

Ομαλοποίηση (Regularization)

Βαθιά Μηχανική Μάθηση

ΔΠΜΣ Επιστήμης Δεδομένων & Μηχανικής Μάθησης

Εθνικό Μετσόβιο Πολυτεχνείο

Γιώργος Αλεξανδρίδης

Δομή της Παρουσίαση

1. Εισαγωγικές Έννοιες
 - Προσαρμογή
 - Ομαλοποίηση
2. Όροι ποινής
 - Προσθήκη όρων ποινής
 - Ομαλοποίηση L^2
 - Ομαλοποίηση L^1
 - Αραιές Αναπαραστάσεις
3. Πρόωρος Τερματισμός
4. Επαύξηση Δεδομένων
5. Μέθοδοι Ensemble
 - Bagging
 - Dropout
6. Βιβλιογραφία

Εισαγωγικές Έννοιες

Προσαρμογή

Γενίκευση

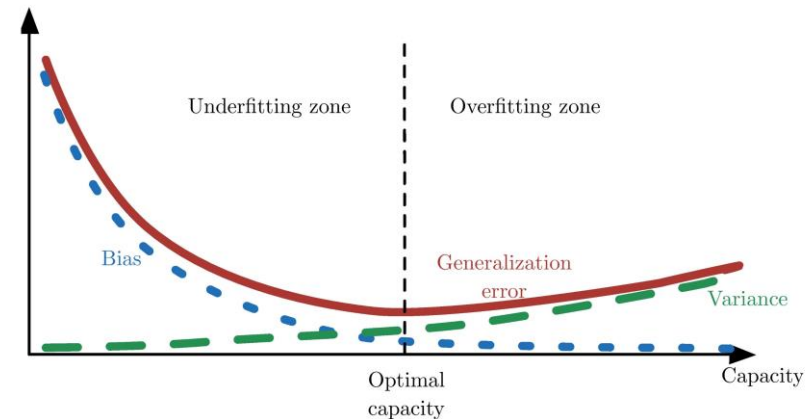
- **Επιθυμητό αποτέλεσμα** κάθε διαδικασίας μάθησης είναι το μοντέλο να έχει **καλή απόδοση** **όχι μόνο** στα δεδομένα που έχει **εκπαιδευτεί** αλλά και σε **νέα δεδομένα**, μια διαδικασία που είναι γνωστή ως **γενίκευση** (*generalization*)
 - Μικρό **σφάλμα εκπαίδευσης** (*training error*)
 - Μικρή διαφορά μεταξύ σφάλματος εκπαίδευσης και **σφάλματος γενίκευσης** (ή αλλιώς **σφάλματος ελέγχου** – *test error*)

Στατιστική Θεωρία μάθησης

- *Statistical Learning Theory*
- Βασική υπόθεση εργασίας
- Κάθε πρόβλημα μοντελοποιείται ως μια **διαδικασία παραγωγής δεδομένων** που προκύπτουν από **άγνωστη** σε εμάς **κατανομή**
- Κάθε πιθανό δείγμα των δεδομένων, είτε ανήκει στο σύνολο εκπαίδευσης είτε στο σύνολο ελέγχου, παράγεται **ανεξάρτητα** (*independent*) από τα υπόλοιπα, από την **ίδια κατανομή** (*identically distributed*) \Rightarrow *i.i.d assumptions*

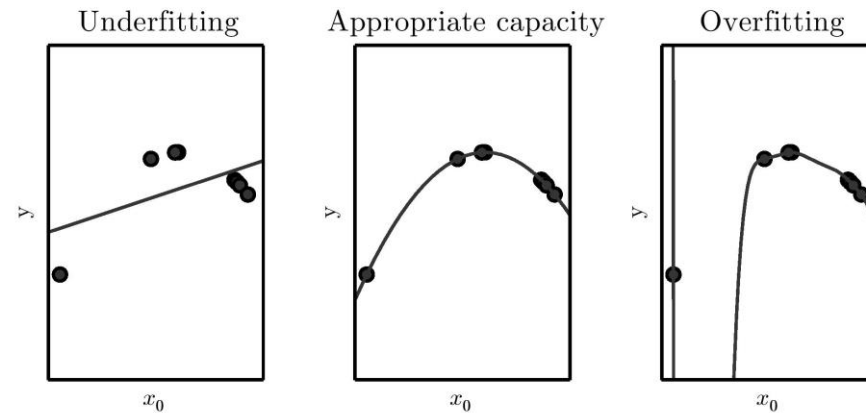
Υπερπροσαρμογή και Υποπροσαρμογή

- **Υποπροσαρμογή (*underfitting*)**
 - Το μοντέλο δεν μπορεί να μειώσει το **σφάλμα εκπαίδευσης**
- **Υπερπροσαρμογή (*overfitting*)**
 - Το μοντέλο δεν μπορεί να μειώσει το «**χάσμα**» μεταξύ **σφάλματος εκπαίδευσης και ελέγχου**
- **Χωρητικότητα (*capacity*)**
 - Καθορίζει την ικανότητα του μοντέλου να γενικεύει
 - Σχετίζεται άμεσα με τον **χώρο υποθέσεων (*hypothesis space*)** του μοντέλου



