ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΕΡΓΑΣΤΗΡΙΟ ΕΠΕΞΕΡΓΑΣΙΑΣ ΣΗΜΑΤΩΝ ΚΑΙ
ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

Θεματα Ορασης Υπολογιστων

Αναλυση Κυριων Σινιστωσων-PCA
Autoencoders-AE
Και
Variational Autoencoders-VAE

Διδάσκων: Αναπλ. Καθηγητής Εμμανουήλ Ζ. Ψαράκης
Επικουρικό έργο: Αριστείδης Μπίφης, Παναγιώτης Κάτσος

Πάτρα Δεκέμβριος 2021

ΣΤΟΙΧΕΙΩΔΗΣ ΘΕΩΡΙΑ

ΑΝΑΛΥΣΗ ΚΥΡΙΩΝ ΣΥΝΙΣΤΩΣΩΝ

Η ανάλυση κυρίων συνιστωσών είναι η διαδικασία υπολογισμού των κύριων συνιστωσών (principal components) και η χρήση τους για την δημιουργία μιας βάσης από τα δεδομένα, ορισμένες φορές χρησιμοποιώντας μόνο μερικές από τις μεγαλύτερες κύριες συνιστώσες, αγνοώντας τις υπόλοιπες.

Η ανάλυση κυρίων συνιστωσών (Principal Component Analysis-PCA) είναι μια τεχνική που χρησιμοποιείται σε εφαρμογές όπως:

- στην μείωση της διαστατικότητας των δεδομένων,
- στην εξαγωγή χαρακτηριστικών και κατά συνέπεια στην ανάλυση δεδομένων,
- στην οπτικοποίηση των δεδομένων και
- στη δημιουργία μοντέλων πρόβλεψης.

Η PCA είναι επίσης γνωστή ως ο μετασχηματισμός των Karhunen-Loeve.

Η PCA μπορεί να ορισθεί ισοδύναμα με τους ακόλουθους δύο τρόπους:

- ως η ορθογώνια προβολή των δεδομένων πάνω σε ένα γραμμικό χώρο μικρότερης διάστασης από αυτό των δεδομένων, ο οποίος είναι γνωστός ως **κύριος υπόχωρος** (principal subspace), με τρόπο ώστε να μεγιστοποιείται η διασπορά των προβολών των δεδομένων στην μείωση της διαστατικότητας των δεδομένων. Αν ορίσουμε την PCA με αυτό τον τρόπο, το πρώτο **κύριο στοιχείο** (principal component) αποτελεί την κατεύθυνση που μεγιστοποιεί τη διακύμανση των προβληθέντων δεδομένων. Γενικεύοντας, το m-οστό **κύριο στοιχείο** μπορεί να θεωρηθεί ως μια κατεύθυνση ορθοκανονική ως προς τα προηγούμενα $m-1$ **κύρια στοιχεία** που μεγιστοποιεί τη διακύμανση των προβληθέντων δεδομένων.
- ως η γραμμική προβολή η οποία ελαχιστοποιεί το μέσο κόστος προβολής, το οποίο ορίζεται ως η μέση τετραγωνική απόσταση ανάμεσα

στα δεδομένα και τις αντίστοιχες προβολές τους. Χρησιμοποιώντας τον ισοδύναμο αυτό ορισμό ορίζουμε μια ακολουθία μοναδιαίων διανυσμάτων, όπου:

- το m-οστό διάνυσμα, είναι η κατεύθυνση μιας ευθείας η οποία περιγράφει καλύτερα τα δεδομένα και
- είναι ορθογώνια με τα υπόλοιπα m-1 διανύσματα. Η ευθεία αυτή ορίζεται, ελαχιστοποιώντας τη μέση τετραγωνική απόσταση της από τα σημεία. Οι κατευθύνσεις αυτές διαμορφώνουν μια ορθοκανονική βάση, στην οποία, οι διαφορετικές ατομικές διαστάσεις των δεδομένων είναι γραμμικά ασυσχέτιστες.

Τους δύο παραπάνω διαφορετικούς, αλλά ισοδύναμους, ορισμούς της PCA θα αναλύσουμε στην συνέχεια.

Μειωση Διαστατικοτητας των Δεδομενων

Όπως αναφέραμε παραπάνω, η PCA μπορεί να θεωρηθεί ως ένας ορθοκανονικός γραμμικός μετασχηματισμός των δεδομένων, σε ένα νέο σύστημα συντεταγμένων, τέτοιο ώστε η μέγιστη διακύμανση, από κάποια βαθμωτή προβολή των δεδομένων, εμπίπτει στην πρώτη συντεταγμένη και καλείται το πρώτο κύριο στοιχείο, η δεύτερη μεγαλύτερη διακύμανση το δεύτερο κύριο στοιχείο, κ.ο.κ..

Για το σκοπό αυτό, έστω:

$$\mathcal{X} = \{\mathbf{x}_m\}_{m=1}^{M}$$

το σύνολο των δεδομένων μας, με πληθικό αριθμό M και $\mathbf{x}_m$ μία διανυσματική ευκλείδια μεταβλητή στο χώρο $\mathbb{R}^N$. Το σύνολο των δεδομένων μπορούμε να το τοποθετήσουμε σε ένα μητρώο X διάστασεων $N \times M$.

Σκοπός μας τώρα είναι να προβάλλουμε τα δεδομένα σε ένα χώρο διάστασης L, με $L < N$, ενώ ταυτόχρονα θέλουμε να μεγιστοποιήσουμε την διασπορά των προβληθέντων δεδομένων.

Θεωρούμε ότι:

- το L μας δίδεται[1],
- τα δεδομένα είναι μηδενικής μέσης τιμής[2], δηλαδή:

$$\bar{\mathbf{x}} = \frac{1}{M}\sum_{m=1}^{M}\mathbf{x}_m = \mathbf{0}$$

και

- το μητρώο συνδιασπορών των δεδομένων, δηλαδή:

$$S = XX^T$$

είναι γνωστό.

