Αυτοκωδικοποιητές (Autoencoders)

Βαθιά Μηχανική Μάθηση

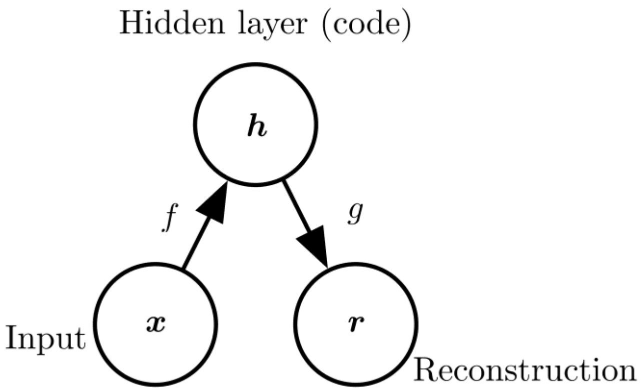
ΔΠΜΣ Επιστήμης Δεδομένων & Μηχανικής Μάθησης

Εθνικό Μετσόβιο Πολυτεχνείο

Γιώργος Αλεξανδρίδης (gealexan@mail.ntua.gr)

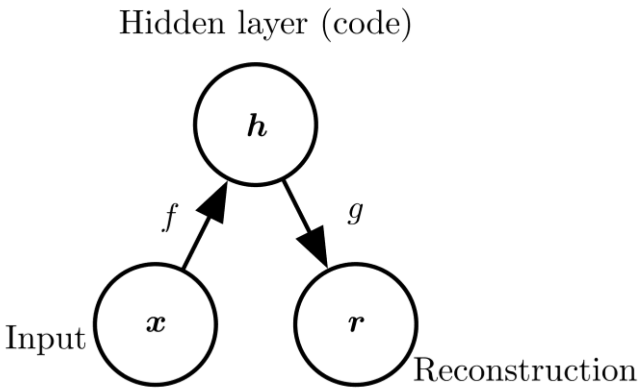
Αυτοκωδικοποιητές

- **Αυτοκωδικοποιητής** (*Autoencoder* ή ΑΚ)
  - Νευρωνικό δίκτυο που εκπαιδεύεται να αντιγράφει την είσοδό του στην έξοδό του
  - Εσωτερικά αποτελείται από κρυφό επίπεδο $\boldsymbol{h}$ στο οποίο αναπαρίσταται **κωδικοποιημένη** (*coded*) η είσοδος
  - Δύο μέρη
    1. Συνάρτηση **κωδικοποίησης** (*encoder function*) $\boldsymbol{h} = f(\boldsymbol{x})$
    2. Συνάρτηση **αποκωδικοποίησης** (*decoder function*) $\boldsymbol{r} = g(\boldsymbol{h})$



Αυτοκωδικοποιητές

- **Εκπαίδευση**
  1. Μέσω προς τα πίσω διάδοσης του σφάλματος
     - Όπως στα ΤΝΔ πρόσθιας τροφοδότησης
  2. Μέσω επανακυκλοφορίας (recirculation)
     - Συγκρίνεται η ενεργοποίηση των νευρώνων στην αρχική είσοδο και στην αναπαράσταση
- Τετριμμένη λύση: Μάθε την $g(f(\boldsymbol{x})) = \boldsymbol{x}, \forall \boldsymbol{x}$
  - Δεν έχει νόημα γι' αυτό οι ΑΚ σχεδιάζονται έτσι ώστε **να μην μπορούν** να αντιγράψουν τέλεια
- Παρουσιάστηκαν στα μέσα του 1980
- Χρήση σε προβλήματα **μείωσης διαστατικότητας** (*dimensionality reduction*) και **εξαγωγής χαρακτηριστικών** (*feature extraction*)



