MultiLayer Perceptron (συμπληρωματικά)

Ιωάννης Μαδεμλής

Οπισθοδιάδοση σφάλματος

MLP

- Τα πολυεπίπεδα νευρωνικά δίκτυα εμπρόσθιας τροφοδότησης (π.χ., MLP, CNN, κλπ.) κατά κανόνα εκπαιδεύονται επαναληπτικά με αλγορίθμους βελτιστοποίησης καθόδου κλίσης (gradient descent).
  - Ελαχιστοποίηση της προκαθορισμένης συνάρτηση κόστους ως προς τις παραμέτρους συναπτικών βαρών του δικτύου.
- Όμως οι εν λόγω αλγόριθμοι απαιτούν τον υπολογισμό των μερικών παραγώγων της συνάρτησης κόστους ως προς τις συναπτικές παραμέτρους.
- Η **οπισθοδιάδοση σφάλματος** (error back-propagation) είναι μία γρήγορη μέθοδος εύκολου υπολογισμού αυτών των παραγώγων, στηριγμένη στον κανόνα της αλυσίδας για την παραγώγιση σύνθετων συναρτήσεων.

Οπισθοδιάδοση σφάλματος

- Η οπισθοδιάδοση εκτελείται ξεχωριστά για κάθε πρότυπο εκπαίδευσης, με βάση την *τρέχουσα κάθε φορά εκδοχή* του δικτύου (τρέχον διάνυσμα παραμέτρων).
- Έτσι συγκροτείται το ζητούμενο **τρέχον διάνυσμα κλίσης** της συνάρτησης κόστους.
  - Μετά την εκτίμηση της κλίσης, επιτελείται μία ενημέρωση των συναπτικών βαρών/παραμέτρων μέσω της εξίσωσης της καθόδου κλίσης.
  - Μετά την ενημέρωση του μοντέλου, δειγματοληπτείται ένα διαφορετικό πρότυπο εκπαίδευσης και η όλη διαδικασία επαναλαμβάνεται.
- Ιδανικά, μετά από αρκετές διαδοχικές επαναλήψεις, η κάθοδος κλίσης συγκλίνει σε ένα αποδεκτό τοπικό ελάχιστο του κόστους.

MLP

Οπισθοδιάδοση σφάλματος

- Με την εύρεση ενός καλού τοπικού ελαχίστου της συνάρτησης κόστους, η εκπαίδευση σταματά και τα βάρη του δικτύου παγώνουν στο τρέχον/τελικό διάνυσμα παραμέτρων.
- Μετά τη λήξη της εκπαίδευσης, έχουμε πλέον ένα στατικό, εκπαιδευμένο νευρωνικό μοντέλο, το οποίο χρησιμοποιείται μόνο για ευθέα περάσματα προτύπων ελέγχου.
- Κάθε εκτέλεση της οπισθοδιάδοσης απαιτεί απλώς **ένα ευθύ** και **ένα** διαδοχικό **αντίστροφο πέρασμα** του τρέχοντος προτύπου εκπαίδευσης, διαμέσου των επιπέδων του δικτύου (με τις τρέχουσες παραμέτρους).
  - Εξ ου και η υψηλή ταχύτητα του αλγορίθμου.

Οπισθοδιάδοση σφάλματος

- Τι συμβαίνει όμως κατά την οπισθοδιάδοση;
- Κατά το ευθύ πέρασμα:
  - Εκτελούνται οι υπολογισμοί κάθε διαδοχικού επιπέδου του δικτύου με βάση τις εξόδους του προηγούμενου επιπέδου,
  - Αποθηκεύονται: i) οι τιμές εισόδου/εξόδου του κάθε νευρώνα, και ii) η παράγωγος της συνάρτησης ενεργοποίησής του ως προς κάθε όρισμά της.
- Σημείωση: Σε πρακτικές υλοποιήσεις, κατά το ευθύ πέρασμα κατασκευάζεται ένας **υπολογιστικός γράφος**.
  - Κάθε κόμβος του αντιστοιχεί σε ένα επίπεδο, κάθε ακμή σε μία σύναψη.
  - Κατευθυνόμενος ακυκλικός γράφος, υποστηρίζει εμπρόσθιες συνάψεις μεταξύ μη διαδοχικών επιπέδων.
  - Χρησιμοποιείται στο αντίστροφο πέρασμα.

Οπισθοδιάδοση σφάλματος

MLP

- Κατά το αντίστροφο πέρασμα:
  - Πρώτα υπολογίζονται οι παράγωγοι του κόστους ως προς τις ενεργοποιήσεις του τελικού επιπέδου εξόδου.
  - Στη συνέχεια, εξάγεται αναδρομικά η μερική παράγωγος του κόστους ως προς την ενεργοποίηση, ή ως προς τη γραμμική έξοδο, κάθε νευρώνα του κάθε επιπέδου («**σφάλμα**» του αντίστοιχου νευρώνα).
    - Αξιοποιούνται τα ήδη υπολογισμένα σφάλματα του επόμενου επιπέδου, αλλά και τα δεδομένα που αποθηκεύτηκαν κατά το ευθύ πέρασμα (π.χ., στον υπολογιστικό γράφο, σε πρακτικές υλοποιήσεις).
    - Διατρέχουμε τα επίπεδα από το τέλος προς την αρχή του δικτύου.
  - Στη συνέχεια εκτιμώνται οι ζητούμενες μερικές παράγωγοι του κόστους ως προς τα συναπτικά βάρη του αντίστοιχου επιπέδου (μία μερική παράγωγος ανά σύναψη).
    - Κάθε τέτοια παράγωγος ισούται με το σφάλμα του μετασυναπτικού νευρώνα επί την ενεργοποίηση του προσυναπτικού.

MLP

Εξισώσεις οπισθοδιάδοσης

Ορισμοί:

- L: συνολικό πλήθος επιπέδων (αρίθμηση επιπέδων: 1, 2, …, L).
- C: η επιλεγμένη συνάρτηση κόστους.
- δ_j^l: το τρέχον σφάλμα του j-οστού νευρώνα του l-οστού επιπέδου.
  - Μπορεί να οριστεί ως η μερική παράγωγος του κόστους είτε ως προς την ενεργοποίηση, είτε ως προς τη γραμμική έξοδο του νευρώνα.
- $\boldsymbol{\delta}^l$: το διάνυσμα των τρεχόντων σφαλμάτων των νευρώνων του l-οστού επιπέδου (μία συνιστώσα δ_j^l ανά νευρώνα).
- $\mathbf{W}_l$: πίνακας βαρών όλων των συνάψεων με μετασυναπτικό νευρώνα στο l-οστό επίπεδο.
- w_{jk}^l: το τρέχον βάρος της σύναψης από τον k-οστό νευρώνα (στο (l-1)-οστό επίπεδο) προς τον j-οστό νευρώνα (στο l-οστό επίπεδο).
- w_{j0}^l : η τρέχουσα πόλωση του j-οστού νευρώνα του l-οστού επιπέδου.