Για να αρχίσουμε, ας υποθέσουμε ότι θέλουμε να λύσουμε το πρόβλημα για $L = 1$. **Αποδείξτε** ότι η επιθυμητή λύση προκύπτει από τη λύση του ακόλουθου με περιορισμούς προβλήματος βελτιστοποίησης:

$$\begin{aligned}&\max_{\mathbf{v}_1}\mathbf{v}_1^t S\mathbf{v}_1\\ &\text{subject to: } ||\mathbf{v}_1||_2 = 1\end{aligned}$$

όπου $||\mathbf{z}||_2$, η l_2 στάθμη του διανύσματος z.

Αποδείξτε επίσης ότι η λύση του παραπάνω προβλήματος είναι το ιδιοδιάνυσμα $\mathbf{v}_1^\star$ που αντιστοιχεί στην μέγιστη ιδιοτιμή (έστω λ_1) του μητρώου S[3].

Ας υποθέσουμε τώρα ότι θέλουμε να λύσουμε το πρόβλημα για $L = 2$. Σε αυτή την περίπτωση **αποδείξτε** ότι η επιθυμητή λύση προκύπτει από τη λύση του ακόλουθου, με περιορισμούς προβλήματος βελτιστοποίησης:

$$\begin{aligned}&\max_{\mathbf{v}_2}\mathbf{v}_2^t S\mathbf{v}_2\\ &\text{subject to: } ||\mathbf{v}_2||_2 = 1\\ &\qquad\qquad < \mathbf{v}_1^\star,\ \mathbf{v}_2 >= 0\end{aligned}$$

[1] Αν δεν μας δίδεται το L, πως θα μπορούσαμε να το εκτιμήσουμε; Δικαιολογήστε την απάντησή σας.

[2] Αν δεν είναι τι μπορούμε να κάνουμε;

[3] Παρατηρήστε ότι, το μητρώο S είναι εξ ορισμού ένα μη αρνητικά ορισμένο μητρώο.

όπου $||\mathbf{z}||_2$, η l_2 στάθμη του διανύσματος z και $< ., . >$ ο τελεστής εσωτερικού γινομένου.

Αποδείξτε επίσης ότι η λύση του παραπάνω προβλήματος είναι το ιδιοδιάνυσμα $\mathbf{v}_2^\star$ που αντιστοιχεί στην δεύτερη μεγαλύτερη ιδιοτιμή (έστω λ_2) του μητρώου S.

Χρησιμοποιώντας την αρχή της επαγωγής, **διατυπώστε** το τελικό ($L = L$ πρόβλημα βελτιστοποίησης και εξηγείστε αναλυτικά τις επιλογές σας.

Ας δούμε τώρα την PCA ως τη γραμμική προβολή των δεδομένων μας, η οποία ελαχιστοποιεί το μέσο κόστος προβολής, το οποίο ορίζεται ως η μέση τετραγωνική απόσταση ανάμεσα στα δεδομένα και τις αντίστοιχες προβολές τους.

Η ανάλυση κυρίων συνιστωσών συνδέεται άμεσα με μια άλλη τεχνική παραγοντοποίησης μητρώων, την **ανάλυση ιδιάζουσων τιμών** (SVD) του μητρώου X, δηλαδή:

$$X = V\Sigma U^T,$$

όπου:

- Σ είναι ένα διαγώνιο $N \times M$ μητρώο μη αρνητικών αριθμών $\sigma_k, k = 1, 2, \cdots, \max\{N, M\}$ οι οποίες ονομάζονται ιδιάζουσες τιμές του μητρώου X,
- V είναι ένα $N \times N$ ορθοκανονικό μητρώο, του οποίου η $n-$οστή στήλη αποτελεί το $n-$οστό αριστερό ιδιάζων διάνυσμα του μητρώου X,
- U είναι ένα $M \times M$ ορθοκανονικό μητρώο, του οποίου η $m-$οστή στήλη αποτελεί το $m-$οστό δεξιό ιδιάζων διάνυσμα του μητρώου X.

Παίρνοντας υπόψην μας τα παραπάνω, το μητρώο συνδιασπορών $S = XX^T$ μπορεί να γραφτεί ως:

$$\begin{aligned} S &= XX^T = V\Sigma\Sigma^T V^T \\ &= V\Sigma_0 V^T \end{aligned}$$

όπου Σ_0 ένα $N \times N$ τετραγωνικό διαγώνιο μητρώο των ιδιαζουσών τιμών του X.

Το προβλημα μας τώρα μπορεί να διατυπωθεί ως το ακόλουθο πρόβλημα βελτιστοποίησης:

$$\min_{A} ||S - A||_F^2$$
subject to: rank{A}=L

όπου $||C||_F$, η Frobenius στάθμη του μητρώου C [4].

Αποδείξτε ότι η λύση του παραπάνω προβλήματος ελαχιστοποίησης είναι η ακόλουθη:

$$A^\star = V_L \Sigma_L V_L^T$$

όπου:

- V_L οι πρώτες L στήλες του μητρώου V και
- Σ_L ένα $L \times L$ διαγώνιο μητρώο που περιέχει τις πρώτες L τιμές της διαγωνίου του μητρώου Σ_0.

ΔΙΑΔΙΚΑΣΙΑ

1. Πειραματιστείτε χρησιμοποιώντας το σύνολο δεδομένων MNIST[5]. Συγκεκριμένα, γιά τουλάχιστον δύο (2) ψηφία της αρεσκείας σας, υπολογίστε:
   (α΄) το μέσο ψηφίο
   (β΄) το μητρώο συνδιασπορών
   (γ΄) τις οκτώ (8) πρώτες κύριες συνιστώσες και
   (δ΄) τις ανακατασκευές του ψηφίου για $L = 1,\ 8,\ 16,\ 64$ και 256 αντίστοιχα. Σχεδιάστε και σχολιάστε κατάλληλα τα αποτελέσματά σας.

[4] **Αποδείξτε** ότι η Frobenius στάθμη ενός μητρώου, ισούται με την l_2 στάθμη του διανύσματος που προκύπτει από τις γραμμές ή τις στήλες.

[5] Το σύνολο δεδομένων MNIST περιλαμβάνει εικόνες από χειρόγραφα ψηφία μεγέθους 28×28, σε μορφή διανυσμάτων διάστασης 784. Το σύνολο δεδομένων εκπαίδευσης περιλαμβάνει 60000 διανύσματα και το σύνολο δεδομένων ελέγχου 10000

(ε΄) Σχεδιάστε τα ιστογράμματα των σφαλμάτων καθώς και το σφάλμα ανακατασκευής για κάθε ένα ψηφίο από αυτά που επιλέξατε.