Διαδικασία Μάθησης

- Δεν μας ενδιαφέρει τόσο η διαδικασία αντιγραφής, όσο να αποτυπωθούν στο $\boldsymbol{h}$ χρήσιμες ιδιότητες του $\boldsymbol{x}$
  - **Υποπλήρης** (*undercomplete*) ΑΚ: Περιορίζουμε το $\boldsymbol{h}$ ώστε να έχει μικρότερες διαστάσεις από το $\boldsymbol{x}$
    - Μαθαίνει τα **προεξέχοντα** (*salient*) χαρακτηριστικά της εισόδου
- **Διαδικασία μάθησης**: Ελαχιστοποίηση συνάρτησης απώλειας $L(x, g(f(x)))$
  - Αν g γραμμική και L Μέσο Τετραγωνικό Σφάλμα (ΜΤΣ), τότε ο υποπλήρης ΑΚ καλύπτει τον ίδιο **υποχώρο** με την PCA (Γιατί; Άσκηση για το σπίτι!)
  - Αν f, g μη γραμμικές, ο υποπλήρης ΑΚ μαθαίνει μια πιο **ισχυρή μη-γραμμική γενίκευση** της PCA
  - Αν f, g έχουν **πολύ μεγάλη χωρητικότητα**, τότε απλά θα μάθουν να αντιγράφουν την είσοδο στην έξοδο χωρίς εξαγωγή χαρακτηριστικών!
- **Υπερπλήρης** (*overcomplete*) ΑΚ: Θέτουμε $\dim(\boldsymbol{h}) \geq \dim(\boldsymbol{x})$
- Η επιλογή της διάστασης του ΑΚ καθώς και της χωρητικότητας των f,g εξαρτάται από την πολυπλοκότητα της **υποκείμενης** (*underlying*) κατανομής των δεδομένων που μοντελοποιούνται

Ομαλοποιημένοι Αυτοκωδικοποιητές

- **Ομαλοποιημένοι** (*regularized*) ΑΚ
  - Περιορισμός της **χωρητικότητάς** τους μέσω συνάρτησης απώλειας
  - Θυμηθείτε το ρόλο της ομαλοποίησης από προηγούμενα μαθήματα και διαλέξεις!
- Συνάρτηση απώλειας επιβάλλει πρόσθετες ιδιότητες πέραν της αντιγραφής
  - **Αραιότητα** (*sparsity*) της αναπαράστασης
  - Μικρό **μέγεθος** της **παραγώγου**
  - **Ανοχή** σε **θόρυβο** ή σε απουσιάζουσες τιμές
- Ένας *ομαλοποιημένος* ΑΚ μπορεί να είναι *μη-γραμμικός* και *υπερπλήρης* αλλά παρόλα αυτά να **μαθαίνει χαρακτηριστικά** της κατανομής των δεδομένων

Αραιοί Αυτοκωδικοποιητές

- **Αραιός** (*sparse*) ΑΚ
  - Προσθήκη όρου ποινής αραιότητας $\Omega(\boldsymbol{h})$ του επιπέδου κωδικοποίησης $\boldsymbol{h}$ στη διαδικασία μάθησης
  - $L(\boldsymbol{x},\ g(f(\boldsymbol{x})) + \Omega(\boldsymbol{h})$
- Προσθήκη όρων ομαλοποίησης στη διαδικασία μάθησης
  - Προσέγγιση της *Μεϋζιανής Συμπερασματολογίας* υπό τη μορφή της *Μέγιστης εκ των Υστέρων Πιθανότητας* (MAP)
    - Όπως και στην περίπτωση της ομαλοποίησης
  - **Ποινή ομαλοποίησης**: η εκ των προτέρων κατανομή $p(\theta)$ των παραμέτρων θ του μοντέλου
    - $p(\theta|x) \approx p(\boldsymbol{x}|\theta)p(\theta)$

Αραιοί Αυτοκωδικοποιητές

- Στους ΑΚ η ποινή ομαλοποίησης **εφαρμόζεται** στα δεδομένα ($\boldsymbol{x}$) και όχι στις παραμέτρους (θ)!
  - Ο αραιός ΑΚ δρα ως εκτιμητής μέγιστης πιθανοφάνειας ενός **παραγωγικού** (*generative*) μοντέλου με **λανθάνουσες** (*latent*) μεταβλητές $\boldsymbol{h}$
  - $p_{model}(\boldsymbol{x}, \boldsymbol{h}) = p_{model}(\boldsymbol{h}) p_{model}(\boldsymbol{x}|\boldsymbol{h})$
  - $p_{model}(\boldsymbol{h})$: εκτίμηση μοντέλου για την εκ των προτέρων κατανομή των λανθανουσών μεταβλητών