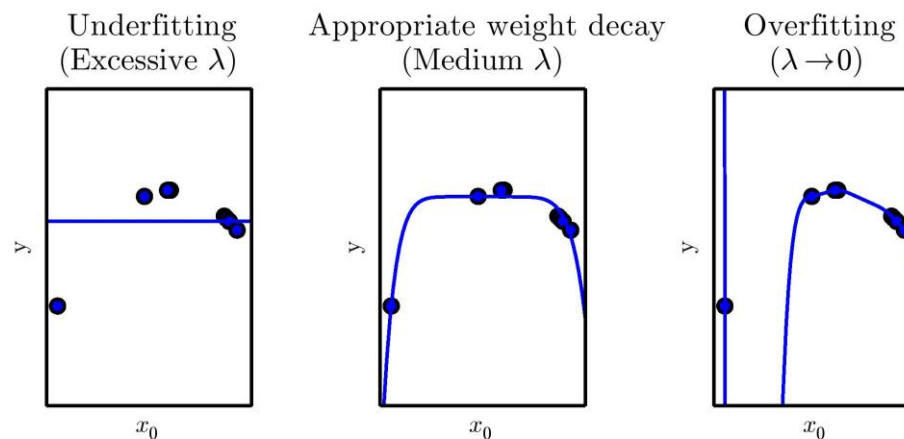
Χωρητικότητα και Δυνατότητα Μάθησης

- Τα μοντέλα μηχανικής μάθησης έχουν τη **βέλτιστη απόδοση** όταν η **χωρητικότητά** τους είναι η **κατάλληλη** για το **πρόβλημα** που τους ζητείται να μάθουν
- Τα μοντέλα με **ανεπαρκή χωρητικότητα** **δεν μπορούν** να επιλύσουν πολύπλοκα προβλήματα
- Τα μοντέλα με **υψηλή χωρητικότητα** **επιλύουν** πολύπλοκα προβλήματα, αλλά ενδέχεται να εμφανίσουν **υπερπροσαρμογή**



Έλεγχος του χώρου υποθέσεων

- Στις περισσότερες περιπτώσεις είναι **δύσκολο** να καθοριστεί επακριβώς εκ των προτέρων ο **χώρος υποθέσεων** ενός μοντέλου
- Μπορούμε, ωστόσο να ξεκινήσουμε από ένα μεγάλο χώρο υποθέσεων και όσο προχωράει η εκπαίδευση να **ενισχύουμε** ή να **εξασθενούμε** συγκεκριμένα χαρακτηριστικά

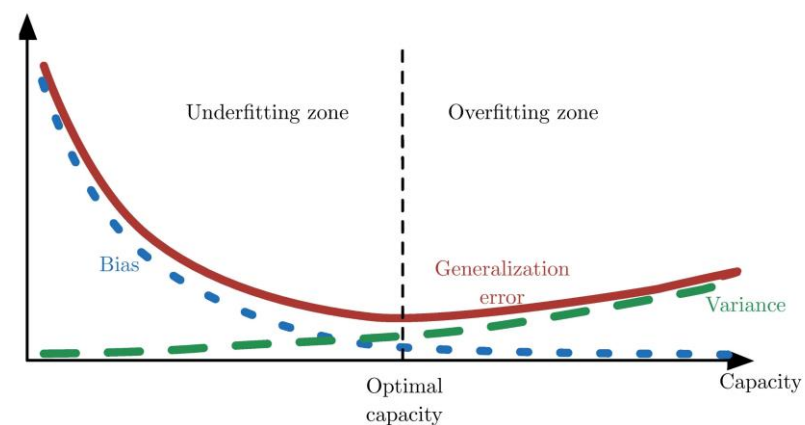


Μεροληψία και Διακύμανση

- Μετρούν δύο διαφορετικά είδη σφάλματος σε έναν εκτιμητή
- **Μεροληψία** (*bias*)
 - Μέτρο της αναμενόμενης απόκλισης της πρόβλεψης του εκτιμητή από την πραγματική τιμή
 - **Υψηλή μεροληψία** σημαίνει ότι το μοντέλο έχει «**υπεραπλουστεύσει**» τα δεδομένα
- **Διακύμανση** (*variance*)
 - Μέτρο της απόκλισης των προβλέψεων του εκτιμητή σε «παραπλήσια» δεδομένα
 - **Υψηλή διακύμανση** σημαίνει ότι το μοντέλο δεν μπορεί να **γενικεύσει**

Ισορροπία μεταξύ μεροληψίας και διακύμανσης

- Bias-variance trade-off
- Στόχος κάθε διαδικασίας μάθησης
- Στόχος της **ομαλοποίησης** είναι να **μειώσουμε** τη διακύμανση του μοντέλου, **χωρίς** να **επιηρεαστεί** η **μεροληψία** του



Εισαγωγικές Έννοιες

Ομαλοποίηση

Ομαλοποίηση

- Ορισμός
 - *Κάθε τροποποίηση ενός αλγορίθμου μάθησης που στόχο έχει τη μείωση του σφάλματος γενίκευσης του και όχι του σφάλματος εκπαίδευσής του καλείται ομαλοποίηση*
- Τεχνικές
 - Προσθήκη περιορισμών στο μοντέλο
 - Επιπλέον όροι στην αντικειμενική συνάρτηση
 - Εισαγωγή εκ των προτέρων γνώσης
 - Προτίμηση απλούστερων μοντέλων
 - Μέθοδοι ensemble
 - Συνδυασμός πολλαπλών υποθέσεων για την περιγραφή των δεδομένων εκπαίδευσης

Όροι ποιηής

Όροι ποινής (penalties)

- Μείωση χωρητικότητας μοντέλου μέσω προσθήκης **όρου ποινής** Ω στην αντικειμενική συνάρτηση J
- Από τις πρώτες προσεγγίσεις ομαλοποίησης στο χώρο της μηχανικής μάθησης
 - Χρησιμοποιείται και στα βαθιά δίκτυα
- Πρόβλημα ταξινόμησης: $\tilde{J}(\theta; \mathbf{X}, \mathbf{y}) = J(\theta; \mathbf{X}, \mathbf{y}) + \alpha\Omega(\theta)$
 - \tilde{J} ομαλοποιημένη αντικειμενική συνάρτηση
 - θ χώρος παραμέτρων, \mathbf{X} δεδομένα εκπαίδευσης, \mathbf{y} ετικέτες εξόδου
 - α **υπερπαραμέτρος** που ρυθμίζει τη **συνεισφορά** του παράγοντα της **ομαλοποίησης** στην αντικειμενική συνάρτηση
 - Για $\alpha = 0$ δεν έχουμε **καθόλου** ομαλοποίηση
 - Όσο το α μεγαλώνει, η ομαλοποίηση παίζει μεγαλύτερο ρόλο
- Στην πράξη, η ομαλοποίηση αφορά τα **βάρη** και όχι τις **πολώσεις**
 - Ομαλοποίηση στις πολώσεις μπορεί να οδηγήσει σε προβλήματα *υποπροσαρμογής*
 - $\tilde{J}(w; \mathbf{X}, \mathbf{y}) = J(w; \mathbf{X}, \mathbf{y}) + \alpha\Omega(w)$