2. Πάνω στο σύνολο δεδομένων εκπαίδευσης του MNIST, το οποίο σας δίνεται, εφαρμόστε αρχικά PCA και υπολογίστε το μητρώο V_L με $L = 128$.

   (α΄) Ανακατασκευάστε, με το περικομμένο αυτό μητρώο, τα συμπιεσμένα (compressed) πλέον δεδομένα εκπαίδευσης.

   (β΄) Χρησιμοποιώντας αυτό το μητρώο που υπολογίσατε, παράξτε και τα συμπιεσμένα δεδομένα ελέγχου και παρουσιάστε χαρακτηριστικά ζεύγη εικόνων (αρχική και συμπιεσμένη εικόνα).

PCA Βασισμενη σε Πυρηνες

Ας υποθέσουμε τώρα ότι διαθέτουμε έναν μη γραμμικό μετασχηματισμό $k(\mathbf{x}) : \mathbb{R}^N \to \mathbb{R}^L$ και επομένως κάθε δεδομένο $\mathbf{x}_m$ απεικονίζεται (προβάλλεται) σε ένα σημείο $k(\mathbf{x}_m)$.

Μπορούμε τώρα να εφαρμόσουμε την τεχνική PCA στο χώρο των χαρακτηριστικών. Είναι προφανές ότι αν:

$$k(X) = [k(\mathbf{x}_1)\ k(\mathbf{x}_2)\ \cdots\ k(\mathbf{x}_M)] \tag{1}$$

είναι το μετασχηματισμένο $L \times M$ μητρώο X μπορούμε να ορίσουμε το μητρώο:

$$S_k = k(X)k(X)^T.$$

και να εφαρμόσουμε την κλασσική PCA.

Καταγράψτε τις βασικές διαφορές από την κλασσική τεχνική. Εμμέσως υποθέσαμε ότι η αναμενόμενη τιμή των μετασχηματισμένων δεδομένων μας είναι μηδενική. Τι συμβαίνει όταν αυτή η υπόθεση δεν ισχύει· Πειραματιστείτε χρησιμοποιώντας το σύνολο δεδομένων MNIST και τον μη γραμμικό μετασχηματισμό:

$$k(\mathbf{x}) = e^{-\frac{||\mathbf{x}||_2^2}{0.1}}. \tag{2}$$

Διαδικασια

1. Γιά τουλάχιστον δύο (2) ψηφία της αρεσκείας σας, υπολογίστε:

   (α′) το μέσο ψηφίο

   (β′) το μητρώο συνδιασπορών

   (γ′) τις οκτώ (8) πρώτες κύριες συνιστώσες και

   (δ′) τις ανακατασκευές του ψηφίου για $L = 1,\ 8,\ 16,\ 64$ και 256 αντίστοιχα. Σχεδιάστε και σχολιάστε κατάλληλα τα αποτελέσματά σας.

   (ε′) Σχεδιάστε τα ιστογράμματα των σφαλμάτων καθώς και το σφάλμα ανακατασκευής για κάθε ένα ψηφίο από αυτά που επιλέξατε.

AUTOENCODERS

Ένας autoencoder είναι ένα νευρωνικό δίκτυο που εκπαιδεύεται για να 'αντιγράφει' την είσοδό του στην έξοδο. Εσωτερικά, αποτελείται από έναν αριθμό κρυφών επιπέδων που σκόπο έχουν την 'αποδοτική' περιγραφή των δεδομένων εισόδου. Το δίκτυο αποτελείται από δύο (2) τμήματα στα οποία υλοποιούνται:

- μια συνάρτηση κωδικοποίησης $\mathbf{z} = E(\mathbf{x}; \theta)$ όπου θ οι παράμετροι του κωδικοποιητή (encoder), με $\mathbf{x} \in \mathbb{R}^N$, $\mathbf{z} \in \mathbb{R}^L$ και
- μια συνάρτηση αποκωδικοποίησης $\hat{\mathbf{x}} = D(\mathbf{z}; \phi)$ όπου ϕ οι παράμετροι του αποκωδικοποιητή (decoder), και $\hat{\mathbf{x}} \in \mathbb{R}^N$ η ανακατασκευή των δεδομένων εισόδου $\mathbf{x}$.

Η 'αντιγραφή' της εισόδου στην έξοδο μπορεί να παρουσιάζεται ως μια διαδικασία χωρίς σημασία. Ωστόσο, δεν μας ενδιαφέρει η απλή μεταφορά των δεδομένων εισόδου στην έξοδο του δικτύου αλλά η αποδοτική περιγραφή τους. Ένας τρόπος για να πετύχουμε τον σκοπό μας, είναι να περιορίζουμε το μέγεθος της εξόδου του κωδικοποιητή, σε σχέση με αυτό της εισόδου του. Ένα τέτοιο δίκτυο ονομάζεται και undercomplete

autoencoder. Η εκμάθηση μιας τέτοιας 'συμπιεσμένης' αναπαράστασης των δεδομένων αναγκάζει το δίκτυο να μάθει τα πιο σημαντικά χαρακτηριστικά των δεδομένων εκπαίδευσης. Η διαδικασία μάθησης μπορεί να περιγραφεί με απλό τρόπο, μέσω της λύσης του ακόλουθου προβλήματος ελαχιστοποίησης:

$$\min_{\theta,\phi} \mathcal{L}(\theta, \phi)$$

με χρήση του κλασικού αλγορίθμου οπισθοδιάδοσης (backpropagation).