MLP

Εξισώσεις οπισθοδιάδοσης

Ορισμοί (συνέχεια):

- $\mathbf{b}^l$: το διάνυσμα των τρεχόντων πολώσεων των νευρώνων του l-οστού επιπέδου (μία συνιστώσα w^l_{j0} ανά νευρώνα).
- a^l_j : η τρέχουσα ενεργοποίηση/τιμή εξόδου (κατά το ευθύ πέρασμα) του j-οστού νευρώνα του l-οστού επιπέδου.
- z^l_j : η τρέχουσα γραμμική έξοδος (κατά το ευθύ πέρασμα) του j-οστού νευρώνα του l-οστού επιπέδου, πριν τον μετασχηματισμό της από τη συνάρτηση ενεργοποίησης.
- $\mathbf{z}^l$: το διάνυσμα των τρεχουσών γραμμικών εξόδων όλων των νευρώνων του l-οστού επιπέδου (μία συνιστώσα z^l_j ανά νευρώνα).
- g' : η παράγωγος συνάρτηση της συνάρτησης ενεργοποίησης.
- $\mathbf{g}'(\mathbf{z}^l)$: διάνυσμα όπου συλλέγονται οι τιμές της g', υπολογισμένες ξεχωριστά στις συνιστώσες του $\mathbf{z}^l$ (μία συνιστώσα για κάθε συνιστώσα του $\mathbf{z}^l$).

MLP

Εξισώσεις οπισθοδιάδοσης

Ορισμοί (συνέχεια):

- $\nabla_a C$: διάνυσμα κλίσης της συνάρτησης κόστους ως προς τις τελικές ενεργοποιήσεις του επιπέδου εξόδου.
  - Οι συνιστώσες του υπολογίζονται για τις συγκεκριμένες τελικές τιμές ενεργοποίησης των νευρώνων εξόδου στο τέλος του ευθέος περάσματος.
- $\odot$: γινόμενο Hadamard μεταξύ δύο διανυσμάτων ή μεταξύ δύο πινάκων ίσης διάστασης.

Εξισώσεις:

- 1. $\boldsymbol{\delta}^L = \nabla_a C \odot \mathbf{g}'(\mathbf{z}^L).$
- 2. $\boldsymbol{\delta}^l = \mathbf{W}_{l+1}^T \boldsymbol{\delta}^{l+1} \odot \mathbf{g}'(\mathbf{z}^l).$
- 3. $\frac{\theta C}{\theta w_{j0}^l} = \delta_j^l.$
- 4. $\frac{\theta C}{\theta w_{jk}^l} = a_k^{l-1} \delta_j^l.$