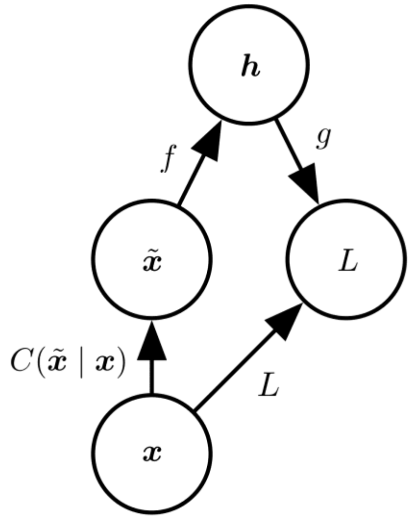
- $\log p_{model}(\boldsymbol{x}) = \log \sum_h p_{model}(\boldsymbol{h}, \boldsymbol{x})$
  - Στους ΑΚ το άθροισμα προσεγγίζεται από σημειακή εκτίμηση μιας πολύ πιθανής τιμής για το $\boldsymbol{h}$
  - $\max_{\boldsymbol{h}} \log p_{model}(\boldsymbol{h}, \boldsymbol{x}) = \max_{\boldsymbol{h}} \log p_{model}(\boldsymbol{h}) + \max_{\boldsymbol{h}} \log p_{model}(\boldsymbol{x}|\boldsymbol{h})$

Εκ των προτέρων κατανομή λανθανουσών μεταβλητών

- **Εισαγωγή αραιότητας**: κατάλληλη επιλογή $p_{model}(\boldsymbol{h})$
  - πχ Laplace, Student-t
- Κατανομή Laplace: Ισοδυναμεί με κανονικοποίηση L_1
  - $p_{model}(h_i) = \frac{\lambda}{2} e^{-\lambda|h_i|} \Rightarrow -\log p_{model}(\boldsymbol{h}) = \sum_i \left(\lambda|h_i| - \log\frac{\lambda}{2}\right) = \Omega(\boldsymbol{h}) + c$
  - Όρος c **σταθερά** (*υπερπαράμετρος* του μοντέλου)
    - Δεν εξαρτάται από το $\boldsymbol{h}$ αλλά μόνο από το λ
- Η ποινή αραιότητας δεν είναι όρος *κανονικοποίησης*
  - Γιατί δεν εφαρμόζεται στις παραμέτρους του μοντέλου αλλά στις λανθάνουσες μεταβλητές
  - Αποτελεί **συνέπεια** της θεώρησης για την **κατανομή** του μοντέλου όσον αφορά τις λανθάνουσες μεταβλητές
- Συνεπώς, η **εκπαίδευση** ενός **ΑΚ** είναι αντίστοιχη της εκπαίδευσης ενός **παραγωγικού μοντέλου**
  - Τα <u>χαρακτηριστικά</u> που μαθαίνει ο **ΑΚ** είναι επί της ουσίας οι λανθάνουσες μεταβλητές που χαρακτηρίζουν την είσοδο