Τυπικές συναρτήσεις κόστους, τις οποίες θα χρησιμοποιήσουμε και εμείς, είναι:

- το Μέσο τετραγωνικό σφάλμα (MSE) οριζόμενο ως:

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{\sim f_{\mathcal{X}}}\left[\frac{1}{N}||\mathcal{X} - D(E(\mathcal{X};\theta);\phi)||_2^2\right] \text{ και}^6$$

- η δυαδική εντροπία (binary entropy) η οποία ορίζεται από την ακόλουθη σχέση:

$$\mathcal{L}(\theta, \phi) = -\mathbb{E}_{\sim f_{\mathcal{X}}}\Big[\mathcal{X}\log(D(E(\mathcal{X};\theta);\phi)) + (1-\mathcal{X})\log(1-D(E(\mathcal{X};\theta);\phi))\Big].$$

Όπως είναι εύκολο να δούμε, και θα το επαληθεύσουμε και πειραματικά, όταν το δίκτυο είναι γραμμικό και η συνάρτηση κόστους είναι το μέσο τετραγωνικό σφάλμα, τότε ο undercomplete autoencoder ουσιαστικά συγκλίνει στην ίδια λύση με την PCA. Ωστόσο, οι Autoencoders με μη γραμμικές συναρτήσεις ενεργοποίησης αποτελούν γενικεύσεις της PCA.

Διαδικασια

Σε όλες τις υλοποιήσεις θα πρέπει αρχικά να κεντράρετε τα δεδομένα προκειμένου να έχουν μηδενική μέση τιμή. Καλείστε να μην χρησιμοποιήσετε bias στα επίπεδα των δικτύων σας. Η επιλογή optimizer καθώς και των τιμών των υπερπαραμέτρων, αφήνεται στην κρίση σας.

1. Κατασκευάστε έναν πλήρως συνδεδεμένο (full-connected) AE, ενός επιπέδου στον κωδικοποιητή & στον αποκωδικοποιητή αντίστοιχα και εκπαιδεύστε τον πάνω στο σύνολο δεδομένων MNIST:

[6] Όπου απαιτείται ο υπολογισμός της αναμενόμενης τιμής, αυτός θα αντικαθίσταται από τον αριθμητικό μέσο όρο πάνω στα δεδομένα που ανήκουν σε ένα batch, βασιζόμενοι στο νόμο των μεγάλων αριθμών.

(α΄) για 40 εποχές, με

(β΄) latent size$= 128$ και

(γ΄) batch size = 250.

Ως συνάρτηση κόστους ανακατασκευής χρησιμοποιήστε τη δυαδική εντροπία.

Σε κάθε εποχή:

- συγκρίνετε τα βάρη του κωδικοποιητή με το μητρώο V_L του προηγούμενου ερωτήματος
- παρουσιάστε διάγραμμα ομοιότητας μεταξύ των δύο, χρησιμοποιώντας οποιαδήποτε μετρική επιθυμείτε.
- ποιά ιδιαιτερότητα πρέπει να έχει το συγκεκριμένο δίκτυο προκειμένου ο ΑΕ να προσεγγίζει την PCA; Καταγράψτε αναλυτικά την εξήγησή σας.

2. Σχεδιάστε ένα μη γραμμικό, πλήρως συνδεδεμένο ΑΕ τριών επιπέδων:

   - στον κωδικοποιητή και
   - στον αποκωδικοποιητή αντίστοιχα

   και εκπαιδεύστε τον στο σύνολο δεδομένων εκπαίδευσης του MNIST όπως και στο Ερώτημα 1.

   Υπολογίστε το πλήθος των παραμέτρων του δικτύου σας (με βάση το μέγεθος των ενδιάμεσων επιπέδων που επιλέξατε για την αρχιτεκτονική σας).

3. Εκπαιδεύστε δίκτυο παρόμοιας αρχιτεκτονικής με του προηγούμενου ερωτήματος, το οποίο όμως χρησιμοποιεί ακριβώς τις μισές παραμέτρους, ακολουθώντας τη λογική ανακατασκευής της PCA.

   (α΄) Τι σχέση πρέπει να έχει ο αποκωδικοποιητής με τον κωδικοποιητή σε αυτή την περίπτωση;

(β΄) Αν στον κωδικοποιητή χρησιμοποιήσουμε συναρτήσεις ενεργοποίησης leaky relu με κλίση 0.2, τι συναρτήσεις ενεργοποίησης πρέπει να χρησιμοποιήσετε στον αποκωδικοποιητή και με ποιά κλίση; Εξηγήστε την απάντησή σας.

4. Επαναλάβετε την παραπάνω διαδικασία, κάνοντας χρήση ψευδοαντιστρόφων των βαρών του κωδικοποιητή για τα βάρη του αποκωδικοποιητή.

5. Παρουσιάστε συγκεντρωτικά παραδείγματα ανακατασκευής καθώς και πίνακα με τις τιμές του μέσου τετραγωνικού σφάλματος ανακατασκευής των δεδομένων ελέγχου για τους τρεις διαφορετικούς ΑΕ που εκπαιδεύσατε.

Πιθανοτικη PCA

Μπορούμε να διατυπώσουμε την πιθανοτική PCA εισάγοντας μια κρυφή ή λανθάνουσα (latent) διανυσματική τ.μ. $\mathcal{Z}$ της οποίας η διάσταση ταυτίζεται με την διάσταση του **κύριου υπόχωρου**, δηλαδή L και της οποίας η συνάρτηση πυκνότητας πιθανότητας (σππ) θεωρούμε ότι είναι κανονική Gaussian (αυτή ουσιαστικά ταυτίζεται με την εκ των προτέρων συνάρτηση πυκνότητας πιθανότητας της τ.μ. $\mathcal{Z}$), δηλαδή:

$$f_{\mathcal{Z}}(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$$

και αντίστοιχα ή υπό συνθήκη σππ της τ.μ. $\mathcal{X}$ που κρύβεται πίσω από τις παρατηρήσεις, με συνθήκη την τιμή της λανθάνουσας μεταβλητής $\mathcal{Z}$, είναι επίσης μία **ισοτροπική** διασποράς σ^2 Gaussian της ακόλουθης μορφής:

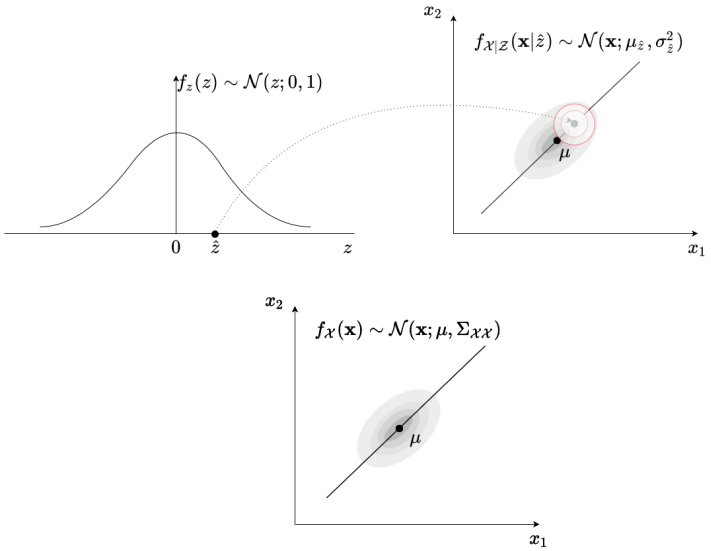
$$f_{\mathcal{X}|\mathcal{Z}}(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}; W\mathbf{z} + \mu, \sigma^2\mathbf{I})$$

όπου η αναμενόμενη τιμή της τ.μ. $\mathcal{X}$ είναι μία γραμμική συνάρτηση της τ.μ. $\mathcal{Z}$ η οποία πολλαπλασιάζεται με το $N \times M$ μητρώο W και αθροίζεται με το N-διάστατο διάνυσμα μ.

Μπορούμε να δούμε την πιθανοτική PCA και σαν ένα αναγεννητικό μηχανισμό των δεδομένων. Συγκεκριμένα, μία N-διάστατη παρατήρηση $\mathcal{X}$ μπορούμε να την ορίσουμε μέσω του ακόλουθου γραμμικού μετασχηματισμού της L-διάστατης λανθάνουσας μεταβλητής $\mathcal{Z}$:

$$\mathcal{X} = W\mathcal{Z} + \mu + \mathcal{W}$$

όπου $\mathcal{W}$ μία ισοτροπική διασποράς σ^2 Gaussian μηδενικής μέσης τιμής τ.μ.. Αυτός ο μηχανισμός γέννεσης φαίνεται στο ακόλουθο σχήμα.



Σχήμα 1: Απεικόνιση του λανθάνοντα χώρου στο χώρο των δεδομένων

Είναι προφανές (αλλά ωστόσο αποδείξτε το) ότι ισχύουν οι ακόλουθες σχέσεις:

$$\begin{aligned} \mathbb{E}\{\mathcal{X}\} &= \mathbb{E}\{W\mathcal{Z} + \mu + \mathcal{W}\} = \mu \\ \Sigma_{\mathcal{X}\mathcal{X}} &= \mathbb{E}\{(W\mathcal{Z} + \mathcal{W})(W\mathcal{Z} + \mathcal{W})^T\} \\ &= W\mathbb{E}\{\mathcal{Z}\mathcal{Z}^T\}W^T + \mathbb{E}\{\mathcal{W}\mathcal{W}^t\} \\ &= WW^T + \sigma^2\mathbf{I}. \end{aligned} \tag{3}$$

και επομένως η σππ της τ.μ. $\mathcal{X}$ θα είναι:

$$f_{\mathcal{X}}(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mu, \Sigma_{\mathcal{X}\mathcal{X}}).$$

Ας υποθέσουμε τώρα ότι θέλουμε να υπολογίσουμε την ακόλουθη υπό συνθήκη σππ:

$$f_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}) \tag{4}$$

η οποία είναι γνωστή ως **εκ των υστέρων** (posterior) σππ.

Αν γνωρίζουμε την σππ $f_{\mathbf{X}}(\mathbf{x})$, τότε από το θεώρημα του Bayes εύκολα βρίσκουμε ότι:

$$f_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}) = \frac{f_{\mathcal{X}|\mathcal{Z}}(\mathbf{x}|\mathbf{z})f_{\mathcal{Z}}(\mathbf{z})}{f_{\mathcal{X}}(\mathbf{x})},$$

όπου $f_{\mathcal{X}|\mathcal{Z}}(\mathbf{x}|\mathbf{z})$ η **συνάρτηση πιθανοφάνειας** (likelihood).

Αποδείξτε ότι η παραπάνω υπό συνθήκη σππ, με δεδομένο αυτό της Σχέσης (4), δίνεται από την ακόλουθη σχέση:

$$f_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \Sigma_1^{-1}W^T(\mathbf{x}-\mu), \sigma^2\Sigma_1^{-1}), \tag{5}$$

όπου $\Sigma_1 = W^TW + \sigma^2 I$. Αξιοσημείωτο είναι ότι η εκ των υστέρων αναμενόμενη τιμή εξαρτάται από το $\mathbf{x}$ ενώ το εκ των υστέρων μητρώο συνδιασπορών όχι.

Kullback-Leibler Divergence

Το Kullback-Leibler Divergence-KLD είναι ένα πολύ γνωστό μέτρο[7] ομοιότητας (similarity measure), το οποίο χρησιμοποιείται στην μέτρηση της ομοιότητας δύο κατανομών. Το KLD για τις σππ $p_{\mathcal{X}}(\mathbf{x})$ και $q_{\mathcal{X}}(\mathbf{x})$ ορίζεται ως ακολούθως:

$$D_{KL}(p_{\mathcal{X}}(\mathbf{x})||q_{\mathcal{X}}(\mathbf{x})) = \mathbb{E}_{\sim\, p_{\mathcal{X}}}\Big(log\Big[\frac{p_{\mathcal{X}}(\mathbf{x})}{q_{\mathcal{X}}(\mathbf{x})}\Big]\Big) = \int p_{\mathcal{X}}(\mathbf{x})log\Big[\frac{p_{\mathcal{X}}(\mathbf{x})}{q_{\mathcal{X}}(\mathbf{x})}\Big]dx$$

Παρατηρήστε ότι:

$$D_{KL}(p_{\mathcal{X}}(\mathbf{x})||q_{\mathcal{X}}(\mathbf{x})) \geq\ 0 \tag{6}$$

με την ισότητα να ισχύει όταν $p_{\mathcal{X}}(\mathbf{x}) = q_{\mathcal{X}}(\mathbf{x})$ (**αποδείξτε το**).

