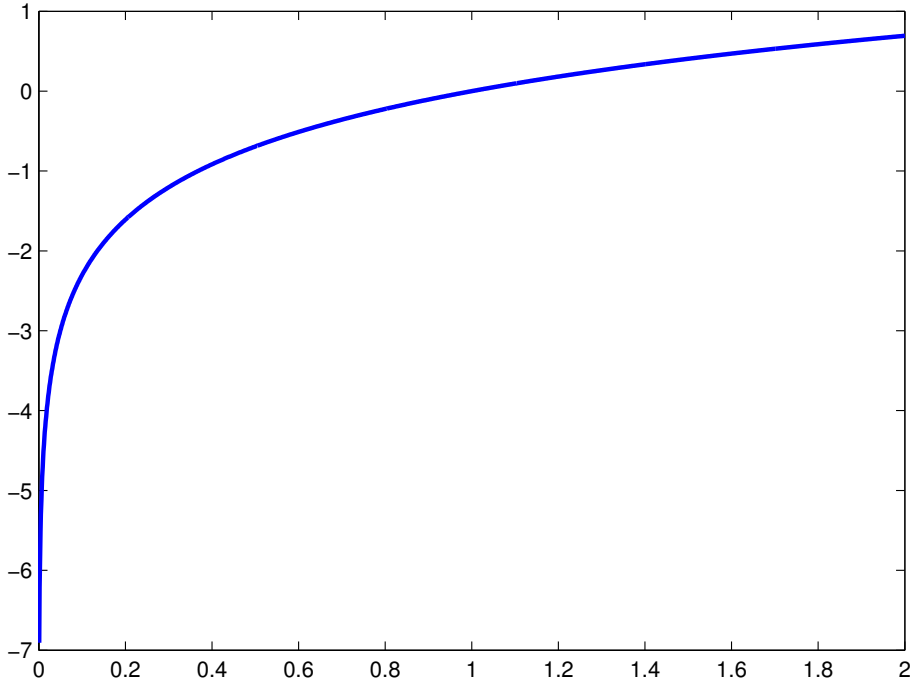


Πιο γενικά μπορεί να δειχθεί αν η $f(x)$ είναι κυρτή, τότε για οποιαδήποτε κατανομή $q(x)$ (δηλ. $q(x) \geq 0$, $\int q(x)dx = 1$) ισχύει

$$f\left(\int q(x)x dx\right) \leq \int q(x)f(x)dx \quad \Rightarrow f\left(\mathbb{E}[x]\right) \leq \mathbb{E}[f(x)]$$

(προφανώς ισχύει επίσης όταν η q είναι διακριτή και το ολοκλήρωμα $\int$ γίνεται άθροισμα $\sum$)

Παράρτημα: Η ανισότητα Jensen



Σχήμα: Η λογαριθμική συνάρτηση

Αν η συνάρτηση είναι κοίλη (concave), τότε η ανισότητα αντιστρέφεται. Π.χ. ο λογάριθμος $\log(x)$ είναι μια κοίλη συνάρτηση και η ανισότητα Jensen μας λέει ότι

$$\log\left(\int q(x)x dx\right) \geq \int q(x)\log(x)dx$$

Παράρτημα: Ομαδοποίηση κειμένων με μίξεις από κατηγορικές κατανομές

Έστω οτι θέλουμε να ομαδοποιήσουμε κείμενα (οι ομάδες σε ένα τέτοιο πρόβλημα ονομάζονται και topics) και βασιζόμαστε σε αναπαράσταση κειμένων με ένα λεξιλόγιο (σύνολο από λέξεις κλειδιά), το οποίο θα μπορούσε π.χ. να αποτελείται από τις ακόλουθες λέξεις

(πολιτική, αθλητισμός, φοιτητής, βιβλίο, διασκέδαση)

Κάθε κείμενο n αναπαριστάται από ένα σύνολο L_n λέξεων

$$\mathbf{x}_n = (x_{n1}, \ldots, x_{nL_n})$$

όπου η κάθε λέξη $x_{n\ell}$ παίρνει μια τιμή από το λεξιλόγιο που έχουμε καθορίσει (οι λέξεις του αρχικού κειμένου που είναι εκτός του λεξιλογίου δεν χρησιμοποιούνται). Π.χ.

$$x_{n\ell} = \text{διασκέδαση}$$

Στην πράξη το λεξιλόγιο αποτελείται από D λέξεις και D μπορεί να είναι κάποιες χιλιάδες

Παράρτημα: Ομαδοποίηση κειμένων με μίξεις από κατηγορικές κατανομές

Θα θέλαμε να ομαδοποιήσουμε τα κείμενα ώστε η κάθε ομάδα/topic να εκφράζει την συχνότητα εμφάνισης λέξεων από το λεξιλόγιο. Υπό αυτό το πρίσμα η κάθε ομάδα περιγράφεται από ένα διάνυσμα πιθανοτήτων

$$\boldsymbol{\theta}_k = (\theta_{k1}, \ldots, \theta_{kD}), \quad \sum_{d=1}^{D} \theta_{kd} = 1$$

όπου θ_{kd} εκφράζει την συχνότητα εμφάνισης της λέξης d σε κείμενα της ομάδας k. Έπειτα η συνιστώσα κατανομή που περιγράφει το τρόπο που παράγεται το διάνυσμα $\mathbf{x}_n$ έχει τη μορφή

$$p(\mathbf{x}_n|k) = \prod_{\ell=1}^{L_n} p(x_{n\ell}|k) = \prod_{\ell=1}^{L_n} \prod_{d=1}^{D} \theta_{kd}^{x_{n\ell d}}$$

όπου αναπαριστούμε το κάθε $x_{n\ell}$ ως D-διάσταοτο διάνυσμα υπόδειξης. Παρατηρούμε ότι η παραπάνω σχέση γράφεται επίσης ως

$$p(\mathbf{x}_n|k) = \prod_{d=1}^{D} \theta_{kd}^{\sum_{\ell=1}^{L_n} x_{n\ell d}} = \prod_{d=1}^{D} \theta_{kd}^{\eta_{nd}}$$

όπου $\eta_{nd} = \sum_{\ell=1}^{L_n} x_{n\ell d}$ είναι ο αριθμός των φορών εμφάνισης της λέξης d στο κείμενο n (word count)

Παράρτημα: Ομαδοποίηση κειμένων με μίξεις από κατηγορικές κατανομές

Οπότε το μοντέλο της μίξης παίρνει τη μορφή

$$p(\mathbf{x}_n) = \sum_{k=1}^{K} \pi_k \prod_{d=1}^{D} \theta_{kd}^{\eta_{nd}}$$

το οποίο έχει παραμέτρους $\{\pi_k, \theta_{kd}\}$. Επιπλέον η λογαριθμική πιθανοφάνεια που θα θέλαμε να μεγιστοποιήσουμε με τον αλγόριθμο ΕΜ έχει τη μορφή

$$\mathcal{L} = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \prod_{d=1}^{D} \theta_{kd}^{\eta_{nd}}$$

Άσκηση: Βγάλε τις εξισώσεις του ΕΜ αλγορίθμου.