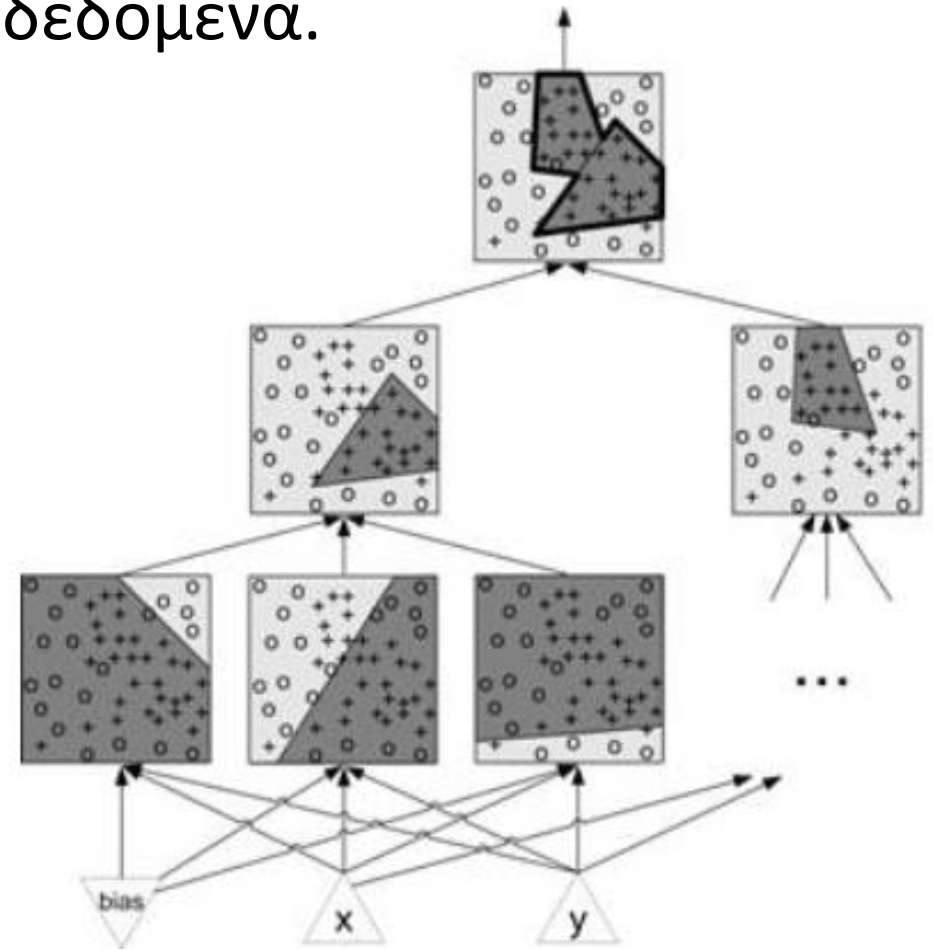
MLP

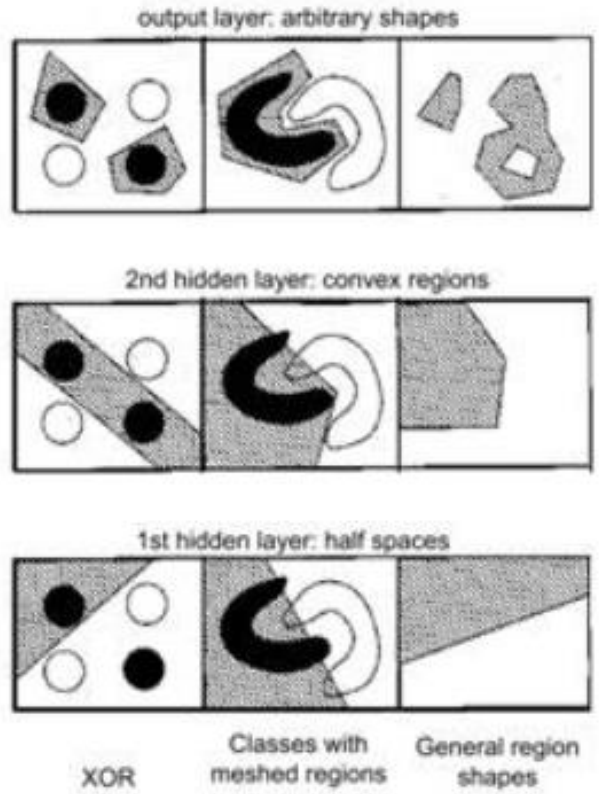
Περιοχές απόφασης

- Έστω MLP εκπαιδευμένο για ταξινόμηση.
- Κάθε κρυμμένος νευρώνας εκπαιδεύεται ώστε να κωδικοποιεί ένα υπερεπίπεδο απόφασης.
  - Παρομοίως με ένα μοντέλο λογιστικής παλινδρόμησης.
- Όμως οι νευρώνες του $(i+1)$-οστού κρυμμένου επιπέδου επεξεργάζονται τις εξόδους του i-οστού επιπέδου, **όχι** τα αρχικά δεδομένα.
  - Υπολογίζουν **τομές** υπερεπιπέδων του i-οστού επιπέδου.
- Έτσι, το ολικό δίκτυο συνθέτει περίπλοκες, μη γραμμικές, ενδεχομένως κλειστές περιοχές απόφασης, λαμβάνοντας υπόψη όλους τους νευρώνες.
- Το $(i+1)$-οστό κρυμμένο επίπεδο κωδικοποιεί πιο εκλεπτυσμένες περιοχές απόφασης από το i-οστό.

Περιοχές απόφασης

- Παράδειγμα: δυαδική ταξινόμηση σε διδιάστατα δεδομένα.





Πηγή: http://www.neural-forecasting.com/mlp_neural_nets.htm

Περιοχές απόφασης

- Όμως η εκπαίδευση γίνεται απλώς με ελαχιστοποίηση του επιλεγμένου κόστους ως προς τις παραμέτρους του δικτύου.
- **Δεν** ορίζουμε εμείς ρητά τι περιοχές απόφασης θα κωδικοποιεί κάθε επίπεδο στο τέλος της εκπαίδευσης.
- Θεμελιώδης ιδιότητα: **αυτο-οργάνωση**.
  - Κατά την εκπαίδευση, αναδύεται αυτομάτως μία πολύπλοκη τάξη από την αλληλεπίδραση απλών συστατικών (μεμονωμένων νευρώνων).
  - Η γνώση την οποία κωδικοποιεί εσωτερικά (στις τιμές παραμέτρων/βαρών του) το εκπαιδευμένο δίκτυο, *κατανέμεται* μεταξύ όλων των επιμέρους συστατικών του.
  - Η απώλεια ορισμένων μεμονωμένων νευρώνων δεν είναι καταστροφική για το εκπαιδευμένο δίκτυο.

Στοχαστική κάθοδος κλίσης

- Η εκπαίδευση με απλή κάθοδο κλίσης αξιοποιεί σε **κάθε** επανάληψη το μέσο διάνυσμα κλίσης, υπολογισμένο **ανεξάρτητα** επί όλων των προτύπων εκπαίδευσης:
  - Το ευθύ πέρασμα εκτελείται **ξεχωριστά για όλα** τα πρότυπα/παρατηρήσεις.
  - Για το καθένα υπολογίζεται κόστος και ατομικό διάνυσμα κλίσης μέσω *οπισθοδιάδοσης σφάλματος* (back-propagation).
  - Υπολογίζουμε τον μέσο όρο όλων των ατομικών διανυσμάτων κλίσης.
  - Μετατοπιζόμαστε στον διανυσματικό χώρο των παραμέτρων σε μία διεύθυνση αντίρροπη προς το μέσο διάνυσμα κλίσης, και….
  - …επόμενη επανάληψη.

Στοχαστική κάθοδος κλίσης

- Αυτό λέγεται **εκπαίδευση δέσμης** (batch training).
  - Έτσι υπολογίζεται πάντα το ακριβές διάνυσμα κλίσης.
- Η παρουσίαση όλων των προτύπων εκπαίδευσης, μία φορά το καθένα, καλείται **εποχή**.
  - Τυχαία αναδιάταξη των προτύπων στην αρχή κάθε εποχής.
  - Αποφυγή μάθησης τυχαίου θορύβου οφειλόμενου στη διάταξή τους.
- Στην εκπαίδευση δέσμης, κάθε εποχή συμπίπτει με μία επανάληψη της καθόδου κλίσης.
- Κριτήρια τερματισμού: π.χ. περίπου μηδενική τιμή κόστους, περίπου μηδενικό διάνυσμα κλίσης της συνάρτησης κόστους, κλπ.
- Εκπαιδεύουμε για όσες εποχές χρειαστεί, μέχρι να ικανοποιηθεί το επιλεγμένο κριτήριο τερματισμού.

Στοχαστική κάθοδος κλίσης

- Αντιδιαμετρική λύση: **on-line εκπαίδευση**.
- Σε κάθε επανάληψη της καθόδου κλίσης λαμβάνεται υπόψη μόνο 1 πρότυπο εκπαίδευσης, διαφορετικό κάθε φορά.
- Αν τελειώσουν τα K πρότυπα εκπαίδευσης και η κάθοδος κλίσης δεν έχει ακόμα συγκλίνει (δεν ικανοποιείται ακόμα το κριτήριο τερματισμού)…
  - Επόμενη εποχή: δίνονται ξανά από την αρχή ως είσοδοι όλα τα πρότυπα εκπαίδευσης, διαδοχικά και ανεξάρτητα.
  - Ένα μόνο πρότυπο ανά επανάληψη/ενημέρωση βαρών.
  - Κάθε εποχή περιέχει K επαναλήψεις.