Υποθέστε οτι θέλετε να ελέγξετε αν οι τ.μ. $\mathcal{X}, \mathcal{Y}$ με από κοινού σππ $f_{\mathcal{X}\mathcal{Y}}(\mathbf{x},\ \mathbf{y})$ είναι στατιστικά ανεξάρτητες. Ορίστε κατάλληλα το KLD που θα δώσει λύση στο πρόβλημά σας. Συνδέστε την απάντησή σας με την έννοια της **αμοιβαίας πληροφορίας** (mutual information).

Ας θεωρήσουμε τώρα ότι η σππ $f_{\mathcal{X}}(\mathbf{x})$ είναι άγνωστη και επομένως δεν μπορούμε να υπολογίσουμε την υπό συνθήκη σππ $f_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x})$:

- είτε χρησιμοποιώντας το θεώρημα του Bayes,

[7] Το KLD δεν είναι συμμετρικό, δηλαδή, $D_{KL}(p_{\mathcal{X}}(\mathbf{x})||q_{\mathcal{X}}(\mathbf{x})) \neq\ D_{KL}(q_{\mathcal{X}}(\mathbf{x})||p_{\mathcal{X}}(\mathbf{x}))$. Αυτός είναι και ο λόγος που αναφέρεται ως **απόκλιση** (divergence) και όχι ως μετρική (metric).

- είτε, εναλλακτικά, μέσω της διαδικασίας της **περιθωριοποίησης** (marginalization):

$$f_{\mathcal{X}}(\mathbf{x}) = \int f_{\mathcal{X}|\mathcal{Z}}(\mathbf{x}|\mathbf{z}) f_{\mathcal{Z}}(\mathbf{z}) dz,$$

γιατί ο υπολογισμός δεν είναι εύκολος.

Σε αυτή την περίπτωση, καταφεύγουμε σε τεχνικές προσέγγισης της άγνωστης σππ από μια παραμετρική οικογένεια σππ (συνήθως Gaussian) την οποία ας συμβολίσουμε με $\hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x};\theta)$, όπου θ οι παράμετροι της σππ και για την ποιότητα της προσέγγισης μας θα χρησιμοποιούμε το KLD, δηλαδή:

$$D_{KL}(\hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x};\theta)||f_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x})). \tag{7}$$

Για μία συγκεκριμένη εικόνα $\mathbf{x}_m$ και εφαρμόζοντας το θεώρημα του Bayes, έχουμε:

$$\begin{aligned}
D_{KL}(\hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}_m;\theta) \quad || \quad & f_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}_m)) = \\
& \int \hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}_m;\theta) \log\Big[\frac{\hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}_m;\theta)}{f_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}_m)}\Big] dz \\
= \quad & \int \hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}_m;\theta) \log\Big[\frac{\hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}_m;\theta) f_{\mathcal{X}}(\mathbf{x}_m)}{f_{\mathcal{X}|\mathcal{Z}}(\mathbf{x}_m|\mathbf{z}) f_{\mathcal{Z}}(\mathbf{z})}\Big] dz \\
= \quad & \int \hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}_m;\theta) \log\Big[\frac{\hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}_m;\theta)}{f_{\mathcal{Z}}(\mathbf{z})}\Big] dz \\
+ \quad & \int \hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}_m;\theta) \log\Big[\frac{f_{\mathcal{X}}(\mathbf{x}_m)}{f_{\mathcal{X}|\mathcal{Z}}(\mathbf{x}_m|\mathbf{z})}\Big] dz \\
= \quad & D_{KL}(\hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}_m;\theta)||f_{\mathcal{Z}}(\mathbf{z})) + \log\big[f_{\mathcal{X}}(\mathbf{x}_m)\big] \\
- \quad & \mathbb{E}_{\sim\, \hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}_m;\theta)}(\log\big[f_{\mathcal{X}|\mathcal{Z}}(\mathbf{x}_m|\mathbf{z})\big])
\end{aligned} \tag{8}$$

Άρα, παίρνοντας υπόψην μας την Σχέση (6) εύκολα καταλήγουμε στην ακόλουθη ανισότητα:

$$\begin{aligned}
\log\big[f_{\mathcal{X}}(\mathbf{x}_m)\big] \quad \geq \quad & -D_{KL}(\hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}_m;\theta)||f_{\mathcal{Z}}(\mathbf{z})) \\
+ \quad & \mathbb{E}_{\sim\, \hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}_m;\theta)}(\log\big[f_{\mathcal{X}|\mathcal{Z}}(\mathbf{x}_m|\mathbf{z})\big]).
\end{aligned} \tag{9}$$

Το δεξιό μέλος της παραπάνω ανισότητας είναι γνωστό ως το Evidence Lower Bound-ELBO, δηλαδή:

$$\mathcal{G}_m(\theta) = -D_{KL}(\hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}_m;\theta)||f_{\mathcal{Z}}(\mathbf{z})) + \mathbb{E}_{\sim\, \hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}_m;\theta)}(\log\big[f_{\mathcal{X}|\mathcal{Z}}(\mathbf{x}_m|\mathbf{z})\big]) \tag{10}$$

και η μεγιστοποίησή του οδηγεί στην μεγιστοποίηση της πιθανότητας των δεδομένων μας. Επομένως, από τη λύση του ακόλουθου προβλήματος βελτιστοποίησης:

$$\max_{\theta} \mathcal{G}_m(\theta)$$

προκύπτει η βέλτιστη λύση του αρχικού μας προβλήματος.

Variational Autoencoders

Οι Variational Autoencoders αποτελούν μια ειδική μορφή των Autoencoders. Από πλευράς αρχιτεκτονικής, τα δύο είδη δικτύων έχουν πολλές ομοιότητες, με την έννοια πως και τα δύο δημιουργούνται με πλήρως συνδεδεμένα (full connected) ή συνελικτικά επίπεδα. Η διαφορά των δύο αρχιτεκτονικών είναι πως, σε αντίθεση με τους autoencoders, τα VAE έχουν μιας ειδικής μορφής αποκωδικοποιητές. Συγκεκριμένα, θέλουμε ο αποκωδικοποιητής να λειτουργεί ως ένα αναγεννητικό (generative) δίκτυο. Με αυτό το σκοπό, ο κωδικοποιητής του VAE, εκπαιδεύεται έτσι ώστε, ο χώρος των κρυφών λανθανουσών (latent) μεταβλητών, εκτός της μικρότερης διάστασης, να έχει επιπλέον και κάποιες ιδιότητες που εξασφαλίζουν τον αναγεννητικό χαρακτήρα του αποκωδικοποιητή.

