L03 – Παλινδρόμηση (Regression)

Σύνολο εκπαίδευσης (Training Set): $D = \{(x_i, y_i)\}_{i=1}^{N} \sim p$
Validation Set: $V = \{(x_i, y_i)\}_{i=1}^{V} \sim p$
Σύνολο Ελέγχου (Test Set): $J = \{(x_i, y_i)\}_{i=1}^{T} \sim p$
Σύνολο Δεδομένων (Dataset): $D \cup V \cup J \sim p$

Γραμμική Παλινδρόμηση (Linear Regression)

Υπόθεση για τα δεδομένα: Γραμμική σχέση ανάμεσα στις μεταβλητές x και y.
$\hat{y}_i = f(x_i; w = [w_0, w_1]^T) = w_0 + w_1 x_i, x_i \in \mathbb{R}$
$\hat{y}_i = f(x_i; w^T) = w^T x_i, w, x_i \in \mathbb{R}^{D+1}, x_i = [1, x_{i1}, \dots, x_{iD}]^T$

Εκπαίδευση: Προσδιορισμός των παραμέτρων του μοντέλου (w), έτσι ώστε η εκτίμηση της συνάρτησης $f(\cdot)$ να είναι κοντά στα δεδομένα εκπαίδευσης.

Υπόλοιπο (Residual): $e_i = y_i - \hat{y}_i = y_i - f(x_i; w)$
Residual Sum of Squares (RSS): $RSS = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} (y_i - f(x_i; w))^2$
Το RSS αποτελεί τη συνάρτηση απώλειας (loss function) της μεθόδου εκτίμησης παραμέτρων που είναι γνωστή ως ελάχιστα τετράγωνα (least squares).

$RSS = \|y - Xw\|_2^2$, όπου $y = [y_1, \dots, y_N]^T, X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1D} \\ \vdots & & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{ND} \end{bmatrix}, w = \begin{bmatrix} w_0 \\ \vdots \\ w_D \end{bmatrix}$

Το βέλτιστο διάνυσμα παραμέτρων αποτελεί λύση του προβλήματος βελτιστοποίησης:
$\min_w l(w) = \min_w \|y - Xw\|_2^2$
Λύση
$l(w) = \|y - Xw\|_2^2 = (y - Xw)^T (y - Xw) = (y^T - w^T X^T)(y - Xw) = y^t y - y^T Xw - w^T X^T y + w^T X^T Xw$
$\frac{\partial l(w)}{\partial w} = 0 - (y^T X)^T - X^T y + X^T Xw + (w^T X^T X)^T = -2 \cdot X^T y + 2 \cdot X^T Xw$
$\frac{\partial l(w)}{\partial w} = 0 \Rightarrow -2 \cdot X^T y + 2 \cdot X^T Xw = 0 \Leftrightarrow X^T Xw = X^T y \Rightarrow w^* \begin{cases} (X^T X)^{-1} X^T y & , if\ X^T X\ is\ invertible \\ X^{\dagger} y & , otherwise \end{cases}$

Πολυωνυμική Παλινδρόμηση

Υπόθεση: Πολυωνυμική σχέση ανάμεσα στα δεδομένα X και Y.
$\hat{y}_i = f(x_i; w = [w_0, \dots, w_M]^T) = \sum_{j=0}^{M} w_j x_i^j, x_i \in \mathbb{R}$
Ο βαθμός του πολυωνύμου (M) αποτελεί υπερ-παράμετρο (hyper-parameter) του μοντέλου.
Συνάρτηση απώλειας (loss function): RSS

$RSS = \|y - Xw\|_2^2$, όπου $y = [y_1, \dots, y_N]^T, X = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^M \\ \vdots & & \ddots & & \vdots \\ 1 & x_N & x_N^2 & \dots & x_N^M \end{bmatrix}, w = \begin{bmatrix} w_0 \\ \vdots \\ w_D \end{bmatrix}$

Η λύση σε κλειστή μορφή είναι ίδια με πριν για προφανείς λόγους.

Ο προσδιορισμός του βαθμού του πολυωνύμου (και γενικά των υπερπαραμέτρων στα στατιστικά μοντέλα) απαιτεί μεθοδολογίες επιλογής μοντέλου (model selection) ώστε να μεγιστοποιήσουμε την γενίκευση.

Η χωρητικότητα ενός μοντέλου (model capacity) μηχανικής μάθησης, αφορά την ικανότητα του να μάθει ένα ευρύ σύνολο συναρτήσεων μέσω των δεδομένων εκπαίδευσης. Ένας τρόπος να την ελέγξουμε είναι με το να τροποποιήσουμε τον χώρο υποθέσεων.
Μοντέλα μικρής χωρητικότητας δεν μπορούν να περιγράψουν σύνθετα δεδομένα ικανοποιητικά (under-fitting)
Μοντέλα μεγάλης χωρητικότητας μπορεί να απομνημονεύσουν ιδιότητες και χαρακτηριστικά του συνόλου εκπαίδευσης τα οποία δεν παρουσιάζονται στο σύνολο ελέγχου (over-fitting).

Ο χώρος υποθέσεων μοντέλου πολυωνυμικής παλινδρόμησης: $H_M = \{y = \sum_{j=0}^{M} w_j x_j^j, \forall w \in \mathbb{R}^{M+1}\}$ και για γραμμική παλινδρόμηση: $H = \{y = w_0 + w_1 x, \forall w \in \mathbb{R}^2\}$

L04 – Παλινδρόμηση (συνέχεια)

Γραμμική παλινδρόμηση με διανυσματικά δεδομένα εισόδου: $y = f(x; w) = w_0 + w_1 x_1 + \cdots + w_D x_D$
Το παραπάνω μοντέλο αποτελεί γραμμική συνάρτηση των παραμέτρων w και των μεταβλητών εισόδου.
Εξετάζοντας γραμμικούς συνδυασμούς μη-γραμμικών συναρτήσεων των μεταβλητών εισόδου, επεκτείνουμε την κλάση των μοντέλων ή ισοδύναμα το χώρο υποθέσεων.

Γραμμικό μοντέλο συναρτήσεων βάσης

$y = f(x; w) = w_0 + \sum_{j=1}^{J-1} w_j \phi_j(x)$
Το παραπάνω μοντέλο αποτελεί γραμμική συνάρτηση των παραμέτρων w, αλλά **όχι** των μεταβλητών εισόδου, εφόσον οι συναρτήσεις βάσεις είναι μη-γραμμικές.

Παραδείγματα συναρτήσεων βάσης: $\phi_j(x) = x^j, \phi(x) = \exp\left(-\frac{(x-\mu)^2}{2\cdot\sigma^2}\right)$
Οι συναρτήσεις βάσης $\varphi(\cdot)$ αποτελούν κάποιας μορφής σταθερής προεπεξεργασίας ή εξαγωγής χαρακτηριστικών από τα αρχικά δεδομένα εισόδου.

Ορίζουμε:
$y = [y_1, \dots, y_N]^T \in \mathbb{R}^N, X = [x_1, \dots, x_N] \in \mathbb{R}^{D\times N}$,
$\varphi_i \doteq \varphi(x_i) = \left[\varphi_0(x_i), \varphi_1(x_i), \dots, \varphi_{j-1}(x_i)\right] \in \mathbb{R}^{1\times J}, \Phi = [\varphi_1, .., \varphi_N]^T \in \mathbb{R}^{N\times J}$
$y = \Phi w$
$l(w) = \|y - \Phi w\|_2^2$ και $w^* \begin{cases} (\Phi^T\Phi)^{-1}\Phi^T y, if\ \Phi^T\Phi\ is\ invertible \\ \Phi^\dagger y \quad\quad , otherwise \end{cases}$
Το w^* ενδέχεται να μην είναι μοναδικό.