Όροι ποινηής

Ομαλοποίηση L^2

Ομαλοποίηση L^2

- Εναλλακτικές ονομασίες
 - Φθορά βαρών (*weight decay*), **αμφικλινής παλινδρόμηση** (*ridge regression*), **κανονικοποίηση Tikhonov** (*Tikhonov regularization*)
- Προσθήκη όρου ομαλοποίησης βαρών $\Omega(\boldsymbol{\theta}) = \frac{1}{2} \|\mathbf{w}\|_2^2 = \frac{1}{2} \mathbf{w}^T \mathbf{w}$ στην J
 - $\tilde{J}(\mathbf{w}; \mathbf{X}, \mathbf{y}) = J(\mathbf{w}; \mathbf{X}, \mathbf{y}) + \alpha \frac{1}{2} \mathbf{w}^T \mathbf{w}$
- Ενημέρωση βαρών μέσω στοχαστικής κατάβασης κλίσης (SGD)
 - Υπολογισμός κλίσης
 - $\nabla_{\mathbf{w}} \tilde{J}(\mathbf{w}; \mathbf{X}, \mathbf{y}) = \nabla_{\mathbf{w}} J(\mathbf{w}; \mathbf{X}, \mathbf{y}) + \alpha \mathbf{w}$
 - Ενημέρωση βαρών
 - $\mathbf{w}^{\tau+1} \leftarrow \mathbf{w}^{\tau} - \epsilon (\nabla_{\mathbf{w}^{\tau}} J(\mathbf{w}^{\tau}; \mathbf{X}, \mathbf{y}) + \alpha \mathbf{w}^{\tau}) \Rightarrow \mathbf{w}^{\tau+1} \leftarrow (1 - \epsilon \alpha) \mathbf{w}^{\tau} - \epsilon \nabla_{\mathbf{w}^{\tau}} J(\mathbf{w}^{\tau}; \mathbf{X}, \mathbf{y})$
 - Το αποτέλεσμα είναι να **μειώνεται** («φθεΐρεται») το εύρος του διανύσματος των βαρών κατά παράγοντα σε κάθε βήμα.

Ομαλοποίηση L^2 : Επίδραση στη διαδικασία μάθησης (1/3)

- Έστω w^* το διάνυσμα βαρών που ελαχιστοποιεί την J
- Αντικατάσταση της J από την τετραγωνική της προσέγγιση \hat{J} γύρω από το w^*
 - Ειδικά για προβλήματα γραμμικής παλινδρόμησης (*linear regression*) όπου η αντικειμενική συνάρτηση υπολογίζει διαφορές τετραγώνων (π.χ. MSE), η προσέγγιση είναι τέλεια
 - $\hat{J}(w) = J(w^*) + \frac{1}{2} (w - w^*)^T H (w - w^*)$
 - **Δεν υπάρχει** πρωτοβάθμιος όρος μιας εξ' ορισμού είναι **ελάχιστο** και άρα η **κλίση** γύρω από το $w - w^*$ είναι (σχεδόν) **μηδενική**
 - **H**: Εσσιανός Πίνακας (*Hessian Matrix*) όλων των $\frac{\partial^2 J}{\partial w_i \partial w_j}$
 - Επειδή βρισκόμαστε γύρω από ελάχιστο, ο **H** είναι **θετικά ημι-καθορισμένος**

Ομαλοποίηση L^2 : Επίδραση στη διαδικασία μάθησης (2/3)

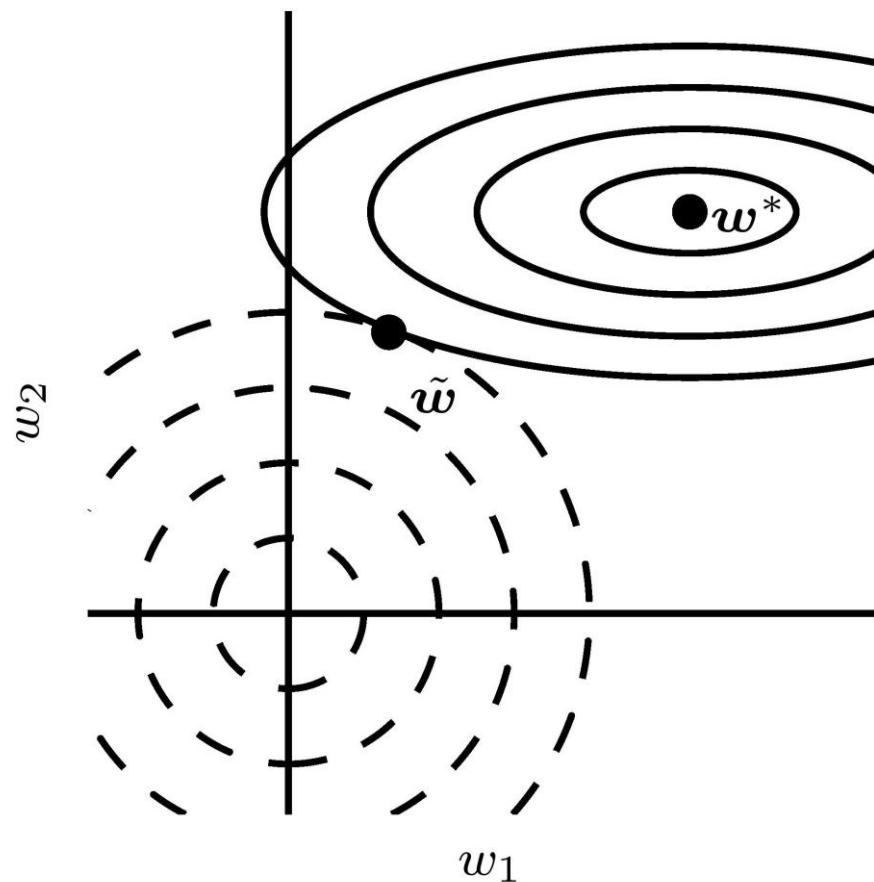
- Η \hat{J} ελαχιστοποιείται στα σημεία όπου η κλίση της $\nabla_w \hat{J}(w; \mathbf{X}, \mathbf{y}) = \mathbf{H}(w - w^*)$ γίνεται ίση με το 0
- Προσθέτοντας τον όρο ποινής $\alpha \frac{1}{2} w^T w$ το τοπικό ελάχιστο αλλάζει και γίνεται πλέον \tilde{w}
 - $\nabla_w \left(\hat{J}(\tilde{w}) + \frac{\alpha}{2} \tilde{w}^T \tilde{w} \right) = 0 \Rightarrow \mathbf{H}(\tilde{w} - w^*) + \alpha \tilde{w} = 0 \Rightarrow \tilde{w} = (\mathbf{H} + \alpha \mathbf{I})^{-1} \mathbf{H} w^*$
- Όταν $\alpha \rightarrow 0$, το \tilde{w} προσεγγίζει το w^*
- Όταν $\alpha \neq 0$, χρησιμοποιούμε αποσύνθεση ιδιοτιμών (eigen decomposition) $\mathbf{H} = \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^T$
 - Εφικτό μιας και \mathbf{H} συμμετρικός με πραγματικές τιμές
 - $\mathbf{\Lambda}$ διαγώνιος πίνακας ιδιοτιμών, \mathbf{Q} ορθο-κανονικός πίνακας ιδιοδιανυσμάτων
 - $\tilde{w} = \mathbf{Q}(\mathbf{\Lambda} + \alpha \mathbf{I})^{-1} \mathbf{\Lambda} \mathbf{Q}^T w^*$