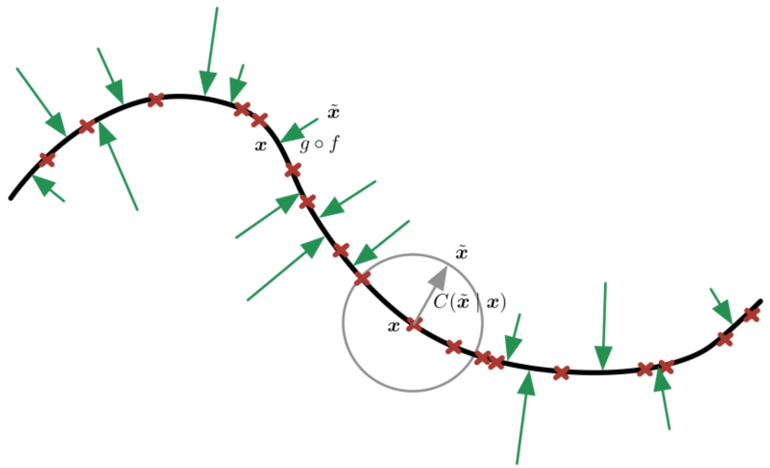
Αυτοκωδικοποιητές απαλοιφής θορύβου

- **ΑΚ απαλοιφής θορύβου** (*Denoising Autoencoders* ή DAE)
  - Ελαχιστοποίηση της $L(\boldsymbol{x}, g(f(\tilde{x}))$
  - $\tilde{x}$ αλλοίωση του $\boldsymbol{x}$ μέσω της προσθήκης θορύβου
  - $C(\boldsymbol{x}|\tilde{x})$: Υπό **συνθήκη κατανομή** παραγωγής αλλοιωμένων $\tilde{x}$ δεδομένου $\boldsymbol{x}$
- Ο ΑΚ μαθαίνει την **κατανομή αποκατάστασης** (*reconstruction distribution*)
  - Λήψη δείγματος $\boldsymbol{x}$ από τα δεδομένα εκπαίδευσης και αλλοιωμένης του μορφής $\tilde{x}$ από την $C(\tilde{x}|\boldsymbol{x} = x)$
  - Χρήση $(\boldsymbol{x}, \tilde{x})$ ως **δείγμα εκπαίδευσης** για την εκτίμηση $p_r(\widetilde{x}|\boldsymbol{x}) = p_d(\boldsymbol{x}|\boldsymbol{h})$
    - p_d ορίζεται από $g(\boldsymbol{h})$
- Αν θέσουμε f ντετερμινιστική, ο DAE συμπεριφέρεται ως ένα ΤΝΔ πρόσθιας τροφοδότησης
  - Συνεπώς μπορούμε να χρησιμοποιήσουμε αντίστοιχές τεχνικές μάθησης (λχ κατάβαση κλίσης)



Συνταίριασμα τιμών (score matching)

- Εναλλακτική μέθοδος στην εκτίμηση μέγιστης πιθανοφάνειας
  - **Παραγωγή εκτιμήσεων** για το μοντέλο, οι οποίες «ενθαρρύνονται» να εμφανίζουν παρόμοιες **τιμές** (*scores*) με την **κατανομή των δεδομένων** σε κάθε δείγμα $\boldsymbol{x}$
  - Οι τιμές προκύπτουν ως η αποτίμηση ενός **πεδίου κλίσεων** (*gradient field*):$\nabla_x \log p(x)$
- Σημαντική ιδιότητα των DAE
  - Το κριτήριο εκπαίδευσής τους κάνει τον ΑΚ να μάθει ένα **διανυσματικό πεδίο** $g(f(\boldsymbol{x})) - \boldsymbol{x}$ που αποτελεί εκτίμηση της τιμής της κατανομής των δεδομένων



Εκμάθηση Πολλαπλότητας

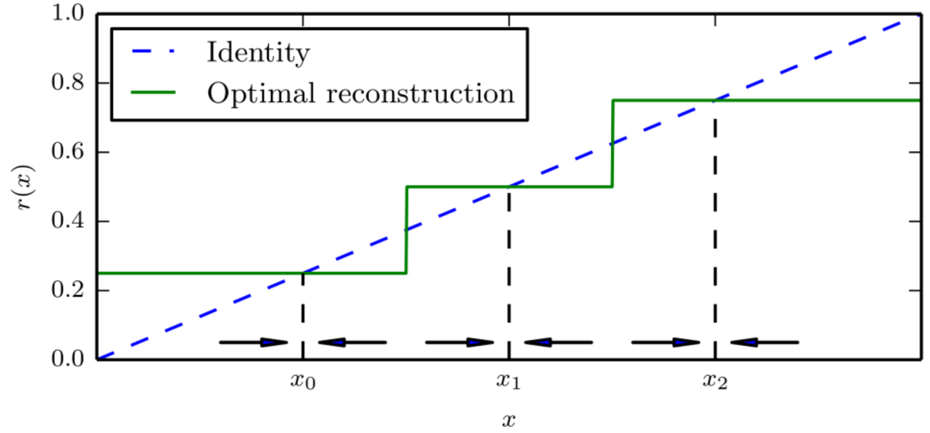
- Γενική υπόθεση ΑΚ
  - Τα δεδομένα είναι συγκεντρωμένα γύρω από **πολλαπλότητες** (*manifolds*) χαμηλότερων διαστάσεων
- Πολλαπλότητες χαρακτηρίζονται από τις **εφαπτόμενες πλευρές** τους (*tangent planes*)
  - Πολλαπλότητα **διάστασης** d ορίζεται από d **διανύσματα βάσης**
  - Διανύσματα βάσης ορίζουν τις επιτρεπόμενες αποκλίσεις εντός της πολλαπλότητας
    - Πόσο μπορεί να «αλλάξει» το x παραμένοντας ενός της πολλαπλότητας