Στοχαστική κάθοδος κλίσης

- Η on-line εκπαίδευση οδηγεί σε μεγαλύτερη τυχαιότητα κατά την κίνηση στον χώρο των παραμέτρων (μέχρι να βρεθεί ένα καλό τοπικό ελάχιστο του κόστους), λόγω αυξημένου θορύβου.
- Σύγκριση με την εκπαίδευση δέσμης:
  - Σύγκλιση σε περισσότερες επαναλήψεις, λόγω θορύβου.
  - Καλύτερη αποφυγή παγίδευσης σε ανεπαρκή τοπικά ελάχιστα, λόγω θορύβου.
  - Πρακτικά πολύ ταχύτερη εκτέλεση της κάθε επανάληψης, σε μεγάλα σύνολα δεδομένων εκπαίδευσης των K παρατηρήσεων/προτύπων.
    - Σε κάθε επανάληψη, απαιτείται **μόνο ένα** ευθύ και αντίστροφο πέρασμα πριν από την ενημέρωση των παραμέτρων του δικτύου/μοντέλου.
    - Αντιθέτως, η εκπαίδευση δέσμης απαιτεί K ευθέα και αντίστροφα περάσματα, ώστε να γίνει μία μόνο ενημέρωση.

Στοχαστική κάθοδος κλίσης

- Ενδιάμεση πρακτική (συνηθέστερη): **στοχαστική κάθοδος κλίσης**.
  - Στο τέλος μίας επανάληψης της εκπαίδευσης, δεν μετακινούμαστε επί του πεδίου ορισμού του κόστους (χώρος των παραμέτρων/βαρών) με βάση την κλίση όλων των K προτύπων εκπαίδευσης (δέσμη), ούτε όμως και με βάση την κλίση μόνον ενός (on-line).
  - Κινούμαστε με βάση τη μέση κλίση M προτύπων εκπαίδευσης, όπου $1 < M << K$.
  - **Μικρή δέσμη** (mini-batch) *M* προτύπων.
- Σύγκριση με την on-line εκπαίδευση:
  - Μικρότερος θόρυβος στην κίνηση επί του χώρου των παραμέτρων.
  - Ταχύτερη εκτέλεση λόγω δυνατοτήτων διανυσματικού παραλληλισμού SIMD στον επεξεργαστή.

Ορμή

- Ο ρυθμός μάθησης η καθορίζει την ταχύτητα κίνησης στον χώρο των παραμέτρων κατά την εκπαίδευση.
  - Το η προσδιορίζει το *μέτρο* της ενημέρωσης των παραμέτρων του δικτύου σε κάθε επανάληψη.
  - Αντιθέτως, η *κατεύθυνση* της ενημέρωσης δίνεται από το τρέχον διάνυσμα κλίσης.
  - Μικρό η: ομαλότερη τροχιά, περισσότερες επαναλήψεις μέχρι τη σύγκλιση.
  - Μεγάλο η: κίνδυνος για ασταθή κίνηση (π.χ., ταλαντώσεις), ενδεχομένως λιγότερες επαναλήψεις μέχρι τη σύγκλιση.
- Λύση: **ορμή** (momentum).
  - Συνθέτει τα πλεονεκτήματα και των δύο προσεγγίσεων.

Ορμή

- Η ορμή τροποποιεί **αυτομάτως** σε κάθε επανάληψη το μέτρο της τρέχουσας ενημέρωσης στον χώρο των παραμέτρων.
  - Κριτήριο: πόσο **συγγραμικό** είναι το τρέχον διάνυσμα κλίσης με τις κλίσεις των προηγούμενων επαναλήψεων.
- Διάνυσμα $\mathbf{u}$ το οποίο προαιρετικά προσθέτουμε στην εξίσωση ενημέρωσης παραμέτρων, σε κάθε επανάληψη της καθόδου κλίσης.

Κινούμενος μέσος προηγούμενων κλίσεων

$$\mathbf{u}^i = \alpha \mathbf{u}^{i-1} - \eta \nabla_x f(\mathbf{x}^i),$$

Τρέχον διάνυσμα κλίσης

$$\mathbf{x}^{i+1} = \mathbf{x}^i + \mathbf{u}^i.$$

- Το α είναι βαθμωτή υπερπαράμετρος (**συντελεστής ορμής**) η οποία λαμβάνει τιμή στο [0, 1).
  - Συνήθως, με το πέρασμα των εποχών αυξάνουμε το α και μειώνουμε το η.

Ορμή

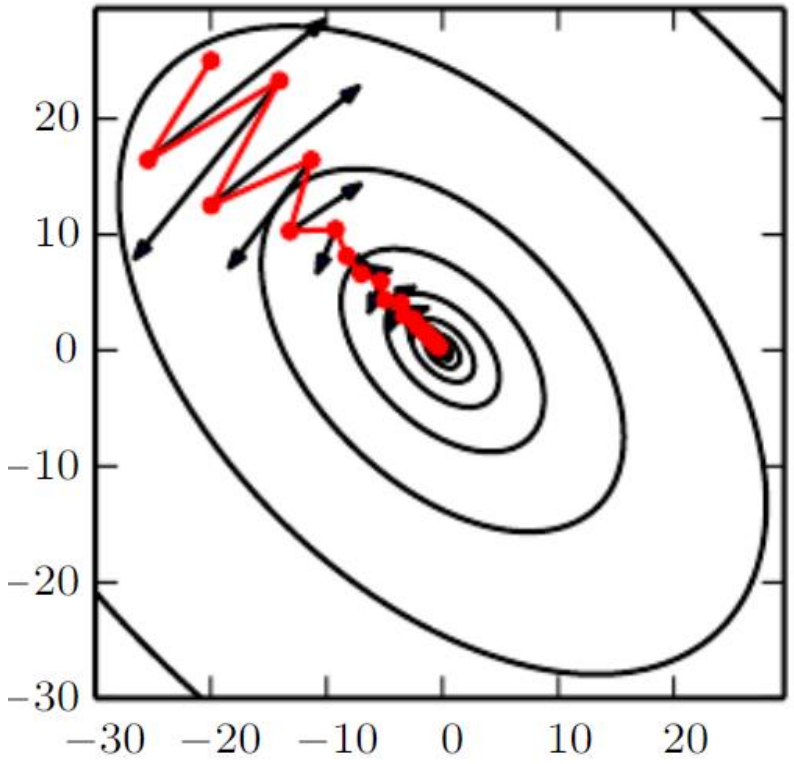
- Η ορμή μας επιτρέπει να κινούμαστε με κάποια «αδράνεια» στον χώρο των παραμέτρων.
  - Κάθε ενημέρωση επηρεάζεται από έναν κινούμενο μέσο όρο των προηγούμενων ενημερώσεων, συσσωρευμένων στο $\mathbf{u}^{i-1}$.
  - Επειδή $\alpha<1$, η συνεισφορά μίας παλιάς ενημέρωσης στο τρέχον κάθε φορά $\mathbf{u}^{i}$ μειώνεται εκθετικά με τον χρόνο.
- Γιατί συμβαίνει αυτό; Με χρήση ορμής, το διάνυσμα ενημέρωσης παραμέτρων της τρέχουσας i-οστής επανάληψης είναι βεβαρυμμένο άθροισμα του $\mathbf{u}^{i-1}$ και του τρέχοντος διανύσματος κλίσης.
  - Υπενθύμιση: το μέτρο του αθροίσματος δύο διανυσμάτων εξαρτάται από το πόσο συγγραμικά είναι τα δύο επιμέρους διανύσματα.
    - Άρα το μέτρο της ενημέρωσης των παραμέτρων στο τέλος της τρέχουσας επανάληψης ρυθμίζεται *αυτομάτως* από τη συνέπεια μεταξύ τρέχουσας κλίσης και αμέσως προηγούμενων κλίσεων.