Στατιστικά μοντέλα, υποθέσεις και κατανομές πιθανότητας

Υπόθεση (assumption): Οι ερμηνείες των παρατηρήσεων (δεδομένων) σχετίζονται με τις υποθέσεις.
Οι καλές υποθέσεις δομούν τον κόσμο με χρήσιμο τρόπο.
Οι λανθασμένες υποθέσεις μπορεί να είναι πολύ "τρομακτικές".

Αβεβαιότητα (uncertainty): Η αβεβαιότητα είναι μία "πραγματοποίηση" μίας υπόθεσης και αποτελεί κεντρική έννοια στη μοντελοποίηση δεδομένων.
Οφείλεται (κυρίως):

1. Στο θόρυβο που υπεισέρχεται στα δεδομένα
2. Στις υποθέσεις που κάνουμε για τα δεδομένα

Οι πιθανότητες αποτελούν ποσοτικοποίηση της αβεβαιότητας

Αναμενόμενη (Μέση) Τιμή (Expected Value): $E[f(x)] = \begin{cases} \sum_x p(x)f(x), & Διακριτή\ Κατανομή \\ \int p(x)f(x)dx, & Συνεχής\ Κατανομή \end{cases}$
Διακύμανση (Variance): $Var[f(x)] = \sigma^2 = E[(f(x) - E[f(x)])^2] = E[f^2(x)] - E^2[f(x)]$
Τυπική απόκλιση (Standard deviation): $\sigma = \sigma(f(x)) = \sqrt{Var[f(x)]}$

Gaussian/Normal Distribution: $p(x; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \cdot exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

Βασική υπόθεση για την παραγωγή δεδομένων: Independently and identically distributed (i.i.d.) data
Τα δεδομένα προέρχονται από την ίδια (πανομοιότυπη), άγνωστη αλλά σταθερή, κατανομή $p(x; y)$
Επιπλέον όλα τα δεδομένα παράγονται ανεξάρτητα μεταξύ τους.
Η ανεξαρτησία κατά την διαδικασία παραγωγής των δεδομένων συνεπάγεται ότι η κατανομή πιθανότητας του συνόλου των διαθέσιμων δεδομένων μπορεί να εκφραστεί ως γινόμενο των πιθανοτήτων του καθενός δείγματος, δηλαδή: $p(Y|X; w) = \prod_{i=1}^N p(y_i|x_i; w)$, όπου $p(y_i|x_i; w)$ κάποια συγκεκριμένη συνάρτηση πυκνότητας πιθανότητας.

Maximum Likelihood Estimation (MLE)

Μια μέθοδος για να εκτιμήσουμε τις άγνωστες παραμέτρους είναι να μεγιστοποιήσουμε την πιθανοφάνεια (likelihood). Η μέθοδος αυτή είναι γνωστή ως Maximum Likelihood Estimation (MLE)

Παράδειγμα

$y_i = f(x_i; w) + \epsilon_i = \varphi(x_i)w + \epsilon_i$

Υπόθεση για τον θόρυβο: $\epsilon \sim N(0, \sigma^2), iid$

Ξέρουμε ότι $p(\epsilon_i) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \cdot exp\left(-\frac{(\epsilon_i-\mu)^2}{2\sigma^2}\right) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \cdot exp\left(-\frac{\epsilon_i^2}{2\sigma^2}\right)$ και $y_i = \varphi(x_i)w + \epsilon_i \Rightarrow \epsilon_i = y_i - \varphi(x_i)w$

Συνεπώς έχουμε ότι $p(y_i|x_i, w) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \cdot exp\left(-\frac{(y_i-\varphi(x_i)w)^2}{2\sigma^2}\right)$

$L(w) = \prod_{i=1}^{N} p(y_i|x_w, w)$ και

$l(w) = \ln L(w) = \ln \prod_{i=1}^{N} p(y_i|x_w, w) = \sum_{i=1}^{N} \ln p(y_i|x_w, w) = \sum_{i=1}^{N} \ln \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} + \ln \exp\left(-\frac{(y_i-\varphi(x_i)w)^2}{2\sigma^2}\right) =$

$= \sum_{i=1}^{N} -\frac{1}{2}\ln 2\pi\sigma^2 - \frac{(y_i-\varphi(x_i)w)^2}{2\sigma^2} = -\frac{N}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \varphi(x_i)w)^2 = -\frac{N}{2}\ln 2\pi\sigma^2 - \|y - \Phi w\|_2^2$

$w^* = \max_{w} l(w) = \min_{w} \|y - \Phi w\|_2^2 \Rightarrow w^* = \begin{cases} (\Phi^T\Phi)^{-1}\Phi^T y, if\ \Phi^T\Phi\ is\ invertible \\ \Phi^\dagger y \quad\quad , otherwise \end{cases}$

L05 – Ταξινόμηση (Classification)

Γραμμικός ταξινομητής: 2 κλάσεις

Υπόθεση για τα δεδομένα: Τα δεδομένα διαχωρίζονται γραμμικά σε 2 κλάσεις.
Αν $y(x; w) > 0$ τότε το x ανήκει στην κατηγορία 1, αλλιώς ανήκει στη κατηγορία 2.

Γραμμικός ταξινομητής: K κλάσεις

Υπόθεση για τα δεδομένα: Τα δεδομένα διαχωρίζονται γραμμικά σε K κλάσεις.
Στατιστικό μοντέλο: Κάθε κλάση περιγράφεται από το δικό της γραμμικό μοντέλο $y_k(x; [w_0, w]_K^T) = w_K^T x + w_0$
One-hot vector: $y_i = [0 \dots 1 \dots 0]^T \in [0,1]^K$

$Y = \bar{W}\bar{X}, Y = [y_1, \dots, y_N] \in [0,1]^{K\times N},\ \bar{X} = [\bar{x}_1, \dots, \bar{x}_N] \in \mathbb{R}^{(D+1)\times N},\ \bar{W} = [\bar{w}_1, \dots, \bar{w}_N] \in \mathbb{R}^{K\times(D+1)}$
$RSS = \sum_{i=1}^{K}\sum_{j=1}^{N}\left(y_{ij} - \widehat{y_{ij}}\right)^2 = \sum_{i=1}^{K}\sum_{j=1}^{N}\left(y_{ij} - (\bar{w}\bar{x})_{ij}\right)^2 = \|y - \bar{W}\bar{X}\|_F^2$ και $\bar{W}^* = \min_{\mathrm{w}} l(w) = Y\bar{X}^{\dagger}$
$\mathrm{l}(W) = \|Y - WX\|_F^2 = tr\left((Y - WX)(Y - WX)^H\right) = tr\left((Y - WX)(Y - WX)^T\right) = tr\left((Y - WX)(Y^T - X^TW^T)\right) =$
$= tr(YY^T - YX^TW^T - WXY^T + WXX^TW^T) = tr(YY^T) - tr(YX^TW^T) - tr(WXY^T) + tr(WXX^TW^T)$
$\frac{\partial l(W)}{\partial W} = 0 - YX^T - (XY^T)^T + (XX^TW^T)^T + WXX^T = 2\cdot(WXX^T - YX^T)$
$\frac{\partial l(W)}{\partial W} = 0 \Rightarrow 2\cdot(WXX^T - YX^T) = 0 \Leftrightarrow WXX^T = YX^T \Leftrightarrow W = \begin{cases} YX^T(XX^T)^{-1}, if\ XX^T\ is\ invertible \\ YX^{\dagger} \quad , otherwise \end{cases}$

Logistic Regression

$sigmoid(z) = f(z) = \frac{1}{1+e^{-z}}$

Στατιστικό μοντέλο: $y(x; \bar{w}) = f(\bar{w}^T x) = \frac{1}{1+e^{-\bar{w}^T x}} \in [0,1]$

Ταξινόμηση: $y = \begin{cases} 1, f(\bar{w}^T x) \geq 0.5 \\ 0, \quad \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases}$

Συνάρτηση κόστους cross-entropy: $l(w) = -\frac{1}{N}\cdot\sum_{i=1}^{N} \mathrm{y_i}\cdot\log f(x_i; w) + (1 - y_i)\cdot\log(1 - f(x_i; w)$

Εκπαιδεύουμε έναν ταξινομητή για κάθε κλάση και δεδομένου ενός νέου δείγματος x εκχωρούμε σε αυτό την κλάση που αντιστοιχεί στον ταξινομητή με τη μεγαλύτερη τιμή.