Εκμάθηση Πολλαπλότητας

- Συμβιβασμός μεταξύ **δύο** αντίρροπων δυνάμεων
  1. Εκμάθησης κωδικοποίησης ***h*** από τα δεδομένα ***x***
     - έτσι ώστε το ***x*** να μπορεί να ανακτηθεί προσεγγιστικά από το ***h*** μέσω του ΑΚ
  2. Ικανοποίησης περιορισμών που περιορίζουν την χωρητικότητα του ΑΚ
     - λχ ποινές ομαλοποίησης
     - ΑΚ μαθαίνει αναπαραστάσεις που είναι **λιγότερο «ευαίσθητες»** στη μεταβολή της εισόδου

- Τελικά, ο ΑΚ μαθαίνει μόνο τις **διακυμάνσεις** που είναι απαραίτητες για την ανακατασκευή των δειγμάτων εκπαίδευσης

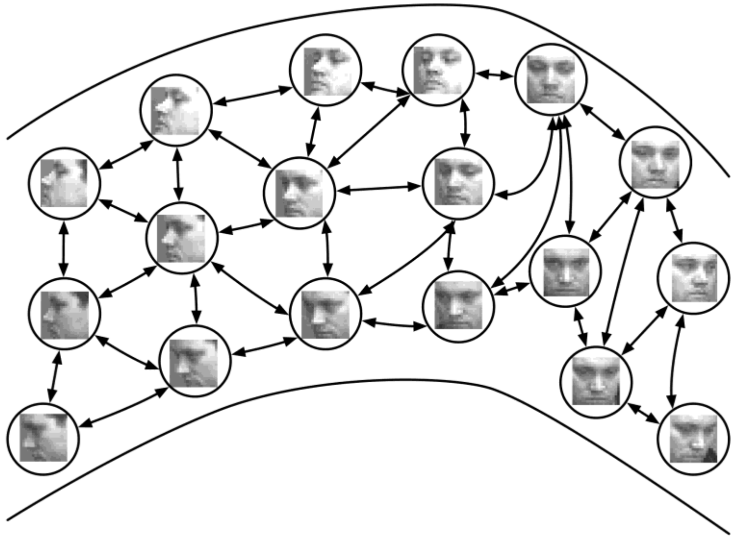
Παράδειγμα Εκμάθησης Πολλαπλότητας



- Χώρος δεδομένων μίας διάστασης
  - Δε μπορεί να μειωθεί άλλο η διάσταση
  - «Σημειακή» πολλαπλότητα
- Διακεκομμένη γραμμή
  - Ιδεατή συνάρτηση ταυτότητας
  - Αυτή επιθυμεί να κατασκευάσει ο ΑΚ
- Βέλη
  - Κατεύθυνση διανύσματος αναπαράστασης $r(x) - x$
  - Δείχνουν προς την πλησιέστερη πολλαπλότητα (εδώ, σημείο στο χώρο)
- Συνεχόμενη γραμμή
  - Ιδεατή συνάρτηση αναπαράστασης
    - Τέμνει την ιδεατή συνάρτηση ταυτότητας στα δείγματα των δεδομένων
    - Μεγάλη παράγωγος στον χώρο μεταξύ των πολλαπλοτήτων
      - Έτσι τα «αλλοιωμένα» σημεία απεικονίζονται πάνω στη «σημειακή» πολλαπλότητα