Τις περισσότερες φορές όταν δειγματοληπτούμε από τον χώρο των λανθανουσών μεταβλητών ενός AE και οδηγούμε αυτά τα δείγματα στην είσοδο του αποκωδικοποιητή, η έξοδος που προκύπτει δεν είναι ποιοτική. Τα VAE είναι εκπαιδευμένα έτσι ώστε, ο χώρος των λανθανουσών μεταβλητών να είναι ομαλός, με την έννοια ότι κάθε σημείο σε αυτό το χώρο παράγει μέσω του κωδικοποιητή μια έγκυρη έξοδο. Μια ενδιαφέρουσα ιδιότητα των VAE είναι πως κάποιος μπορεί να παράξει την λανθάνουσα αναπαράσταση δύο εικόνων και επιλέγοντας τα σημεία στην ευθεία που ενώνει αυτές τις δύο αναπαραστάσεις, να κάνει interpolation μεταξύ των δύο αυτών εικόνων εισάγοντας κάθε ένα από τα σημεία της ευθείας με τη σειρά στον αποκωδικοποιητή.

Όπως και στην περίπτωση της πιθανοτικής PCA, θα καταφύγουμε σε τεχνικές προσέγγισης της άγνωστης σππ από μια παραμετρική οικογένεια σππ (συνήθως Gaussian) την οποία ας συμβολίσουμε με $\hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x};\theta)$, όπου θ οι παράμετροι της σππ και για την ποιότητα της προσέγγισής μας

Θα χρησιμοποιήσουμε το KLD, δηλαδή:

$$D_{KL}(\hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x};\theta)||f_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x})). \tag{11}$$

Θα πρέπει επίσης να τονίσουμε στο σημείο αυτό ότι στην περίπτωση των VAE και η συνάρτηση πιθανοφάνειας $f_{\mathcal{X}|\mathcal{Z}}(\mathbf{x}|\mathbf{z})$ θα προσεγγιστεί, από των αποκωδικοποιητή, από μία παραμετρική οικογένεια σππ την οποία ας συμβολίσουμε με $\hat{f}_{\mathcal{X}|\mathcal{Z}}(\mathbf{x}|\mathbf{z};\phi)$, όπου ϕ οι παράμετροι της σππ. Για μία συγκεκριμένη εικόνα $\mathbf{x}_m$ και εφαρμόζοντας το θεώρημα του Bayes, έχουμε:

$$\begin{aligned}
D_{KL}(\hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}_m;\theta) \quad || \quad & f_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}_m)) = \\
& \int \hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}_m;\theta)\log\Big[\frac{\hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}_m;\theta)}{f_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}_m)}\Big]dz \\
= & \int \hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}_m;\theta)\log\Big[\frac{\hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}_m;\theta)f_{\mathcal{X}}(\mathbf{x}_m)}{f_{\mathcal{X}|\mathcal{Z}}(\mathbf{x}_m|\mathbf{z};\phi)f_{\mathcal{Z}}(\mathbf{z})}\Big]dz \\
= & \int \hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}_m;\theta)\log\Big[\frac{\hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}_m;\theta)}{f_{\mathcal{Z}}(\mathbf{z})}\Big]dz \\
+ & \int \hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}_m;\theta)\log\Big[\frac{f_{\mathcal{X}}(\mathbf{x}_m)}{f_{\mathcal{X}|\mathcal{Z}}(\mathbf{x}_m|\mathbf{z};\phi)}\Big]dz \\
= & D_{KL}(\hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}_m;\theta)||f_{\mathcal{Z}}(\mathbf{z})) + \log\big[f_{\mathcal{X}}(\mathbf{x}_m)\big] \\
- & \mathbb{E}_{\sim\, \hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}_m;\theta)}(\log\big[f_{\mathcal{X}|\mathcal{Z}}(\mathbf{x}_m|\mathbf{z};\phi)\big])
\end{aligned} \tag{12}$$

Άρα, λαμβάνοντας υπόψην την Σχέση (6) εύκολα καταλήγουμε στην ακόλουθη ανισότητα:

$$\begin{aligned}
\log\big[f_{\mathcal{X}}(\mathbf{x}_m)\big] \quad \geq \quad & -D_{KL}(\hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}_m;\theta)||f_{\mathcal{Z}}(\mathbf{z})) \\
+ \quad & \mathbb{E}_{\sim\, \hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}_m;\theta)}(\log\big[f_{\mathcal{X}|\mathcal{Z}}(\mathbf{x}_m|\mathbf{z};\phi)\big])
\end{aligned} \tag{13}$$

Το δεξιό μέλος της παραπάνω ανισότητας αποτελεί και σε αυτή την περίπτωση το (Evidence Lower Bound-ELBO), δηλαδή:

$$\mathcal{G}_m(\theta,\phi) = -D_{KL}(\hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}_m;\theta)||f_{\mathcal{Z}}(\mathbf{z}))+\mathbb{E}_{\sim\, \hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}_m;\theta)}(\log\big[\hat{f}_{\mathcal{X}|\mathcal{Z}}(\mathbf{x}_m|\mathbf{z};\phi)\big]) \tag{14}$$

και η μεγιστοποίησή του οδηγεί στην μεγιστοποίηση της πιθανότητας των δεδομένων μας. Επομένως, από τη λύση του ακόλουθου προβλήματος

βελτιστοποίησης:

$$\max_{\theta,\phi} \mathcal{G}_m(\theta, \phi)$$

προκύπτει η βέλτιστη λύση του αρχικού μας προβλήματος.