Ορμή

- Άρα, με χρήση ορμής, το μέτρο της ενημέρωσης των βαρών στο τέλος της τρέχουσας επανάληψης είναι:
  - *Μεγαλύτερο* από όσο ορίζει το η, αν το τρέχον διάνυσμα κλίσης έχει παρόμοια διεύθυνση με τις κλίσεις των αμέσως προηγούμενων επαναλήψεων, και
  - *Μικρότερο* από όσο ορίζει το η, όταν τα διαδοχικά διανύσματα κλίσης έχουν μη συμβατές διευθύνσεις.
  - Και στις δύο περιπτώσεις προκύπτει ταχύτερη σύγκλιση σε τοπικό ελάχιστο του κόστους.
    - Εξουδετερώνονται (σε κάποιον βαθμό) οι ισχυρές ταλαντώσεις επί του χώρου των παραμέτρων (πεδίο ορισμού της υπερεπιφάνειας του κόστους).
    - Επιταχύνεται η σύγκλιση όταν έτσι κι αλλιώς δεν υπάρχουν ταλαντώσεις.

Ορμή

- Παράδειγμα: στενή χαράδρα.
  - Περιοχή της επιφάνειας κόστους με πολύ διαφορετικές καμπυλότητες σε διαφορετικές διευθύνσεις του πεδίου ορισμού.
- Η απλή κάθοδος κλίσης πάσχει από ταλαντώσεις.
  - Τα διανύσματα κλίσης είναι σχεδόν κάθετα στην «ορθή» διεύθυνση (αυτήν της συντομότερης οδού προς το τοπικό ελάχιστο).
  - Αργή σύγκλιση.
- Η χρήση ορμής (κόκκινη τροχιά) περιορίζει τις άσκοπες κινήσεις.



Πηγή: Goodfellow, 2016.

Καθολική προσέγγιση

- Έχει δειχθεί πως ένα MLP με 1 κρυμμένο επίπεδο έχει την ιδιότητα της **καθολικής προσέγγισης**.
  - Μπορεί να εκπαιδευτεί ώστε να προσεγγίζει κάθε δυνατή συνεχή συνάρτηση (απεικόνιση εισόδων σε εξόδους) με οσοδήποτε υψηλή ακρίβεια θέλουμε.
- Θεωρητικό μόνο συμπέρασμα.
  - Δεν μας πληροφορεί για:
    - Απαιτούμενο μέγεθος συνόλου δεδομένων εκπαίδευσης.
    - Απαιτούμενο πλήθος κρυμμένων νευρώνων.
    - Κατάλληλο κόστος, ή
    - Κατάλληλες τιμές υπερπαραμέτρων.
  - Ενδεχομένως να απαιτούνται άπειροι κρυμμένοι νευρώνες (άπειρη πολυπλοκότητα μοντέλου).

Ρηχά MLP

- Η αύξηση του πλήθους των κρυμμένων επιπέδων με **μη γραμμική** συνάρτηση ενεργοποίησης μειώνει την αναγκαία πολυπλοκότητα.
  - Μειώνει το απαιτούμενο ολικό πλήθος κρυμμένων νευρώνων, ώστε το δίκτυο/μοντέλο να μάθει την ίδια συνάρτηση με την ίδια ακρίβεια.
- Ωστόσο, μέχρι (περίπου) το 2010, ήταν εξαιρετικά δύσκολη η επιτυχής εκπαίδευση MLP με περισσότερα από δύο κρυμμένα επίπεδα.
  - **Υπερεκπαίδευση** λόγω υπερβολικά πολλών παραμέτρων.
  - Πρόβλημα **εξαφάνισης παραγώγων** κατά την εκπαίδευση με οπισθοδιάδοση + κάθοδο κλίσης.
- Τέχνασμα 1: ReLU ως συνάρτηση ενεργοποίησης των κρυμμένων νευρώνων, αντί της σιγμοειδούς.

Dropout

- Τέχνασμα 2: Επιπρόσθετη κανονικοποίηση, πέραν απλώς της προσθήκης ενός όρου φθοράς βαρών ($\mathcal{L}_2$-κανονικοποίηση) στην αντικειμενική συνάρτηση κόστους.

- **Κανονικοποίηση dropout**: τροποποίηση της τοπολογίας του δικτύου και όχι της αντικειμενικής συνάρτησης.
  - Προσωρινή απενεργοποίηση διαφόρων τυχαίων υποσυνόλων των κρυμμένων νευρώνων, από κοινού με τις συνάψεις τους («*σβήσιμο νευρώνων*»), κατά τη διάρκεια διαφορετικών επαναλήψεων ή εποχών της εκπαίδευσης.
  - Οι εν λόγω νευρώνες και οι συνάψεις από/προς αυτούς, όσο είναι ανενεργοί, αγνοούνται σε όλους τους υπολογισμούς ευθέως περάσματος, αντίστροφου περάσματος ή κόστους.