Μη-παραμετρικές μέθοδοι εκμάθησης πολλαπλότητας

- Τεχνικές μη-επιβλεπόμενης μάθησης
  - Γράφος πλησιέστερων γειτόνων
- Ένας κόμβος για κάθε δείγμα εκπαίδευσης
  - Ένωση με ακμές με τους γειτονικούς κόμβους
- Ανάκτηση εφαπτόμενων πλευρών κάθε κόμβου και της διανυσματικής του θέσης
  - **Embedding**
- Γενίκευση σε νέα δείγματα μέσω παρεμβολής
  - Αρκεί το αρχικό πλήθος των δειγμάτων να μπορεί να καλύψει πλήρως τα χαρακτηριστικά της πολλαπλότητας
    - Καμπυλότητες, πτυχώσεις κλπ



Συσταλτικοί Αυτοκωδικοποιητές

- *Contractive Autoencoders* ή CAE
  - Δεν πρέπει να μεταβάλλεται **πολύ** η αναπαράσταση όταν η είσοδος μεταβάλλεται ελαφρά
  - $L(\boldsymbol{x},\ g(f(\boldsymbol{x})) + \Omega(\boldsymbol{h}, x),$ $\qquad \Omega(\boldsymbol{h}, x) = \lambda \|\nabla_x \boldsymbol{h}\|^2$
  - **Όρος ποινής** $\Omega(\boldsymbol{h})$: Το τετράγωνο της νόρμας Frobenius του Ιακωβιανού πίνακα της συνάρτησης κωδικοποίησης

- Σχέση DAE και CAE
  - Το σφάλμα αναπαράστασης όταν έχει προστεθεί στην είσοδο μικρή ποσότητα γκαουσιανού θορύβου είναι ισοδύναμο με μια συσταλτική ποινή στη συνάρτηση αναπαράστασης $g \circ f$

Συσταλτικοί Αυτοκωδικοποιητές

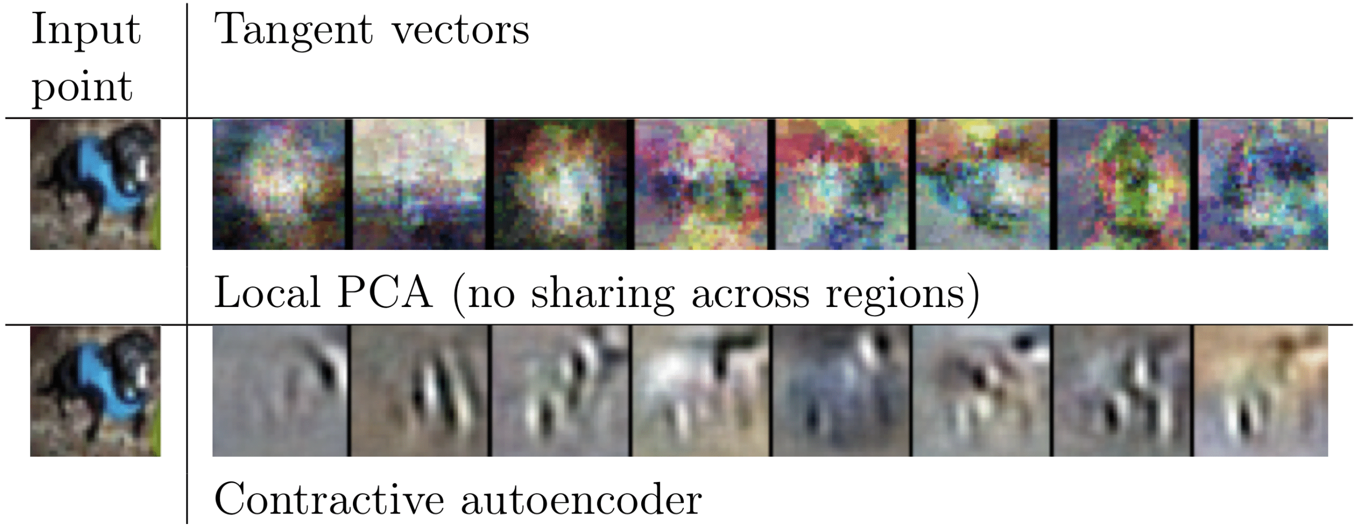
- Ιδιότητα συστολής είναι τοπική
  - Ο Ιακωβιανός πίνακας *προσεγγίζει γραμμικά* γύρω από σημείο $\boldsymbol{x}$ μια μη-γραμμική συνάρτηση