Ταξινομητής πλησιέστερου γείτονα (K-NN)

$S = \{(x_i, y_i)\}_{i=1}^{N},\ x_i \in \mathbb{R}^D, y_i \in \mathbb{R}$
$\hat{y} = y_j$, where $j = \arg\min_{i=1,\dots,N} \|x - x_i\|$

Η επιφάνεια απόφασης: $f(x) = 0$ εξομαλύνεται όσο αυξάνει το K
Μεταβάλλοντας την τιμή του K, το αποτέλεσμα αλλάζει πολύ.
Η τιμή του K αποτελεί υπερ-παράμετρο του αλγορίθμου.

L06 – Αξιολόγηση επίδοσης και επιλογή μοντέλου

Αξιολόγηση μοντέλων παλινδρόμησης

Η αξιολόγηση των μοντέλων παλινδρόμησης εστιάζει κυρίως σε δύο πτυχές:

1. Πόσο καλά το μοντέλο που έχει εκτιμηθεί μπορεί να εξηγήσει τη διακύμανση της εξαρτημένης μεταβλητής Y στο σύνολο δεδομένων εκπαίδευσης, δηλαδή πόσο καλά περιγράφει το μοντέλο τα δεδομένα.
2. Πόσο κοντά στην πραγματική τιμή βρίσκεται η τιμή που προβλέπει το μοντέλο παλινδρόμησης

Η αξιολόγηση του πόσο καλά ένα μοντέλο γραμμικής παλινδρόμησης εξηγεί τη διακύμανση της τιμής της μεταβλητής στόχου γίνεται με υπολογισμό του συντελεστή προσδιορισμού (coefficient of determination) R^2, λαμβάνει τιμές από το 0 έως και το 1 και εκφράζει το ποσοστό της διακύμανσης που κατορθώνει να εξηγήσει το μοντέλο παλινδρόμησης: $R^2 = 1 - \frac{RSS}{TSS}$, $RSS = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$, $TSS = \sum_{i=1}^{M}(y_i - \bar{y})^2$

Όσο πιο υψηλή η τιμή του συντελεστή R^2 τόσο καλύτερα επιτυγχάνει το μοντέλο να εξηγήσει τη διακύμανση της ανεξάρτητης μεταβλητής και επομένως να εξηγήσει τα δεδομένα. Μπορεί να χρησιμοποιηθεί για έλεγχο υποθέσεων.

Η αξιολόγηση ενός μοντέλου παλινδρόμησης με στόχο τη πρόβλεψη γίνεται με τη χρήση μετρικών οι οποίες μετρούν το σφάλμα που κάνει το μοντέλο παλινδρόμησης στην πρόβλεψη της τιμής της εξαρτημένης μεταβλητής και συνεπώς οι μετρικές αυτές καλούνται και μετρικές σφάλματος.

Μετρικές εκτίμησης σφάλματος πρόβλεψης

Mean Absolute Error: $MAE = \frac{1}{N} \cdot \sum_{i=1}^{N}|y_i - \hat{y}_i|$

Το Μέσο Απόλυτο Σφάλμα δίνει την ίδια βαρύτητα σε όλα τα υπόλοιπα (residuals) που προκύπτουν και δεν διακρίνει εάν η προβλεπόμενη τιμή υπερτιμά ή όχι την πραγματική τιμή.

Mean Squared Error: $MSE = \frac{1}{N} \cdot \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$

Το Μέσο Τετραγωνισμένο Σφάλμα δίνει διαφορετική βαρύτητα στο μέγεθος των υπολοίπων (residual) με τα μεγαλύτερα υπόλοιπα να συνεισφέρουν στο συνολικό σφάλμα παραπάνω εξαιτίας της ύψωσης στο τετράγωνο και προτιμάται εάν είναι επιθυμητό να “τιμωρηθούν” μεγάλες τιμές καταλοίπων και ακραίες προβλεπόμενες τιμές.

Root Mean Squared Error: $RMSE = \sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$

Το Μέσο Τετραγωνικό Σφάλμα από την άλλη έχει καλύτερη ερμηνευτική δύναμη καθότι το σφάλμα εκφράζεται σε μονάδες της ανεξάρτητης μεταβλητής.

Mean Absolute Percentage Error: $MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$

Το Μέσο Απόλυτο Εκατοστιαίο Σφάλμα μεροληπτεί υπέρ προβλεπόμενων τιμών που είναι συστηματικά μικρότερες από την πραγματική τιμή. Ειδικότερα, η μετρική αυτή θα είναι μικρότερη όπου οι προβλεπόμενες τιμές είναι μικρότερες από την πραγματική εάν συγκριθεί με άλλο μοντέλο το οποίο παρουσιάζει τιμές που προβλέπει τιμές μεγαλύτερες από τις πραγματικές κατά το ίδιο μέγεθος.

Αξιολόγηση μοντέλων ταξινόμησης

Υποθέτοντας ότι υπάρχουν 2 κλάσεις δεδομένων:

True Negatives (TN): οι περιπτώσεις όπου η πραγματική κλάση του σημείου δεδομένων είναι ίδια με την προβλεπόμενη και είναι A.
True Positives (TP): οι περιπτώσεις όπου η πραγματική κλάση των δεδομένων είναι διαφορετική από την προβλεπόμεν και είναι B.
False Positives (FP): οι περιπτώσεις όπου η πραγματική κλάση των δεδομένων είναι A και η προβλεπόμενη είναι B.
False Negatives (FN): οι περιπτώσεις όπου η πραγματική κλάση των δεδομένων είναι A και η προβλεπόμενη είναι B.

Μετρικές εκτίμησης εσφαλμένης ταξινόμησης

Η ακρίβεια ταξινόμησης (classification accuracy): $Accuracy = \frac{TN+TP}{TN+TP+FP+FN} = \frac{correct\ predictions}{total\ predictions}$

Η ακρίβεια δεν πρέπει χρησιμοποιείται ως μέτρο όταν πολλά δεδομένα ανήκουν σε μία κλάση (class imbalance problem).

Η ακρίβεια (precision): $\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$, είναι ένα καλό μέτρο για να προσδιοριστεί πότε το κόστος των FP είναι υψηλό. Πόσα από αυτά που προέβλεψα ως A, είναι όντως A.

Το recall: $\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$, πόσα από τα συνολικά A προέβλεψα ως A.

$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}$

Τα μέτρα αξιολόγησης ορίζονται και για K>2 κλάσεις με βάση το confusion matrix.

Επιλογή μοντέλου

Επιλογή μοντέλου: Εκτίμηση της επίδοσης διαφορετικών μοντέλων μάθησης ώστε να επιλέξουμε το πιο αποδοτικό.
Η ιδιότητα των αλγορίθμων μηχανικής μάθησης να είναι αποδοτικοί σε νέα δεδομένα εκτός συνόλου εκπαίδευσης, δηλαδή στο σύνολο ελέγχου, ονομάζεται γενίκευση (generalization).
Η i.i.d. υπόθεση επί της διαδικασίας παραγωγής των δεδομένων εκπαίδευσης και ελέγχου μας επιτρέπουν να μελετήσουμε τη σχέση μεταξύ σφάλματος εκπαίδευσης και ελέγχου.