Ομαλοποίηση L^2 : Επίδραση στη διαδικασία μάθησης (3/3)

- Ουσιαστικά το w^* **αναπροσαρμόζεται** στην κατεύθυνση των αξόνων που ορίζονται από τα ιδιοδιανύσματα του H κατά ένα **παράγοντα** $\frac{\lambda_i}{\lambda_i + \alpha}$
 - Για μεγάλες ιδιοτιμές ($\lambda_i \gg \alpha$), η επίδραση της ομαλοποίησης είναι **πολύ μικρή**
 - Για μικρές ιδιοτιμές ($\lambda_i \ll \alpha$), η επίδραση της ομαλοποίησης είναι **πολύ μεγάλη**
 - Συρρικνώνει την επίδραση των αντίστοιχων ιδιοδιανυσμάτων στο 0

Ομαλοποίηση L^2 : 1^ο Παράδειγμα

- **Συνεχείς** καμπύλες: **Ίδιες** τιμές J
- **Διακεκομμένες** καμπύλες: **Ίδιες** τιμές \hat{J}
- **Σημείο ισορροπίας**: \tilde{w}
- 1^η διάσταση: Ιδιοτιμή του \mathbf{H} χαμηλή
 - Η τιμή της J δεν μεταβάλλεται σημαντικά σε αυτό τον άξονα.
 - Επίδραση ομαλοποιητής **σημαντική**
 - «Τραβάει» το w_1 προς το 0
- 2^η διάσταση: Ιδιοτιμή του \mathbf{H} υψηλή
 - Η τιμή της J είναι **«ευαίσθητη»** σε **μικρές μεταβολές** του w_2
 - Επίδραση ομαλοποιητή **μικρή**



Ομαλοποίηση L^2 : 2^ο Παράδειγμα

- Γραμμική Παλινδρόμηση
 - **Συνάρτηση κόστους**: άθροισμα τετραγώνων σφάλματος:
 - $J(\mathbf{w}) = (\mathbf{X}\mathbf{w} - \mathbf{y})^T (\mathbf{X}\mathbf{w} - \mathbf{y})$
 - Βέλτιστο διάνυσμα βαρών: $\mathbf{w}^* = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$
 - Προσθήκη όρου ομαλοποίησης L^2
 - $J(\mathbf{w}) = (\mathbf{X}\mathbf{w} - \mathbf{y})^T (\mathbf{X}\mathbf{w} - \mathbf{y}) + \frac{\alpha}{2} \mathbf{w}^T \mathbf{w}$
 - Βέλτιστο διάνυσμα βαρών: $\tilde{\mathbf{w}} = (\mathbf{X}^T \mathbf{X} + \alpha \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$
- Ο όρος $\mathbf{X}^T \mathbf{X}$ είναι **ανάλογος** του **πίνακα συνδιασποράς** των χαρακτηριστικών της εισόδου
 - **Διαγώνιες τιμές**: Αντιστοιχούν στη διακύμανση των χαρακτηριστικών της εισόδου
 - Η ομαλοποίηση L^2 **αναγκάζει** τον αλγόριθμο μάθησης να θεωρήσει ότι η είσοδος παρουσιάζει **μεγαλύτερη διακύμανση** και συνεπώς να **μικρύνει** τα **βάρη** εκείνα που εμφανίζουν **μικρότερη συνδιασπορά**.

Όροι ποινηής

Ομαλοποίηση L^1

Ομαλοποίηση L^1

- Προσθήκη όρου ομαλοποίησης βαρών $\Omega(\boldsymbol{\theta}) = \|\mathbf{w}\|_1$ στην J
 - Ίσο με το άθροισμα των απόλυτων τιμών των βαρών
- Ομαλοποιημένη αντικειμενική συνάρτηση
 - $\tilde{J}(\mathbf{w}; \mathbf{X}, \mathbf{y}) = J(\mathbf{w}; \mathbf{X}, \mathbf{y}) + \alpha\|\mathbf{w}\|_1$
- Υπολογισμός κλίσης: $\nabla_{\mathbf{w}} \tilde{J}(\mathbf{w}; \mathbf{X}, \mathbf{y}) = \nabla_{\mathbf{w}} J(\mathbf{w}; \mathbf{X}, \mathbf{y}) + \alpha \text{sgn}(\mathbf{w})$
- **Διαφορά** ως προς ομαλοποίηση L^2
 - Η συνεισφορά της ομαλοποίησης στην κλίση **δεν** είναι πλέον **ανάλογη** του εύρους w_i της κάθε παραμέτρου, αλλά του προσήμου της
 - Για να προχωρήσουμε, κάνουμε την **επιπλέον παραδοχή** ότι ο εσσιανός πίνακας είναι **διαγώνιος** με $H_{i,i} > 0$
 - Έχουν αφαιρεθεί οι συσχετίσεις μεταξύ των χαρακτηριστικών της εισόδου λ.χ. με την προεπεξεργασία μέσω PCA