MLP

Dropout

- Σε κάθε επίπεδο, ορίζουμε το ποσοστό p των νευρώνων προς σβήσιμο.
  - Διαφορετικό υποσύνολο νευρώνων επιλέγεται τυχαία προς σβήσιμο σε κάθε επανάληψη
  - Πάντα όμως το πλήθος τους συνιστά το p% όλων των νευρώνων του επιπέδου.
- Στη φάση ελέγχου, το τελικό μοντέλο συμπεριλαμβάνει **όλους** τους κρυμμένους νευρώνες.
  - Όμως λειτουργεί σαν σύμπλεγμα διαφορετικών δικτύων, υπερεκπαιδευμένων κατά διαφορετικό τρόπο.
  - Τα διαφορετικά σφάλματα λόγω υπερεκπαίδευσης αλληλοεξουδετερώνονται και η ικανότητα γενίκευσης του δικτύου αυξάνεται.

Συνάψεις συντόμευσης

- Τέχνασμα 3: Συνάψεις συντόμευσης.
- Η εξαφάνιση παραγώγων οφείλεται σε διαδοχικούς πολλαπλασιασμούς μεταξύ μικρών τιμών εγγύς του 0, κατά την οπισθοδιάδοση σφάλματος.
  - Οι μικρές αυτές τιμές προκύπτουν συνήθως από τις μη γραμμικές συναρτήσεις ενεργοποίησης των κρυμμένων νευρώνων.
- Μία **σύναψη συντόμευσης** (skip connection) έχει στατικό βάρος 1 και συνδέει απευθείας κάθε νευρώνα του τρέχοντος επιπέδου με τον αντίστοιχης θέσης νευρώνα του μεθεπόμενου επιπέδου.
- Οι εν λόγω συνάψεις προσπερνούν ένα κρυμμένο επίπεδο και δεν προσθέτουν παραμέτρους στο δίκτυο, αφού έχουν γνωστό στατικό βάρος.

MLP

Συνάψεις συντόμευσης

- Το ευθύ πέρασμα τροποποιείται.
  - Προστίθεται η ενεργοποίηση του κάθε νευρώνα του προπερασμένου επιπέδου l-2 στη γραμμική έξοδο του αντίστοιχου νευρώνα του τρέχοντος επιπέδου l, πριν την εφαρμογή της συνάρτησης ενεργοποίησης.
- Πρέπει να τροποποιηθούν καταλλήλως οι εξισώσεις οπισθοδιάδοσης σφάλματος, ώστε να γίνεται σωστά η εκπαίδευση.
  - Κατά το αντίστροφο πέρασμα, παρέχεται μία *δίοδος απευθείας μετάδοσης του σφάλματος κάθε νευρώνα*.
  - Μετάδοση προς τα πιο πρώιμα επίπεδα (π.χ., από το l στο l-2), αντί μόνο για το αμέσως προηγούμενο επίπεδο (l-1).

MLP

Συνάψεις συντόμευσης

- Υπενθύμιση οπισθοδιάδοσης:
  - Έστω δ η τρέχουσα παράγωγος της συνάρτησης κόστους ως προς την ενεργοποίηση, ή ως προς τη γραμμική έξοδο ενός συγκεκριμένου νευρώνα (*σφάλμα του νευρώνα*).
  - Κατά την οπισθοδιάδοση, το τρέχον κάθε φορά σφάλμα δ ενός νευρώνα του επιπέδου l υπολογίζεται λαμβάνοντας υπόψη:
    - Τη γραμμική του έξοδο κατά το τελευταίο ευθύ πέρασμα,
    - Τα σφάλματα δ του επόμενου επιπέδου l+1, και
    - Τα συναπτικά βάρη από αυτόν προς το επίπεδο l+1.
  - Στη συνέχεια, έχοντας πλέον υπολογίσει το σφάλμα δ, οι παράγωγοι του κόστους ως προς τα βάρη των συνάψεων από το επίπεδο l-1 προς το l υπολογίζονται με βάση το δ.
    - Αυτές είναι οι μερικές παράγωγοι που μας ενδιαφέρουν, αφού αποτελούν συνιστώσες του τρέχοντος διανύσματος κλίσης.

Συνάψεις συντόμευσης

- Μέσω της σύναψης συντόμευσης, το ίδιο δ μεταδίδεται αυτούσιο και προς το επίπεδο l-2 κατά τον υπολογισμό των δικών του δ.
  - Απλώς προστίθεται στα σφάλματα του επιπέδου l-1, **χωρίς τροποποίηση** διαμέσου συναπτικών βαρών ή συνάρτησης ενεργοποίησης.
  - Μία τέτοια τροποποίηση θα συνεπαγόταν πολλαπλασιασμό με αριθμό εγγύς του 0 (π.χ., παράγωγος της σιγμοειδούς σε κορεσμένη περιοχή του πεδίου ορισμού της).
- Άρα το κύριο όφελος προκύπτει από την επιπρόσθετη οπισθοδιάδοση των νευρωνικών σφαλμάτων, χωρίς παρεμβολή της συνάρτησης ενεργοποίησης.

Βαθιά μάθηση

MLP

- Τα εν λόγω τεχνάσματα επιτρέπουν πλέον την εκπαίδευση βαθιών MLP με πολλαπλά κρυμμένα επίπεδα.
  - **Βαθιά μάθηση:** εκπαίδευση νευρωνικών δικτύων εμπρόσθιας τροφοδότησης (π.χ., MLP, CNN, κλπ.) τα οποία έχουν πολύ περισσότερα από 2 κρυμμένα επίπεδα.
- Το μεγάλο βάθος είναι πολύ σημαντικό για την επίτευξη καλής γενίκευσης, αφού επιτρέπει τη μάθηση *αφηρημένων χαρακτηριστικών / αναπαραστάσεων υψηλότερου επιπέδου* από τα δεδομένα.
  - Τα διαδοχικά επίπεδα μετασχηματίζουν τα δεδομένα, έτσι ώστε οι διανυσματικές αναπαραστάσεις τους σταδιακά να προσαρμοστούν στο συγκεκριμένο πρόβλημα το οποίο κωδικοποιεί η συνάρτηση κόστους.