Δεδομένης μιας συνάρτησης κόστους, η ποιότητα της εκτίμησης μπορεί να ποσοτικοποιηθεί μέσω του πραγματικού ρίσκου (αναμενόμενου σφάλματος ή σφάλματος ελέγχου ή σφάλματος γενίκευσης - true risk ή expected error ή test error ή generalisation error): $R(f) = E_{(x,y)\sim p}[(f(x) - y)^2]$
Στόχος της μηχανικής μάθησης με επίβλεψη είναι η επίλυση του προβλήματος βελτιστοποίησης: $f^* = \min_{f \in H} R(f)$

Επειδή δεν μπορούμε να υπολογίσουμε αναλυτικά το πραγματικό ρίσκο εφόσον η κατανομή από την οποία προέρχονται τα δεδομένα είναι άγνωστη, στη πράξη έχουμε πρόσβαση μόνο στο σφάλμα εκπαίδευσης (training error) ή εμπειρικό ρίσκο (empirical risk): $\hat{R}(f) = E_{(x,y)\sim D}[(f(x) - y)^2]$
Μαθαίνουμε την f^* ελαχιστοποιώντας το $\hat{R}(f)$ ως υποκατάστατο του $R(f)$: $\hat{f} = \min_{f \in H} \hat{R}(f)$

Ιδανικά επιθυμούμε $R(f^*) \approx \hat{R}(\hat{f})$

Ένας αλγορίθμους μηχανικής μάθησης έχει καλή επίδοση όταν:

1. Το σφάλμα εκπαίδευσης είναι μικρό.
2. Η απόσταση μεταξύ σφάλματος εκπαίδευσης και γενίκευσης είναι μικρή. Αυτή η απόσταση είναι γνωστή ως χάσμα γενίκευσης (generalization gap)

Το φαινόμενο underfitting παρατηρείται όταν ο αλγόριθμος μηχανικής μάθησης δεν μπορεί να επιτύχει ικανοποιητικά μικρό σφάλμα εκπαίδευσης.
Το φαινόμενο overfitting παρατηρείται όταν το χάσμα γενίκευσης είναι πολύ μεγάλο.
Στη πράξη ελέγχουμε τη συμπεριφορά ενός αλγορίθμου ως προς τα φαινόμενα underfitting και overfitting τροποποιώντας τη χωρητικότητα του μοντέλου

Το βέλτιστο ως προς τη πολυπλοκότητα και γενίκευση μοντέλο υπάρχει ωστόσο δεν μπορεί να προσδιοριστεί μελετώντας μόνο το σφάλμα εκπαίδευσης.

Το σφάλμα ελέγχου μπορεί να γραφεί ως άθροισμα τριών συνιστωσών, οι οποίες εκφράζουν διαφορετικές πηγές σφάλματος στην εκτιμώμενη συνάρτηση.
Παράδειγμα
$y_i = f^*(x) + \epsilon_i, \epsilon \sim N(0, \sigma^2)$
$R(\hat{f}) = E\left[(\hat{f}(x) - y)^2\right] = E\left[(E[\hat{f}(x)] - f^*(x))^2\right] + E\left[(\hat{f}(x) - E[\hat{f}(x)])^2\right] + \sigma^2 = Bias^2(\hat{f}) + Variance(\hat{f}) + Noise$
Η μεροληψία (bias) μετράει την αναμενόμενη απόκλιση της εκτίμησης από την πραγματική συνάρτηση.
Η διακύμανση (variance) ποσοτικοποιεί την απόκλιση του εκτιμητή από τη μέση τιμή του, η οποία προκαλείται από τη χρήση διαφορετικών συνόλων εκπαίδευσης ή/και μοντέλων διαφορετικής πολυπλοκότητας.
Ο παράγοντας θορύβου σχετίζεται με το θόρυβο που εγγενώς υπάρχει στα δεδομένα.

Υπερπαράμετροι μοντέλων μάθησης

Οι περισσότεροι αλγόριθμοι μηχανικής μάθησης έχουν υπερ-παραμέτρους (hyperparameters), δηλαδή παραμέτρους οι οποίες προσδιορίζουν την συμπεριφορά ενός αλγορίθμου.

Έστω ένα σύνολο μοντέλων μάθησης διαφορετικής πολυπλοκότητας (χωρητικότητας) τα οποία διατάσσονται ως προς αύξουσα πολυπλοκότητα: $\{H_\lambda\}_{\lambda\in\Lambda}, H_1 \prec H_2 \prec \cdots$

Το επιθυμητό μοντέλο είναι αυτό έχει την μικρότερη πολυπλοκότητα και ελαχιστοποιεί το σφάλμα γενίκευσης (πραγματικό ρίσκο): $\hat{f} = \min_{\lambda} \min_{f\in H_\lambda} R(f)$

Πρακτικά, μπορούμε να επιλέξουμε το μοντέλο κατάλληλης πολυπλοκότητας χωρίζοντας τα διαθέσιμα δεδομένα σε υποσύνολα εκπαίδευσης και επικύρωσης (validation set). Τα σύνολο επικύρωσης χρησιμοποιούνται για να εκτιμήσουμε την τιμή του πραγματικού ρίσκου.

Μέθοδος Hold-out

Χωρίζουμε τα διαθέσιμα δεδομένα σε δύο σύνολα: $D = \{(x_i, y_i)\}_{i=1}^{M}, D_V = \{(x_i, y_i)\}_{i=M+1}^{N}$
Χρησιμοποιούμε το σύνολο εκπαίδευσης για να εκπαιδεύσουμε ένα μοντέλο για κάθε κλάση πολυπλοκότητας (υποθέσεων): $\hat{f}_\lambda = \arg\min_{f\in H_\lambda} \hat{R}(f) \rightarrow \{\hat{f}_{\lambda=1}, \hat{f}_{\lambda=2}, \dots\}$
Χρησιμοποιούμε το σύνολο επικύρωσης για να διαλέξουμε το μοντέλο με το μικρότερο σφάλμα γενίκευσης:
$\hat{\lambda} = \arg\min_{\lambda\in\Lambda} \hat{R}_V(\hat{f}_\lambda)$ και $\hat{f} = \hat{f}_{\hat{\lambda}}$
Μειονεκτήματα:

1. Το πλήθος των διαθέσιμων δεδομένων μπορεί να είναι μικρό
2. Το σύνολο επικύρωσης μπορεί να μην είναι αντιπροσωπευτικό του συνόλου ελέγχου

Μέθοδος Cross-Validation

Χωρίζουμε τα διαθέσιμα δεδομένα σε K σύνολα (folds).

Για κάθε κλάση πολυπλοκότητας λ εκπαιδεύουμε K εκτιμήσεις χρησιμοποιώντας τα $K-1$ σύνολα για εκπαίδευση και 1 σύνολο για επικύρωση. Ουσιαστικά εκτελούμε την μέθοδο Hold out K φορες.

Χρησιμοποιούμε τα Κ σύνολα επικύρωσης για να διαλέξουμε το μοντέλο με το μικρότερο κατά μέσο όρο σφάλμα γενίκευσης: $\hat{\lambda} = \arg\min_{\lambda\in\Lambda} \frac{1}{K}\sum_{k=1}^{K} \hat{R}_{V_k}(\hat{f}_\lambda^k)$

Η μέθοδος επιστρέφει το μοντέλο που κατά μέσο όρο στα K πειράματα έχει μικρότερο σφάλμα γενίκευσης.