Ομαλοποίηση L^1 : Χαρακτηριστικά

- Τετραγωνική προσέγγιση \hat{J} της J
 - $\hat{J}(\mathbf{w}) = J(\mathbf{w}^*) + \sum_i \left[\frac{1}{2} H_{i,i} (\mathbf{w} - \mathbf{w}^*)^2 + \alpha |w_i| \right]$
- Τοπικό ελάχιστο: $\tilde{w}_i = \text{sgn}(w_i^*) \max \left(|w_i^*| - \frac{\alpha}{H_{i,i}}, 0 \right)$
 - Όταν $|w_i^*| \leq \frac{\alpha}{H_{i,i}}$, τότε το \tilde{w}_i γίνεται 0
 - Όταν $|w_i^*| > \frac{\alpha}{H_{i,i}}$, τότε το \tilde{w}_i «σύρεται» προς το 0 κατά έναν όρο $\frac{\alpha}{H_{i,i}}$
- Η ομαλοποίηση L^1 οδηγεί σε πιο **αραιές** (*sparse*) αναπαραστάσεις σε σύγκριση με την L^2
 - Υπό την έννοια ότι **περισσότερες παράμετροι** έχουν **μηδενικές τιμές**
 - Χρησιμοποιείται ως μηχανισμός **επιλογής χαρακτηριστικών** (*feature selection*)

Χαρακτηριστικά L^1 και L^2 ομαλοποίησης

- Είναι ισοδύναμες με την **εκτίμηση της μέγιστης εκ των υστέρων πιθανότητας** (MAP) της *Μπεϋζιανής συμπερασματολογίας*
- Ομαλοποίηση L^1
 - Εκ των προτέρων πιθανότητα ακολουθεί μια **ισοτροπική κατανομή Laplace**
- Ομαλοποίηση L^2
 - Εκ των προτέρων πιθανότητα ακολουθεί μια **κανονική κατανομή**

Αραιές Αναπαραστάσεις

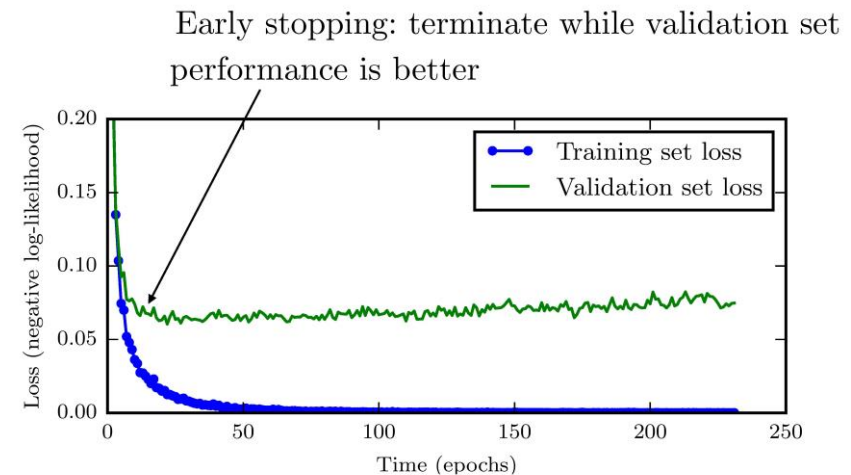
Αραιές Αναπαραστάσεις

- Sparse Representations
- Τοποθέτηση **όρου ποινής** στις **συναρτήσεις ενεργοποίησης** των νευρώνων αντί για τα βάρη
- Οι ενεργοποιήσεις τους γίνονται **αραιές**
 - Αντικατάσταση των εισόδων x των νευρώνων από αραιό διάνυσμα \mathbf{h}
- Ομαλοποιημένη αντικειμενική συνάρτηση:
 - $\tilde{J}(\theta; \mathbf{X}, \mathbf{y}) = J(\theta; \mathbf{X}, \mathbf{y}) + \alpha \Omega(\mathbf{h})$
- Επιλογή συνάρτησης ομαλοποίησης
 - Παρόμοια με ομαλοποίηση L^1 : $\Omega(\mathbf{h}) = \|\mathbf{h}\|_1 = \sum_i |h_i|$
 - Απόκλιση KL
 - Student-t priors
 - ...

Πρόωρος Τερματισμός

Πρόωρος Τερματισμός

- Early stopping
- **Αποθήκευση** των παραμέτρων του μοντέλου σε **κάθε βήμα**
- **Επιστροφή** αυτών που εμφανίζουν το **χαμηλότερο σφάλμα επαλήθευσης** στο τέλος της εκπαίδευσης



Πρόωρος Τερματισμός: Χαρακτηριστικά

- Από τις πιο **διαδεδομένες μορφές** ομαλοποίησης στη βαθιά μηχανική μάθηση
 - Λόγω της απλότητας και της αποτελεσματικότητάς της
- Τα **βήματα εκπαίδευσης** γίνονται μια ακόμα **υπερπαράμετρος** της διαδικασίας μάθησης
- **Πλεονεκτήματα**
 - **Δεν μεταβάλλει** καθόλου τη **διαδικασία μάθησης** (αντικειμενική συνάρτηση, σύνολο παραμέτρων κλπ)
 - Μπορεί να χρησιμοποιηθεί **σε συνδυασμό** με άλλες τεχνικές ομαλοποίησης
- **Μειονεκτήματα**
 - Απαιτεί την **αποθήκευση** των **παραμέτρων** του μοντέλου σε **κάθε βήμα** εκπαίδευσης
 - Απαιτεί τη **χρήση** ενός **μέρους** των δεδομένων εκπαίδευσης ως **δεδομένα επαλήθευσης** (*validation data*)