Είναι προφανές ότι αν ορίσουμε την ακόλουθη συνάρτηση:

$$\mathcal{L}_m(\theta, \phi) = -\mathcal{G}_m(\theta, \phi) \tag{15}$$

αυτή αποτελεί συνάρτηση κόστους και είναι η συνάρτηση που χρησιμοποιούμε για την εκπαίδευση του δικτύου με χρήση του αλγορίθμου οπισθοδρόμησης. Είναι φανερό ότι:

- ο πρώτος όρος εκφράζει την απόσταση της εκ των υστέρων κατανομής των λανθανουσών μεταβλητών από την κανονική $N(\mathbf{z}; \mathbf{0}, I)$, ενώ
- ο δεύτερος όρος αφορά την ποιότητα ανακατασκευής.

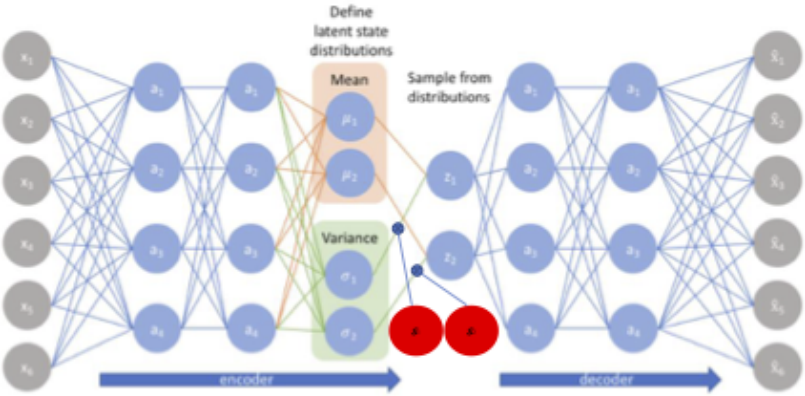
Ο αποκωδικοποιητής δειγματοληπτεί το $\mathcal{Z}$ από την $\hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}; \theta)$. Γνωρίζοντας πως μια μεταβλητή που ακολουθεί κανονική κατανομή με μέση τιμή $\mu_{\mathbf{x}_m}(\theta)$ και τυπική απόκλιση $\sigma_{\mathbf{x}_m}(\theta)$ μπορεί να δειγματοληπτηθεί ως:

$$\mathcal{Z} = \mu_{\mathbf{x}_m}(\theta) + \sigma_{\mathbf{x}_m}(\theta) \circledcirc \mathcal{E}$$

όπου το $\mathcal{E}$ δειγματοληπτείται από μια τυπική πολυδιάστατη κανονική κατανομή κατάλληλων διαστάσεων. Αυτό είναι γνωστό και ως 'reparametrization trick', το οποίο χρησιμοποιείται προκειμένου να μεταφέρουμε τη στοχαστικότητα στον όρο $\mathcal{E}$, ώστε να μπορεί ο αλγόριθμος οπισθοδιάδοσης να υπολογίζει τις παραγώγους σε αυτό το σημείο.

Ο KLD όρος της συνάρτησης κόστους υπολογίζεται μέσω κλειστού τύπου. Η πρώτη κατανομή υποθέτουμε πως είναι πολυδιάστατη κανονική, με μέση τιμή $\mu_{\mathbf{x}_m}(\theta)$ και μητρώο συνδιακύμανσης $\sigma_{\mathbf{x}_m}(\theta)I$ ή αλλιώς $N(\mu_{\mathbf{x}_m}(\theta), \sigma_{\mathbf{x}_m}(\theta)I)$, ενώ η δεύτερη είναι η τυπική πολυδιάστατη κανονική κατανομή $N(0, I)$. Σε αυτή την περίπτωση ο KLD όρος γίνεται:

$$\begin{aligned} -D_{KL}(\hat{f}_{\mathcal{Z}|\mathcal{X}}(\mathbf{z}|\mathbf{x}_m; \theta)||f_{\mathcal{Z}}(\mathbf{z})) &= -\frac{L}{2} + \frac{1}{2}\sum_{l=1}^{L} \sigma^2_{\mathbf{x}_m}(l; \theta) \\ &+ \frac{1}{2}\sum_{l=1}^{L} \left(\mu^2_{\mathbf{x}_m}(l; \theta) - \ln(\sigma^2_{\mathbf{x}_m}(l; \theta))\right) \end{aligned} \tag{16}$$



Σχήμα 2: Παράδειγμα αρχιτεκτονικής ενός VAE

ΔΙΑΔΙΚΑΣΙΑ

1. Σχεδιάστε έναν VAE τριών επιπέδων[8]:

   - στον κωδικοποιητή και
   - στον αποκωδικοποιητή αντίστοιχα

   και εκπαιδεύστε τον:

   - για 100 εποχές, με
   - latent size = 2 και
   - batch size = 250.

   Για τον όρο ανακατασκευής στη συνάρτηση κόστους, χρησιμοποιήστε τη δυαδική εντροπία.

   (α΄) Δημιουργήστε ένα batch θορύβου από την $\mathcal{N}(\mathbf{z}; \mathbf{0}, I)$ το οποίο θα δίνετε σε κάθε εποχή, μετά τη φάση εκπαίδευσης και ελέγχου, στον αποκωδικοποιητή. Παρουσιάστε πως εξελίσεται η ανακατασκευή που προκύπτει για αυτό το batch (figures ανακατασκευών για τις εποχές 1, 50 και 100).

[8] Η επιλογή optimizer καθώς και των τιμών των υπερπαραμέτρων, αφήνεται στην κρίση σας.

(β΄) Μετά το πέρας της εκπαίδευσης, δημιουργήστε την λανθάνουσα αναπαράσταση όλων των δεδομένων ελέγχου. Εκτυπώστε σε κοινό scatter plot τις αναπαραστάσεις αυτές, χρωματίζοντας με κοινό χρώμα όσες αντιστοιχούν σε ίδια ψηφία. Παρουσιάστε το plot και καταγράψτε τις παρατηρήσεις σας.

Βιβλιογραφία

[1] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[2] Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *CoRR*, abs/1906.02691, 2019.

[3] Yunfei Teng and Anna Choromanska. Invertible autoencoder for domain adaptation. *Computation*, 7(2), 2019.

[4] Philip R. Merrifield. Book reviews : Modern factor analysis by harry h. harman. chicago: University of chicago press, 1960. *Educational and Psychological Measurement*, 21(4):1043–1047, 1961.