L07 – Μέθοδοι κανονικοποίησης

Ελαχιστοποίηση εμπειρικού σφάλματος – Empirical Risk Minimization ERM

Γενική ιδέα: Θωρούμε ότι το εμπειρικό ρίσκο υποκαθιστά το πραγματικό ρίσκο

$\hat{R}(f) = E_{(x,y)\sim D}\left[l(y, f(x))\right] = \frac{1}{N}\sum_{i=1}^{N} l(y_i, f(x_i))$

Για να μεταβούμε από την γενική ιδέα σε αλγόριθμο μηχανικής μάθησης:

1. Επιλέγουμε εκ των προτέρων ένα κατάλληλο χώρο υποθέσεων H.
2. Μαθαίνουμε την συνάρτηση f (προσδιορίζουμε τις παραμέτρους της) ελαχιστοποιώντας το εμπειρικό σφάλμα στο χώρο υποθέσεων που επιλέξαμε: $\hat{f} = \min_{f\in H} \hat{R}(f)$

ERM σε σύνθετους χώρους υποθέσεων

Εάν ο χώρος υποθέσεων είναι αρκετά πλούσιος, η επίλυση του ERM συχνά οδηγεί σε overfitting και συνεπώς μεγάλο σφάλμα γενίκευσης.

Οι τεχνικές κανονικοποίησης (regularization techniques) επιτρέπουν ευσταθείς (stable) εκτιμήσεις και τη μείωση του σφάλματος γενίκευσης. Επιπλέον κάποιες τεχνικές κανονικοποίησης επιτρέπουν και μείωση διαστάσεων και συνεπώς συνδράμουν στο περιορισμό της κατάρας των μεγάλων διαστάσεων (curse of dimensionality).

Κανονικοποίηση (Regularization) της πολυπλοκότητας

Structural risk minimization (SRM): Εξισορρόπηση της πολυπλοκότητα του μοντέλου επιβάλλοντας “ποινή” στα μοντέλα που αποκλίνουν από το πραγματικό ρίσκο (άνω φράγμα).

Ισχύει (concentration bounds): $\left|R(f) - \hat{R}(f)\right| \leq C(f), \forall f \in H$

SRM: $\hat{f} = \min_{f\in H}\left\{\hat{R}(f) + \lambda \cdot C(f)\right\} = \min_{f\in H}\left\{\frac{1}{N}\sum_{i=1}^{N} l(y_i, f(x_i)) + \lambda \cdot C(f)\right\}$

Η συνάρτηση $C(\cdot)$ καλείται regularizer ενώ το λ αποτελεί υπερ-παράμετρο.

Οι πιο συνηθισμένες συναρτήσεις κανονικοποιήσης (regularizers):

1. $l2 - norm$: $\|x\|_2^2 = x^T x = \sum_{i=1}^{d} x_i^2$
2. $l1 - norm$: $\|x\|_1 = \sum_{i=1}^{d} |x_i|$

Κανονικοποιήση Tikhonov (Tikhonov regularization)

Το σχήμα κανονικοποίσης του Tikhonov, χρησιμοποιεί την Ευκλείδεια απόσταση ως regularizer στο φορμαλισμό του SRM.

$w^* = \min_{w}\{\sum_{i=1}^{N} l(y_i, f(x_i; w)) + \lambda \cdot \|w\|_2^2\}$

Τα μοντέλα (παραμετρικές συναρτήσεις) που προκύπτουν ως λύση του παραπάνω προβλήματος είναι γνωστά και ως δίκτυα κανονικοποίησης (regularization networks)

Η Ευκλείδεια νόρμα ελέγχει την ευστάθεια της λύσης εμποδίζοντας το φαινόμενο overfitting.

Η υπερπαράμετρος λ ισορροπεί το εμπειρικό σφάλμα και τον regulariser και προσδιορίζεται μέσω cross-validation.

Δεν υπάρχει γενικός αλγόριθμος επίλυσης του παραπάνω προβλήματος. Η λύση εξαρτάται από την επιλογή της συνάρτησης κόστους.

Στη περίπτωση που η συνάρτηση κόστους είναι η τετραγωνική απόσταση, η μέθοδος που προκύπτει ονομάζεται ρυθμισμένα ελάχιστα τετράγωνα (regularized least squares).

Ridge regression

$w^* = \min_w \left\{ \frac{1}{N} \cdot \sum_{i=1}^{N} (y_i - w^T x_i)^2 + \lambda \cdot \|w\|_2^2 \right\} = \min_w \{ \|y - Xw\|_2^2 + \lambda \cdot \|w\|_2^2 \}$, $X \in \mathbb{R}^{Nxd}$, $y \in \mathbb{R}^N$, $w \in \mathbb{R}^d$

Πώς βρίσκουμε την λύση του προβλήματος ridge regression:

$$l(w) = \frac{1}{N} \cdot \|y - Xw\|_2^2 + \lambda \cdot \|w\|_2^2 = \frac{1}{N} \cdot (y - Xw)^T (y - Xw) + \lambda \cdot w^T w = \frac{1}{N} \cdot (y^T - w^T X^T)(y - Xw) + \lambda \cdot w^T w =$$
$$= \frac{1}{N} \cdot (y^t y - y^T Xw - w^T X^T y + w^T X^T Xw) + \lambda \cdot w^T w$$
$$\frac{\partial l(w)}{\partial w} = \frac{1}{N} \cdot (0 - (y^T X)^T - X^T y + X^T Xw + (w^T X^T X)^T) + \lambda \cdot (w + (w^T)^T) = \frac{1}{N} \cdot (-2 \cdot X^T y + 2 \cdot X^T Xw) + 2 \cdot \lambda \cdot w$$
$$\frac{\partial l(w)}{\partial w} = 0 \Rightarrow -2 \cdot X^T y + 2 \cdot X^T Xw + 2 \cdot \lambda \cdot N \cdot w = 0 \Leftrightarrow$$
$$\Leftrightarrow (X^T X + \lambda \cdot N \cdot I) w = X^T y \Rightarrow w^* \begin{cases} (X^T X + \lambda \cdot N \cdot I)^{-1} X^T y & , if\ X^T X\ is\ invertible \\ X^\dagger y & , otherwise \end{cases}$$

Η υπερπαράμετρος λ ελέγχει την αντιστρεψιμότητα του πίνακα: $X^T X + \lambda \cdot N \cdot I$

Ελάχιστα τετράγωνα σε πολλές διαστάσεις

Όταν το πλήθος των δεδομένων εισόδου N, είναι μεγαλύτερο από τη διάσταση τους d, δηλαδή ισχύει $d < N$, τότε η λύση των ελαχίστων τετραγώνων είναι μοναδική και υπολογίζεται σε κλειστή μορφή.
Πολλές φορές στη πράξη λόγω της διαστατικότητας των δεδομένων και του υψηλού κόστους συλλογής τους ισχύει $d > N$. Σε αυτή τη περίπτωση υπάρχει πολύ μεγάλη πιθανότητα να εμφανιστεί το φαινόμενο της συγγραμμικότητας (collinearity): $x_i = \sum_{j \in S} x_j a_j$

Εάν έχουμε ισχυρή συγγραμμικότητα τότε ο πίνακας $X^T X$ είναι χαμηλής τάξης και συνεπώς ιδιάζων (μη αντιστρέψιμος), οπότε το σύστημα κανονικών εξισώσεων δεν μπορεί να λυθεί.
Εάν έχουμε προσεγγιστική συγγραμμικότητα: $x_i \approx \sum_{j \in S} x_j a_j$, τότε ο $X^T X$ είναι κοντά στο να είναι ιδιάζων, ωστόσο αντιστρέψιμος.

Σε αυτή τη περίπτωση η διακύμανση της εκτίμησης των παραμέτρων είναι μεγάλη και συνεπώς ο εκτιμητής είναι ασταθής (unstable).