Ομαλοποίηση μέσω Πρόωρου Τερματισμού (1/2)

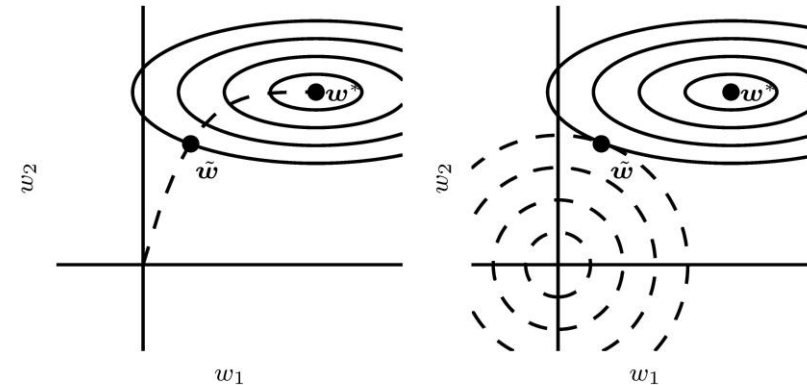
- Ο πρόωρος τερματισμός **περιορίζει** την αναζήτηση των βέλτιστων παραμέτρων σε ένα **χώρο γύρω** από τις **αρχικές θ_0**
 - Στην περίπτωση που έχουμε τ βήματα εκπαίδευσης και ρυθμό μάθησης ϵ , το γινόμενο $\epsilon\tau$ αποτελεί μια μετρική της **πραγματικής χωρητικότητας** του μοντέλου.
- Στην περίπτωση γραμμικού μοντέλου με τετραγωνική συνάρτηση σφάλματος και **απλή κατάβαση κλίσης**, ο πρόωρος τερματισμός είναι **ισοδύναμος** με την ομαλοποίηση L^2
 - Αντικειμενική συνάρτηση: $\hat{J}(\mathbf{w}) = J(\mathbf{w}^*) + \frac{1}{2}(\mathbf{w} - \mathbf{w}^*)^T \mathbf{H}(\mathbf{w} - \mathbf{w}^*)$
 - Κλίση: $\nabla_{\mathbf{w}} \hat{J}(\mathbf{w}; \mathbf{X}, \mathbf{y}) = \mathbf{H}(\mathbf{w} - \mathbf{w}^*)$

Ομαλοποίηση μέσω Πρόωρου Τερματισμού (2/2)

- Έστω ότι στην αρχή $w^{(0)} = 0$. Την χρονική στιγμή τ
 - $w^{(\tau)} \leftarrow w^{(\tau-1)} - \epsilon \nabla_w \hat{J}(w) \Rightarrow w^{(\tau)} - w^* \leftarrow (\mathbf{I} - \epsilon \mathbf{H})(w^{(\tau-1)} - w^*)$
- Χρησιμοποιώντας αποσύνθεση ιδιοτιμών για τον εσσιανό πίνακα $\mathbf{H} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$
 - $\mathbf{Q}^T(w^{(\tau)} - w^*) \leftarrow (\mathbf{I} - \epsilon \mathbf{\Lambda})\mathbf{Q}^T(w^{(\tau-1)} - w^*)$
- Υποθέτοντας ότι το ϵ επιλέγεται τόσο μικρό ώστε $|1 - \epsilon \lambda_i| < 1$
 - $\mathbf{Q}^T w^{(\tau)} = [\mathbf{I} - (\mathbf{I} - \epsilon \mathbf{\Lambda})^T]\mathbf{Q}^T w^*$
 - Παραπλήσια μορφή με ομαλοποίηση L^2 : $\mathbf{Q}^T \tilde{w} = [\mathbf{I} - \alpha(\mathbf{\Lambda} + \alpha \mathbf{I})^{-1}]\mathbf{Q}^T w^*$
 - Η ισοδυναμία ισχύει όταν $\mathbf{I} - (\mathbf{I} - \epsilon \mathbf{\Lambda})^T = \mathbf{I} - \alpha(\mathbf{\Lambda} + \alpha \mathbf{I})^{-1}$
 - Ειδικότερα, αν $\epsilon \lambda_i \ll 1$ και $\frac{\lambda_i}{\alpha} \ll 1$, τότε $\tau \approx \frac{1}{\epsilon \alpha}$ και $\alpha \approx \frac{1}{\tau \epsilon}$
- Συνεπώς ο **αριθμός των βημάτων** τ παίζει ένα **ρόλο ανάστροφο** ως προς τον βαθμό α της ομαλοποίησης L^2 και το **ανάστροφο του γινομένου** $\tau \epsilon$ παίζει το ρόλο του **βαθμού της ομαλοποίησης** L^2

Σύγκριση Πρόωρου Τερματισμού και Ομαλοποίησης L^2

- Ομαλοποίηση L^2
 - Παράμετροι που επηρεάζουν πολύ την αντικειμενική συνάρτηση **ομαλοποιούνται λιγότερο**
 - **Πολλά πειράματα** για διαφορετικές τιμές της υπερπαραμέτρου
- Πρόωρος Τερματισμός
 - Παράμετροι που επηρεάζουν πολύ την αντικειμενική συνάρτηση **ομαλοποιούνται γρηγορότερα**
 - **Εύκολος** και **γρήγορος** εντοπισμός του ποσοστού ομαλοποίησης που χρειάζεται, ακολουθώντας την **τροχιά** της κλίσης