Βαθιά μάθηση

MLP

- Στη ρηχή μάθηση, το διάνυσμα χαρακτηριστικών εισόδου συνήθως είναι μία χειροκίνητα κατασκευασμένη περιγραφή της εισόδου.
  - Π.χ., ένας «χειροποίητος» περιγραφέας 500 διαστάσεων, υπολογισμένος με κάποιον αλγόριθμο, ο οποίος αναπαριστά μία ολόκληρη εικόνα RGB.
  - Οι αλγόριθμοι περιγραφής και οι χειροποίητοι περιγραφείς που παράγουν έχουν μειονεκτήματα, δεν κατορθώνουν πάντα να συλλάβουν όλα τα σημαντικά γνωρίσματα της εισόδου.
- Στη βαθιά μάθηση, το διάνυσμα εισόδου είναι η αρχική, ακατέργαστη παρατήρηση και όχι κάποια σύνοψή της.
  - Π.χ., ένα διάνυσμα διάστασης K x N x 3, για μία RGB εικόνα με ανάλυση K x N pixel.

Βαθιά μάθηση

MLP

- Κάθε επίπεδο του δικτύου μετασχηματίζει **μη γραμμικά** το διάνυσμα εισόδου του και το προβάλει σε έναν διαφορετικό διανυσματικό χώρο.
- Το επόμενο / διάδοχο επίπεδο λαμβάνει ως είσοδο την εν λόγω προβολή και την επεξεργάζεται.
- Με σωστή εκπαίδευση, το τελικό επίπεδο εξόδου επεξεργάζεται πλέον «ιδανικές» περιγραφές / αναπαραστάσεις των αρχικών δεδομένων.
  - Π.χ., γραμμικά διαχωρίσιμες κλάσεις, στην περίπτωση ταξινόμησης.
- Δεν χρειαζόμαστε πια χειροποίητους αλγορίθμους περιγραφής!
  - Κατά την εκπαίδευση, το δίκτυο μαθαίνει να παράγει μόνο του, εσωτερικά, κατάλληλες περιγραφές από τα αρχικά ακατέργαστα δεδομένα.

Βαθιά μάθηση

MLP

- Με εκπαίδευση σε πολύ μεγάλα σύνολα δεδομένων, το δίκτυο σταδιακά αυτο-οργανώνεται ώστε, κατά το ευθύ πέρασμα, κάθε επίπεδο να παράγει μία νέα, κατάλληλη περιγραφή / αναπαράσταση της εισόδου του.
- Καθώς προχωρούμε από πιο πρώιμα επίπεδα σε μεταγενέστερα επίπεδα, οι αναπαραστάσεις αυτές είναι όλο και πιο αφηρημένες:
  - Όλο και πιο απομακρυσμένες από το αρχικό διάνυσμα εισόδου.
  - Όλο και καταλληλότερες για το ζητούμενο πρόβλημα (π.χ., ταξινόμηση) που λύνει το επίπεδο εξόδου.

Συναρτήσεις κόστους

- Ως συναρτήσεις κόστους κατά την εκπαίδευση, συνήθως χρησιμοποιούνται Εκτιμητές Μέγιστης Πιθανοφάνειας (MLE).
- Η γενική μορφή είναι: $C(\mathbf{w}) = -\mathbb{E}_{\mathbf{x},\mathbf{y}\sim\hat{p}_{\text{data}}} \log p_{\text{model}}(\boldsymbol{y} \mid \boldsymbol{x})$.
  - Ισοδυναμεί με ελαχιστοποίηση της διασταυρούμενης εντροπίας μεταξύ της εμπειρικής υπό συνθήκη κατανομής και της υπό συνθήκη κατανομής του επιλεγμένου παραμετρικού μοντέλου.
- Η χρήση συγκεκριμένης υπό συνθήκη κατανομής ως μοντέλου δίνει ως αποτέλεσμα μία συγκεκριμένη συνάρτηση κόστους.
  - Παράδειγμα: Όπως στη γραμμική παλινδρόμηση, η χρήση **γκαουσιανού μοντέλου** με αναμενόμενη τιμή την τελική έξοδο/πρόβλεψη του δικτύου για πρότυπο εισόδου $\mathbf{x}$, δίνει εν τέλει το Μέσο Τετραγωνικό Σφάλμα (MSE).
  - *Διαφορά από τη γραμμική παλινδρόμηση*: η πρόβλεψη εδώ είναι σύνθεση μη γραμμικών μετασχηματισμών του $\mathbf{x}$, όχι απλώς μία γραμμική πράξη (εσωτερικό γινόμενο) $\mathbf{x}^T\mathbf{w}$.
  - *Διαφορά από τη γραμμική παλινδρόμηση*: ισχύει και για πολυδιάστατες ετικέτες.

MLP

Συναρτήσεις κόστους

- Ατομικές συναρτήσεις κόστους (loss), *για μεμονωμένο πρότυπο εκπαίδευσης* $\mathbf{x}$ με αντίστοιχη N-διάστατη ετικέτα $\mathbf{y}$.
- **Μέσο Τετραγωνικό Σφάλμα** (MSE), συνήθως για παλινδρόμηση και με N γραμμικούς νευρώνες εξόδου:
  - $C(\mathbf{w}) = \frac{1}{N}||\mathbf{y} - \mathbf{f}(\mathbf{x}; \mathbf{w})||_2^2$.
- **Δυαδική Διασταυρούμενη Εντροπία** (BCE), για δυαδική ταξινόμηση (κλάση 0, κλάση 1) με έναν σιγμοειδή νευρώνα εξόδου (N = 1):
  - Η έξοδος είναι η πιθανότητα το $\mathbf{x}$ να ανήκει στην κλάση 1.
  - Προκύπτει από MLE με μοντέλο Bernoulli.
  - $C(\mathbf{w}) = -[y \log\big(f(\mathbf{x}; \mathbf{w})\big) + (1 - y) \log\big(1 - f(\mathbf{x}; \mathbf{w})\big)]$.
- **Κατηγορική Διασταυρούμενη Εντροπία** (CE) για ταξινόμηση K κλάσεων, με K softmax νευρώνες εξόδου (N = K):
  - Η έξοδος προσεγγίζει one-hot διάνυσμα, η ετικέτα είναι one-hot.
  - $C(\mathbf{w}) = -\sum_{i=1}^{N}[yi \ \log\big(f_i(\mathbf{x}; \mathbf{w})\big)]$.