Lasso regression, αραιότητα και επιλογή χαρακτηριστικών

Ένα διάνυσμα είναι αραιό (sparse) όταν τα περισσότερα στοιχεία του είναι μηδενικά και μόνο ένα μικρό πλήθος στοιχείων είναι μη μηδενικό.
Για να βρούμε αραιά διανύσματα παραμέτρων αρκεί να λύσουμε ένα σταθμισμένο πρόβλημα ελαχίστων τετραγώνων με regularizer την l1-νόρμα. Το πρόβλημα αυτό είναι γνωστό ως Lasso regression και εκφράζεται μαθηματικά ως εξής: $w^* = \min_w \left\{ \frac{1}{N} \cdot \sum_{i=1}^{N} (y_i - w^T x_i)^2 + \lambda \cdot \|w\|_1 \right\} = \min_w \{ \|y - Xw\|_2^2 + \lambda \cdot \|w\|_1 \}$
Το διάνυσμα στήλη w αναπαριστά τις άγνωστες παραμέτρους του μοντέλου, και μόνο ένα μικρό υποσύνολο τους έχουν μη-μηδενικές τιμές.

Εύρωστη αναγνώριση προσώπου με Lasso

$$y = Xw + e = [X \quad I] \begin{bmatrix} w \\ e \end{bmatrix} = Ac$$

Το διάνυσμα στήλη w έχει διάσταση N και αναπαριστά τους αγνώστους συντελεστές του αραιού γραμμικού συνδυασμού.
Το διάνυσμα στήλη e έχει διάσταση d και αναπαριστά τον αραιό θόρυβο (π.χ. occlusions).

$c^* = \min_c \|y - Xc\|_2^2 + \lambda \cdot \|c\|_1$

Συμπεράσματα

- Παλινδρόμηση ελαχίστων τετραγώνων = πρόβλεψη
- Παλινδρόμηση ελαχίστων τετραγώνων + Tikhonov reg. = πρόβλεψη + ευστάθεια
- Παλινδρόμηση ελαχίστων τετραγώνων + l1 reg.(lasso) = πρόβλεψη + ευστάθεια + επιλογή χαρακτηριστικών (μείωση διαστάσεων)

L08 – Μέθοδοι Μείωσης Διαστάσεων

Η διαστατικότητα των δεδομένων και οι συνέπειες της

Δεν υπάρχουν δεδομένα-διανύσματα μεγάλων διαστάσεων που να βρίσκονται κοντά με βάση κάποια απόσταση στις μεγάλες διαστάσεις. Χρειαζόμαστε ϵ^{-d} δεδομένα εκπαίδευσης για να καλύψουμε ομοιόμορφα το χώρο διάστασης d.
Παράδειγμα: 10^d σημεία χρειάζονται για το χώρο $[0,1]^d$ σε απόσταση 0.1

Όσο μεγαλώνει η διάσταση των δεδομένων, χρειαζόμαστε εκθετικό στη διάσταση των δεδομένων πλήθος δεδομένων έτσι ώστε να υπάρχουν εγγυήσεις γενίκευσης.
Τα δεδομένα μεγάλης διάστασης απαιτούν αυξημένο χώρο μνήμης για την αποθήκευση τους και αυξημένο χρόνο για την εξεργασία τους.
Το σύνολο των συνεπειών της μεγάλης διάστασης των δεδομένων αναφέρεται συχνά ως κατάρα της διαστατικότητας (curse of dimensionality).

Εξαγωγή χαρακτηριστικών

Στόχος είναι η μείωση της διάστασης των δεδομένων, εφαρμόζοντας το μετασχηματισμό $\Phi : X \to R$.
$x \in X \subseteq \mathbb{R}^D \xrightarrow{\Phi} \Phi(x) \in R \subseteq \mathbb{R}^d$, ώστε $d \ll D$
Αν ο μετασχηματισμός Φ είναι προκαθορισμένος, τότε αναφερόμαστε σε εξαγωγή χαρακτηριστικών, πχ. SIFT, HOGs, MFCCs.
Αν τα δεδομένα έχουν κάποια λανθάνουσα γεωμετρική δομή μπορούμε να μάθουμε το μετασχηματισμό Φ από τα δεδομένα, χρησιμοποιώντας κυρίως μεθόδους μάθησης χωρίς επίβλεψη. Στη περίπτωση αυτή αναφερόμαστε σε μάθηση αναπαραστάσεων (representation learning).

Μείωση διαστάσεων (Dimensionality reduction)

Το πρόβλημα της μείωσης διαστάσεων μπορεί να θεωρηθεί μαθηματικά ως πρόβλημα εύρεσης μίας απεικόνισης ή μετασχηματισμού M από ένα χώρο πολλών διαστάσεων σε ένα χώρο (πολύ) μικρότερης διάστασης έτσι ώστε ικανοποιείται κάποιο κατάλληλο κριτήριο (π.χ. ακρίβεια ανακατασκευής των δεδομένων): $M : \mathbb{R}^D \to \mathbb{R}^k$, $k \ll D$

Principal Component Analysis – PCA

Η ανάλυση κύριων συνιστωσών αποτελεί μη επιτηρούμενη μέθοδο (unsupervised) μείωσης διαστάσεων η οποία αξιοποιώντας την υποκείμενη δομή των δεδομένων εξάγει μια γραμμική απεικόνιση από τον D-διάστατο χώρο σε ένα (υπο) χώρο k διαστάσεων. Οι k διαστάσεις του χώρου μειωμένης διάστασης αποτελούν τις κύριες συνιστώσες των δεδομένων και ορίζουν ένα υπερεπίπεδο F.

Προβολή σε υποχώρους

Εάν $w, x \in \mathbb{R}^D$, $\|w\|_2 = 1$, τότε η ορθογώνια προβολή του x στο w είναι η $(w^t x)w = ww^t x = Px$
Ο τετραγωνικός πίνακας P καλείται πίνακας προβολής και προβάλει τα δεδομένα D διαστάσεων στον υποχώρο μίας διάστασης που ορίζεται από το w.

Εύρεση κύριας συνιστώσας

Ελαχιστοποίηση σφάλματος ανακατασκευής
Πρόβλημα: Προσδιόρισε ένα διάνυσμα w του οποίου η κατεύθυνση επιτρέπει τη βέλτιστη ανακατασκευή του x.
$w^* = \arg\min_w \|x - ww^T x\|_2^2, subject\ to\ \|w\|_2 = 1$
Ο περιορισμός μοναδιαίου μήκους που επιβάλουμε στο w είναι απαραίτητος για να αποφύγουμε την τετριμμένη λύση.
Το σφάλμα ανακατασκευής ποσοτικοποιεί τη πληροφορία που χάνουμε εάν προβάλουμε τα δεδομένα x στο w.
Κωδικοποίηση: $z = w^T x$
Αποκωδικοποίηση: $\tilde{x} = wz = ww^T x$
Επιθυμούμε το σφάλμα ανακατασκευής να είναι μικρό: $\|x - \tilde{x}\|_2^2$

Για να βρούμε τη κύρια συνιστώσα ενός συνόλου N δεδομένων, αρκεί να ελαχιστοποιούμε το αναμενόμενο σφάλμα ανακατασκευής: $w^* = \arg\min_w E[\|x - ww^T x\|_2^2], subject\ to\ \|w\|_2 = 1$

Στη πράξη ελαχιστοποιούμε το εμπειρικό σφάλμα ανακατασκευής:

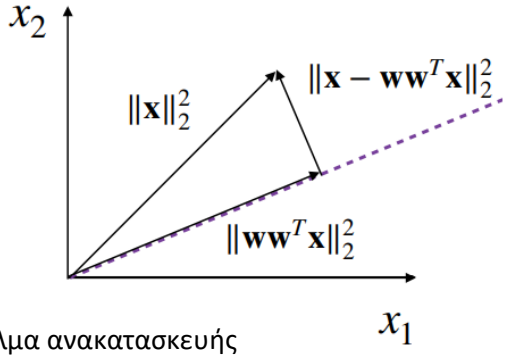
$w^* = \arg\min_w \sum_{n=1}^N \|x_n - ww^T x_n\|_2^2, subject\ to\ \|w\|_2 = 1$

Μεγιστοποίηση διακύμανσης

Μια διαφορετική προσέγγιση στην εύρεση της κύριας συνιστώσας είναι να βρούμε τη προβολή w η οποία συλλαμβάνει (εξηγεί) τη διακύμανση των δεδομένων. (Υποθέτουμε ότι τα δεδομένα έχουν μηδενική μέση τιμή)

Από Πυθαγόρειο Θεώρημα: $\|x\|_2^2 = \|ww^T x\|_2^2 + \|x - ww^T x\|_2^2$

Επειδή το w έχει μήκος ίσο με 1 $= \|w^T x\|_2^2 + \|x - ww^T x\|_2^2$



$E[\|x\|_2^2] = E[\|ww^T x\|_2^2] + E[\|x - ww^T x\|_2^2]$

Διακύμανση των δεδομένων = διακύμανση προβολής + σφάλμα ανακατασκευής

Σταθερή Επιθυμούμε να είναι μεγάλη Επιθυμούμε να είναι μικρό

Αντικειμενική συνάρτηση:

$$w^* = \arg\max_{\|w\|_2=1} E[\,\|w^T x\|_2^2] = \arg\max_{\|w\|_2=1} \frac{1}{N}\sum_{n=1}^N \|w^T x\|_2^2 = \arg\max_{\|w\|_2^2=1} \frac{1}{N}\|w^T X\|_2^2 = \arg\max_{\|w\|_2^2=1} \frac{1}{N} w^T X (w^T X)^T$$

$$= \arg\max_{\|w\|_2^2=1} \frac{1}{N} w^T X X^T w = \arg\max_{\|w\|_2^2=1} w^T \left(\frac{1}{N} X X^T\right) w = \arg\max_w \frac{w^T C w}{w^T w} = \arg\max_w w^T C w - \lambda \cdot w^T w$$

Πίνακας Συνδιακύμανσης

$$C = XX^T = \begin{pmatrix} var(x_1) & \cdots & cov(x_1, x_N) \\ \vdots & \ddots & \vdots \\ cov(x_N, x_1) & \cdots & var(x_N) \end{pmatrix}$$

Ο πίνακας συνδιακύμανσης είναι συμμετρικός: $C = C^T$ και περιγράφει τόσο την διασπορά (διακύμανση) όσο και τον προσανατολισμό (συνδιακύμανση) των δεδομένων μας.

Στο πίνακα συνδιακύμανσης (όπως και σε κάθε τετραγωνικό πίνακα), αντιστοιχούν ιδιοδιάνυσματα (eigenvectors) και ιδιοτιμές (eigenvalues) τέτοια ώστε να ισχύει η παρακάτω εξίσωση: $Cv_i = \lambda_i v_i$

Για να βρούμε την κύρια συνιστώσα των δεδομένων αρκεί να λύσουμε το παρακάτω πρόβλημα μεγιστοποίησης:

$w^* = \arg\max_w l(w) = \arg\max_w w^T C w - \lambda \cdot w^T w$

$\frac{\partial l(w)}{\partial w} = Cw + (w^T C)^T - \lambda \cdot w (w^T)^T = Cw + C^T w - 2 \cdot \lambda \cdot w = Cw + Cw - 2 \cdot \lambda \cdot w = 2 \cdot Cw - 2 \cdot \lambda \cdot w$

$\frac{\partial l(w)}{\partial w} = 0 \Rightarrow Cw^* = \lambda \cdot w^*$

Θα δείξουμε ότι η μεγιστοποίησης της διακύμανσης κατά μήκος της κύριας συνιστώσας είναι ισοδύναμη με την ελαχιστοποίηση του σφάλματος ανακατασκευής.

$$Σφάλμα\ Ανακατασκευής = \sum_{n=1}^N \|x_n - ww^T x_n\|_2^2 = \sum_{n=1}^N (\|x_n\|_2^2 - \|w^T x_n\|_2^2)$$
$$= \sum_{n=1}^N \|x_n\|_2^2 - \sum_{n=1}^N \|w^T x_n\|_2^2 = \sum_{n=1}^N \|x_n\|_2^2 - \|w^T X\|_2^2$$

Συνεπώς: $w^* = \arg\min_w \sum_{n=1}^N \|x_n - ww^T x_n\|_2^2\ subject\ to\ \|w\|_2 = 1 \Leftrightarrow \arg\max_{\|w\|_2^2=1} w^T \left(\frac{1}{N} XX^T\right) w$

Εύρεση k κύριων συνιστωσών

Για να βρούμε k ορθογώνιες μεταξύ τους κύριες συνιστώσες ελαχιστοποιούμε το σφάλμα ανακατασκευής υπό τον περιορισμό της ορθογωνιότητας. $W^* = \arg\min_w \sum_{n=1}^N \|x_n - ww^T x_n\|_2^2 = \|X - WW^T X\|_F^2\ subject\ to\ W^T W = I$,

Ο W έχει ως στήλες k ιδιοδιανύσματα που αντιστοιχούν στις k μεγαλύτερες ιδιοτιμές του πίνακα συνδιακύμανσης και αποτελούν μια ορθοκανονική βάση.

SVD: Singular Value Decomposition (ανάλυση ιδιαζουσών τιμών)

Ως βαθμός (rank) ενός πίνακα A ορίζεται ο μέγιστος αριθμός των γραμμικά ανεξάρτητων στηλών (γραμμών) του A.
$A = U\Sigma V^T, UU^T = V^TV = I \in \mathbb{R}^{r\times r}$
Το πλήθος των μη-μηδενικών ιδιαζουσών τιμών ισούται με το βαθμό του πίνακα A.
Πώς βρίσκουμε μια χαμηλού βαθμού k προσέγγιση (low-rank approximation) A_k ενός πίνακα A;
$A_k^* = \arg\min_{A_k} \|A - A_k\|_2^2 \;\; s.t. rank(A_k) \le k$
Λύση: Truncated Singular Value Decomposition: $A_k^* = U_k\Sigma_k V_k^T = U_k U_k^T A$
Στη περίπτωση της PCA ο πίνακας WW^TX έχει βαθμό μικρότερο ή ίσο του k. Κατα συνέπεια, η PCA είναι ισοδύναμη με τη χαμηλού βαθμού προσέγγιση του πίνακα δεδομένων, και μπορεί να υπολογιστεί μέσω της SVD.
Για να προσδιορίσουμε την απεικόνιση M που προβάλει τα δεδομένα D διαστάσεων στο γραμμικό υποχώρο k διαστάσεων επιλύουμε: $\min_W \|X - WW^TX\|_F^2 \;\; s.t. W^TW = I$ μέσω ιδιοανάλυσης
$\min_W \|X - W\|_F^2 \;\; s.t. rank(W) \le k$ μέσω SVD
Χαμηλόβαθμη αναπαράσταση: $W^Tx \in \mathcal{F}, W \in \mathbb{R}^k, j \ll D$