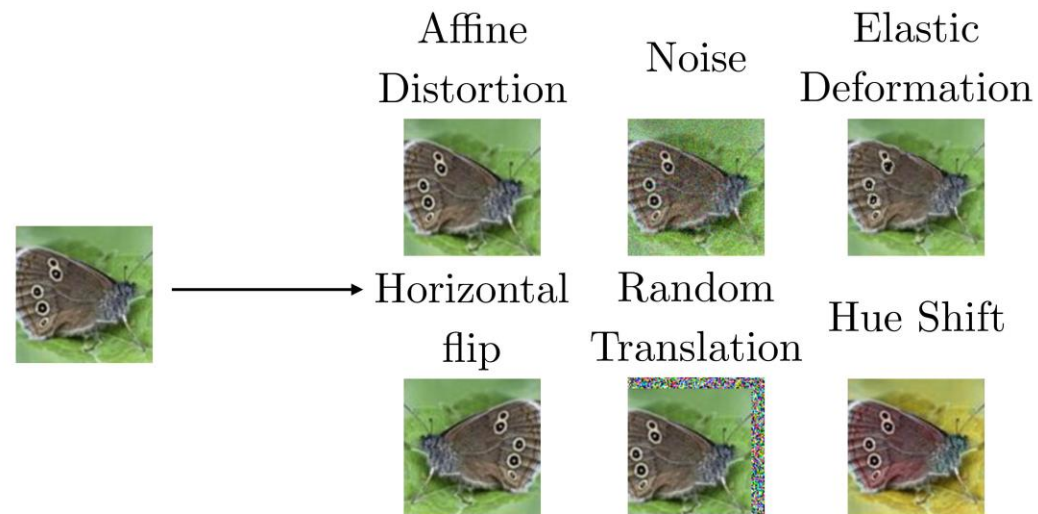


Επαύξηση δεδομένων

Επαύξηση δεδομένων

- Ένα μοντέλο βαθιάς μηχανικής μάθησης μπορεί να **γενικεύσει καλύτερα** αν του δοθούν **περισσότερα δεδομένα**
- Σε ορισμένες περιπτώσεις μπορούμε να **δημιουργήσουμε «τεχνητά» δεδομένα** και να τα προσθέσουμε στο σύνολο εκπαίδευσης
 - λ.χ. σε ένα πρόβλημα ταξινόμησης μπορούμε να παράξουμε **επιπρόσθετα ζεύγη** $(X; y)$ **μετασχηματίζοντας** την είσοδο X
 - Μπορούμε επίσης να προσθέσουμε **θόρυβο** στα δεδομένα εισόδου
- Παράδειγμα: Αναγνώριση αντικειμένων σε εικόνες
 - **Μετακινώντας** τις εικόνες κατά **ορισμένα pixel** ανά κατεύθυνση έχειδειχθεί ότι **αυξάνει** κατά πολύ τη **δυνατότητα γενίκευσης** των αλγορίθμων μηχανικής μάθησης
 - Χρειάζεται **προσοχή** έτσι ώστε να **μην αλλοιωθεί** η φύση του προβλήματος
 - π.χ. σε ένα πρόβλημα αναγνώρισης χαρακτήρων να μην κάνουμε το 'b' 'd'

Επαύξηση δεδομένων: Παράδειγμα



Bagging

Bagging (1/2)

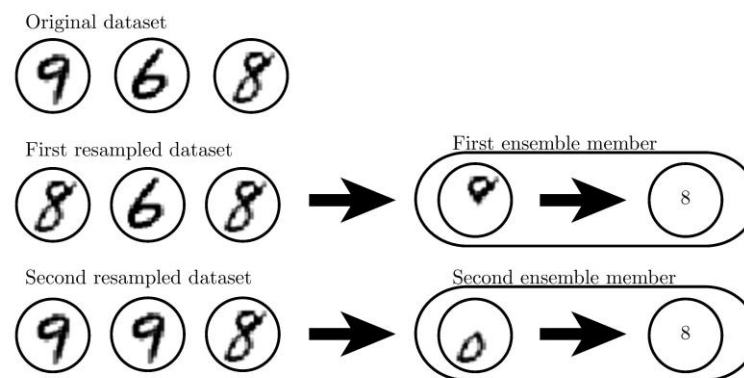
- Συντομογραφία του *bootstrap aggregating*
- **Μείωση** σφάλματος γενίκευσης μέσω συνδυασμού πολλαπλών μοντέλων
 - Το κάθε ένα εκπαιδεύεται **χωριστά**
 - Τα μοντέλα «**ψηφίζουν**» για την έξοδο (*model averaging*)
 - Επίσης αυτές οι στρατηγικές είναι γνωστές και ως **ensemble methods**
- Βασική αρχή: Τα **διαφορετικά μοντέλα** **δε** θα έχουν όλα **το ίδιο σφάλμα** στο σύνολο ελέγχου
 - k διαφορετικά μοντέλα,
 - e_i το σφάλμα του i -στού μοντέλου
 - Προέρχεται από **κανονική κατανομή πολλαπλών μεταβλητών** με **διασπορά** $\mathbb{E}[e_i^2] = v$ και **συνδιασπορά** $\mathbb{E}[e_i e_j] = c$

Bagging (2/2)

- Σφάλμα πρόβλεψης συνολικού μοντέλου: $e = \frac{1}{k} \sum_i e_i$
- Αναμενόμενη τιμή τετραγώνου του σφάλματος
 - $\mathbb{E}[e^2] = \frac{1}{k} V + \frac{k-1}{k} C$
- Αν τα σφάλματα είναι **πλήρως συσχετισμένα** μεταξύ τους
 - $c = v$ και $\mathbb{E}[e^2] = v$
 - Δεν έχουμε **κανένα κέρδος**
 - Το συνολικό μοντέλο είναι τόσο καλό όσο και τα μέλη του
- Τα σφάλματα είναι **πλήρως ασυσχέτιστα** μεταξύ τους
 - $c = 0$ και $\mathbb{E}[e^2] = \frac{1}{k} v$
 - Το μέσο σφάλμα **μειώνεται** κατά ένα παράγοντα $\frac{1}{k}$