- Αντίρροπες δυνάμεις
  1. Σφάλμα αναπαράστασης
  2. Ποινή συστολής $\Omega(\boldsymbol{h})$

- Ισορροπία στα σημεία που οι περισσότερες *μερικές* παράγωγοι είναι **μηδέν**
  - Μόνο ένα μικρό πλήθος χαρακτηριστικών (διαστάσεων) θα έχει μεγάλες τιμές
    - CAE απεικονίζει χαρακτηριστικά σε μια **πολλαπλότητα**

- Γενικά, CAE μαθαίνει χαρακτηριστικά σταθερά ως προς $\boldsymbol{x}$

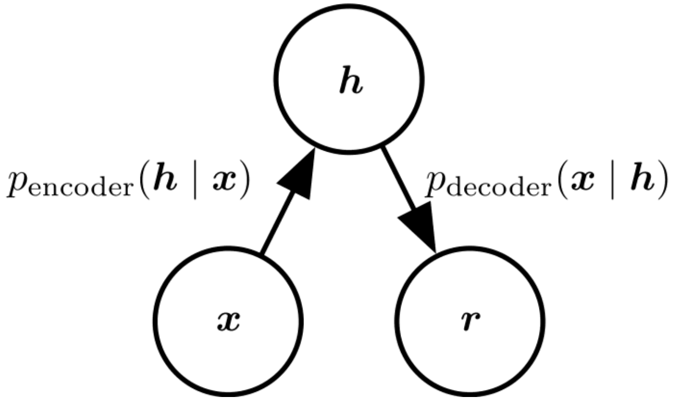
Συσταλτικοί Αυτοκωδικοποιητές: Παράδειγμα



- Εικόνα από το CIFAR-10 dataset
- Εφαπτόμενα διανύσματα υπολογίζονται από τα **κυρίαρχα ιδιάζοντα διανύσματα** Ιακωβιανής μήτρας $\frac{\boldsymbol{\vartheta h}}{\boldsymbol{\vartheta x}}$
- CAE καλύτερη αναπαράσταση από PCA
  - Εκμεταλλεύεται ιδιότητα **διαμοιρασμού παραμέτρων** μεταξύ **διαφορετικών περιοχών**

Στοχαστικοί αυτοκωδικοποιητές

- Διαδικασία κωδικοποίησης και αποκωδικοποίησης στοχαστική
  - Συνάρτηση Κωδικοποίησης $f \Longrightarrow$ κατανομή κωδικοποίησης $p_{encoder}(\boldsymbol{h}|\boldsymbol{x})$
  - Αντίστοιχα $g \Longrightarrow p_{decoder}(\boldsymbol{x}|\boldsymbol{h})$
- Κάθε μοντέλο λανθανουσών μεταβλητών $p_m(\boldsymbol{x}, \boldsymbol{h})$ μπορεί να ορίσει στοχαστικό κωδικοποιητή και αποκωδικοποιητή
  - $p_{encoder}(\boldsymbol{h}|\boldsymbol{x}) = p_m(\boldsymbol{h}|\boldsymbol{x})$
  - $p_{decoder}(\boldsymbol{x}|\boldsymbol{h}) = p_m(\boldsymbol{x}|\boldsymbol{h})$
- Στη γενική περίπτωση, ωστόσο, οι $p_{encoder}$ και $p_{decoder}$ δεν αποτελούν υπό συνθήκη κατανομές μιας ενιαίας κοινής κατανομής



Βιβλιογραφία

- Ian Goodfellow, Yoshua Bengio, Aaron Courville “Deep Learning” – MIT Press (https://www.deeplearningbook.org/)
  - Εισαγωγή (§14.1)
  - Ομαλοποιημένοι Αυτοκωδικοποιητές (§14.2, §14.5, §14.7)
  - Στοχαστικοί Αυτοκωδικοποιητές (§14.4)