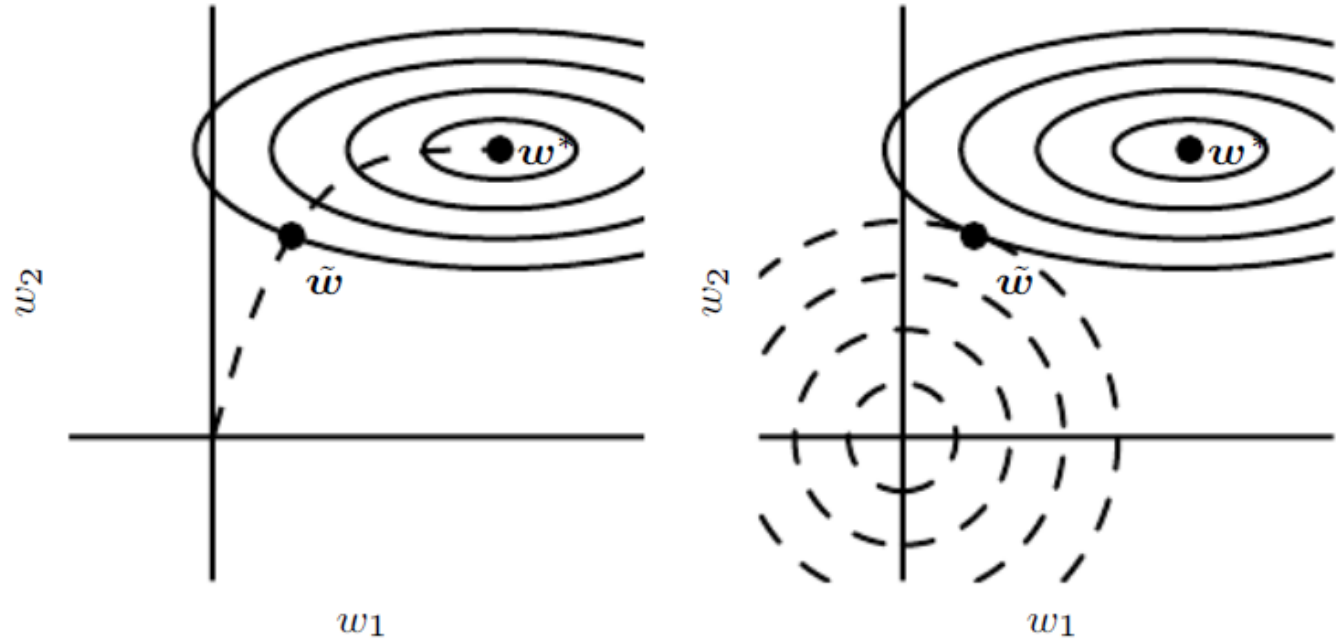
MLP

Εναλλακτικοί τρόποι κανονικοποίησης

- Το πλήθος E των εποχών εκπαίδευσης είναι υπερπαράμετρος προς βελτιστοποίηση.
  - Υπερβολικά λίγες εποχές ⟶ υποεκπαίδευση.
  - Υπερβολικά πολλές εποχές ⟶ υπερεκπαίδευση.
- Λύση: **πρόωρο σταμάτημα** (early stopping).
  - Όπως όλες τις υπερπαραμέτρους, βελτιστοποιούμε το E στο σύνολο επικύρωσης.
    - Το τρέχον μέσο κόστος στο σύνολο *εκπαίδευσης* υπολογίζεται στο τέλος κάθε εποχής.
    - Ανά M εποχές (συνηθισμένες τιμές $M = 1$, 5, ή 10), υπολογίζουμε επιπροσθέτως και το μέσο κόστος του τρέχοντος μοντέλου στο σύνολο *επικύρωσης*.
    - Αρχικώς, το κόστος εκπαίδευσης και το κόστος επικύρωσης μειώνονται ταυτόχρονα με το πέρασμα των εποχών.
    - Από ένα σημείο κι έπειτα, το πρώτο συνεχίζει να μειώνεται ενώ το δεύτερο αρχίζει να αυξάνεται (εμφάνιση υπερεκπαίδευσης, μειωμένη ικανότητα γενίκευσης).
    - Τότε αποθηκεύουμε το τρέχον μοντέλο (έστω στην εποχή E), συνεχίζουμε την εκπαίδευση και επαναλαμβάνουμε.
    - Στο τέλος (έστω στην εποχή $F > E$) επιστρέφουμε ως οριστικό αποτέλεσμα το αποθηκευμένο μοντέλο με το συνολικά μικρότερο κόστος επικύρωσης.

Εναλλακτικοί τρόποι κανονικοποίησης

MLP



Πηγή: Goodfellow, 2016.

MLP

Εναλλακτικοί τρόποι κανονικοποίησης

- Ενίσχυση συνόλου εκπαίδευσης.
  - Π.χ., με **επιπρόσθετα συνθετικά δεδομένα**, παρόμοια (αλλά όχι πανομοιότυπα) με τα υπάρχοντα.
  - Μεγαλύτερα σύνολα δεδομένων απαιτούν πολυπλοκότερα μοντέλα.
  - Άρα, κρατώντας σταθερή την πολυπλοκότητα του μοντέλου και τους μηχανισμούς κανονικοποίησης (π.χ., φθορά βαρών, dropout, πρόωρο σταμάτημα, κλπ.), η εναπομένουσα υπερεκπαίδευση μπορεί να μειωθεί με απλή αύξηση του μεγέθους του συνόλου εκπαίδευσης.
- Εκπαίδευση πολλαπλών προβλημάτων.
  - Το **ίδιο μοντέλο** εκπαιδεύεται σε **ένα κοινό σύνολο δεδομένων**, έτσι ώστε να λύνει **ταυτόχρονα δύο διαφορετικά προβλήματα** (π.χ., ένα ανεπίβλεπτο και ένα επιβλεπόμενο).
  - Συνήθως αθροίζονται δύο συναρτήσεις κόστους.
    - Στον υπολογισμό των τιμών τους, λαμβάνονται υπόψη ως προβλέψεις του δικτύου οι έξοδοι είτε του ίδιου, είτε διαφορετικού επιπέδου.
  - Το μοντέλο αναγκάζεται να εξάγει περισσότερη γνώση από τα ίδια δεδομένα, με τις ίδιες παραμέτρους.
    - Περιορίζεται το περιθώριο απομνημόνευσης των ιδιαιτεροτήτων/θορύβου του συνόλου εκπαίδευσης.

Thank you for your attention!

Q & A

Contact: imademlis@aueb.gr