Επεκτάσεις

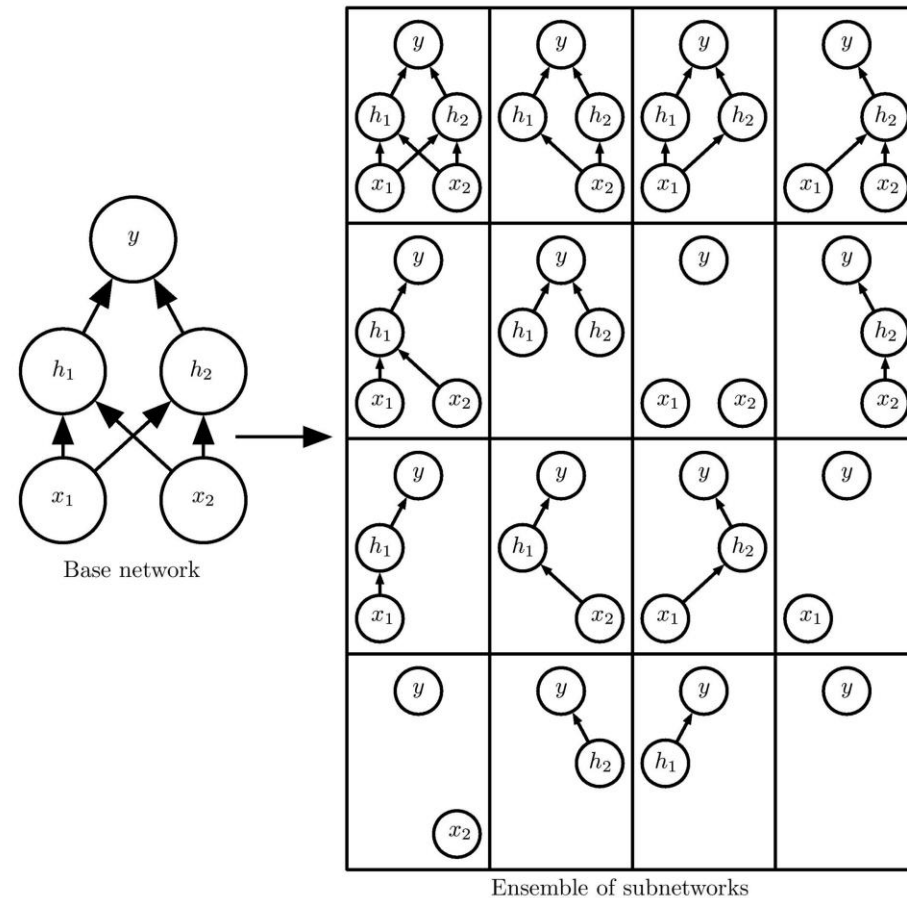
- Γενικότερη μεθοδολογία που **επιτρέπει** την **επαναχρησιμοποίηση** μοντέλων, αλγορίθμων εκπαίδευσης και αντικειμενικών συναρτήσεων
- Κατασκευή k **διαφορετικών** συνόλων δεδομένων
 - Μέσω **δειγματοληψίας με επανατοποθέτηση**
 - Με αρκετά μεγάλη πιθανότητα, κάθε σύνολο δεδομένων έχει **δείγματα** με **πολλαπλότητα μεγαλύτερη** του 1, ενώ άλλα δεν υπάρχουν **καθόλου**
 - Κατ' αυτόν τον τρόπο, τα μέλη του ensemble θα εμφανίζουν σφάλματα που είναι **μερικώς ανεξάρτητα** μεταξύ τους



Dropout

Dropout

- Επέκταση του bagging σε πολύ μεγάλα δίκτυα
- Προϋποθέτει την ύπαρξη ενός δικτύου βάσης
- Σχηματίζει και εκπαιδεύει όλα τα πιθανά υποδίκτυα που προκύπτουν από το δίκτυο βάσης
 - Συνήθως επιλέγονται οι κόμβοι εισόδου με πιθανότητα 0.8 και οι κρυφοί κόμβοι με πιθανότητα 0.5



Ομαλοποίηση μέσω Dropout

- **Διάνυσμα μάσκας** μ ορίζει ποιους κόμβους θα συμπεριληφθούν
- Αντικειμενική συνάρτηση $J(\theta, \mu)$
- Εκπαίδευση
 - Ελαχιστοποίηση $\mathbb{E}_{\mu} J(\theta, \mu)$
 - Περιέχει εκθετικά πολλούς όρους αλλά μπορεί να προσεγγιστεί μέσω δειγματοληψίας του μ
- Πολλές **ομοιότητες** αλλά και ορισμένες **διαφορές** με bagging
 - **Διαμοιρασμός παραμέτρων** μεταξύ των υποδικτύων και όχι ανεξάρτητα μοντέλα
 - Ένα **μικρό υποσύνολο** όλων των πιθανών μοντέλων εκπαιδεύεται κάθε φορά (στο bagging όλα)
- Παραγωγή προβλέψεων
 - Bagging: Αριθμητικός μέσος των κατανομών εξόδου του κάθε μοντέλου: $\frac{1}{k} \sum_{i=1}^k p^{(i)}(y|x)$
 - Dropout: Αριθμητικός μέσος των κατανομών εξόδου του κάθε υπο-δικτύου που ορίζεται από μάσκα μ : $\sum_{\mu} p(\mu) p(y|x, \mu)$

Βιβλιογραφία

Βιβλιογραφία (1/2)

- Πηγή: *Ian Goodfellow, Yoshua Bengio, Aaron Courville* “Deep Learning” – MIT Press (<https://www.deeplearningbook.org/>)
- Εισαγωγικές Έννοιες
 - Χωρητικότητα, Υπερπροσαρμογή, υποπροσαρμογή (§5.2, §5.2.1)
 - Ισορροπία Μεροληψίας και Διακύμανσης (§5.4.4)
 - Προσαρμογή (§5.2.2, §7)
- Ομαλοποίηση μέσω της προσθήκης όρων ποινής (§7.1)
 - Ομαλοποίηση L^2 (§7.1.1)
 - Ομαλοποίηση L^1 (§7.1.2)
 - Αραιές Αναπαραστάσεις (§7.10)

Βιβλιογραφία (2/2)

- Πρόωρος Τερματισμός (§7.8)
- Επαύξηση Δεδομένων (§7.4)
- Μέθοδοι Ensemble
 - Bagging (§7.11)
 - Dropout (§7.12)