Γενικεύσεις της PCA

Στο αρχικό μοντέλο της PCA θεωρήσαμε ότι το μοντέλο δεδομένων είναι: $X \approx WW^TX$
Θέτοντας τη χαμηλόβαθμη αναπαράσταση ίση με $C = W^TX: X \approx WC, W \in \mathbb{R}^{D\times k}, C \in \mathbb{R}^{k\times N}, X \in \mathbb{R}^{D\times N}$
Συνεπώς μπορούμε να σχεδιάσουμε γενικευμένα μοντέλα PCA, όπως για παράδειγμα κανονικοποιημένη PCA (regularized PCA) για να ενισχύσουμε την ικανότητα γενίκευσης: $\min_{W,C} \|X - WC\|_F^2 + \gamma \cdot \sum_{i=1}^D \|w_i\|_2^2 + \gamma \cdot \sum_{j=1}^k \|c_j\|_2^2$

Εναλλασσόμενη ελαχιστοποίηση
$choose\ initial\ starting\ points\ W^{(0)}\ and\ C^{(0)}$
$n \leftarrow 0$
$while\ not\ converged$:
    $W^{(n+1)} \leftarrow minimize\ over\ W\ while\ holding\ C = C^{(n)}\ constant$
    $C^{(n+1)} \leftarrow minimize\ over\ C\ while\ holding\ W = W^{(n+1)}\ constant$
    $n \leftarrow n + 1$

Non-negative Matrix Factorization

Η PCA και οι σχετικές με αυτή μέθοδοι αναπαριστούν τα δεδομένα ως γραμμικό συνδυασμό k διανυσμάτων βάσης τα οποία συλλαμβάνουν ολιστική, δύσκολα ερμηνεύσιμη πληροφορία για τα δεδομένα. Επιπλέον, δεν διατηρεί την μη-αρνητικότητα των τιμών των δεδομένων.

Μπορούμε να σχεδιάσουμε μια μέθοδο μείωσης διαστάσεων που να διατηρεί τη μη- αρνητικότητα των δεδομένων και να εξάγει αραιές και εύκολα ερμηνεύσιμες συνιστώσες: $X \approx WC,\ X \in \mathbb{R}_+^{D\times N}, W \in \mathbb{R}_+^{D\times k},\ C \in \mathbb{R}_+^{k\times N}$

NMF: πρόβλημα βελτιστοποίησης: $\min_{W,C} l(W,C) = \min_{W,C} \|X - WC\|_F^2 \;\; s.t. W, C \ge 0$
Αρχικοποιήση: Αρχικοποιούμε τις άγνωστες μεταβλητές W και C με τυχαίες μη-αρνητικές τιμές.
Gradient descent steps
$C_{[t+1]} = C_{[t]} - n_t \cdot \nabla_{C_{[t]}} l(W, C_{[t]})$
$W_{[t+1]} = W_{[t]} - n_t \cdot \nabla_{W_{[t]}} l(W_{[t]}, C)$
$t = t + 1$
Οι κανόνες επανάληψης δεν επιβάλλουν την απαιτούμενη μη-αρνητικότητα των μεταβλητών.
Παρατήρηση: Κάθε αριθμός μπορεί να εκφραστεί ως διαφορά μη αρνητικών αριθμών:
$\nabla_C l(W,C) = \nabla_C l(W,C)^+ - \nabla_C l(W,C)^- = W^TWC - W^TX$
$n = \frac{C}{\nabla_C l(W,C)^+} = \frac{C}{WW^TC}$

$C = C - \frac{C}{WW^TC} * \nabla_C l(W,C) = C * \frac{W^TX}{WW^TC}$

$W = W * \frac{XC^T}{WCC^T}$

Αποδείξεις

$l(W,C) = \|X - WC\|_F^2 = tr\big((X-WC)(X-WC)^H\big) = tr((X-WC)(X-WC)^T) = tr\big((X-WC)(X^T - C^T W^T)\big)$
$= tr(XX^T - XC^T W^T - WCX^T + WCC^T W^T) = tr(XX^T) - tr(XC^T W^T) - tr(WCX^T) + tr(WCC^T W^T)$

$\nabla_C l(W,C) = 0 - W^T X - W^T (X^T)^T + W^T (C^T W^T)^T + W^T WC = 2 \cdot (W^T WC - W^T X)$

$\nabla_W l(W,C) = 0 - XC^T - (CX^T)^2 + (CC^T W^T)^T + WCC^T = 2 \cdot (WCC^T - XC^T)$

Προσδιορίζουμε (θέτουμε) τα step sizes με τρόπο τέτοιο ώστε να εξαρτούνται από τα δεδομένα ως εξής:

$\eta = \frac{1}{2} \cdot \frac{C}{\nabla_C l(W,C)^+} = \frac{1}{2} \cdot \frac{C}{W^T WC}$

$\zeta = \frac{1}{2} \cdot \frac{W}{\nabla_W l(W,C)^+} = \frac{1}{2} \cdot \frac{W}{WCC^T}$

$C = C - \eta \cdot \nabla_C l(W,C) = C - \frac{1}{2} \cdot \frac{C}{W^T WC} * 2 \cdot (W^T WC - W^T X) = C - C + C * \frac{W^T X}{W^T WC} = C * \frac{W^T X}{W^T WC}$

$W = W - \zeta \cdot \nabla_W l(W,C) = W - \frac{1}{2} \cdot \frac{W}{WCC^T} * 2 \cdot (WCC^T - XC^T) = W - W + W * \frac{XC^T}{WCC^T} = W * \frac{XC^T}{WCC^T}$

Γραμμική διακριτή ανάλυση – LDA

Οι μέθοδοι μείωσης διάστασης και εξάγουν χαρακτηριστικά (αναπαραστάσεις) χωρίς να λαμβάνουν υπόψιν την πληροφορία κλάσης που πιθανόν συνοδεύουν τα δεδομένα. Συνεπώς, δεν εξάγουν χαρακτηριστικά που να είναι βέλτιστα για προβλήματα ταξινόμησης.

Μπορούμε να σχεδιάσουμε μια μέθοδο μείωσης διαστάσεων που να μεγιστοποιεί τον διαχωρισμό των δεδομένων ανάλογα με τη κλάση που ανήκουν και συνεπώς να εξάγει αναπαραστάσεις (χαρακτηριστικά) οι οποίες έχουν διακριτική ικανότητα

Η γραμμική διακριτική ανάλυση (linear discriminant analysis - LDA ή Fisher discriminant analysis - FDA) στοχεύει στο να μετασχηματίσει τα δεδομένα, προβάλλοντάς τα σε ένα χώρο μικρότερης διάστασης, έτσι ώστε ο διαχωρισμός των δεδομένων διαφορετικών κλάσεων να μεγιστοποιείται ενώ παράλληλα η διακύμανση των δεδομένων κάθε κλάσης ελαχιστοποιείται ώστε η πιθανότητα σύμπτωσης διαφορετικών κλάσεων να είναι μικρή.

Πίνακας συνδιακύμανσης μεταξύ κλάσεων: $S_B = \sum_{classes\ c} N_c \cdot (\mu_c - \mu)\,(\mu_c - \mu)^T = \sum_{classes\ c} N_c \cdot \|\mu_c - \mu\|_2^2$

Πίνακας συνδιακύμανσης εντός κλάσεων: $S_W = \sum_{classes\ c} \sum_{j \in c} \big(x_j - \mu_c\big)\,\big(x_j - \mu_c\big)^T = \sum_{classes\ c} \sum_{j \in c} \big\|x_j - \mu_c\big\|_2^2$

Η LDA βρίσκει το βέλτιστο μετασχηματισμό W διάστασης $D \times (c-1)$ (το πολύ) ο οποίος να μεγιστοποιεί τη διακύμανση μεταξύ κλάσεων και να ελαχιστοποιεί την διακύμανση εντός κλάσεων των δεδομένων όταν αυτά προβάλλονται στο χώρο διάστασης $c-1$, λύνοντας το πρόβλημα βελτιστοποίησης: $W^* = \arg\max_W tr\left(\frac{W^T S_B W}{W^T S_W W}\right)$

Το πρόβλημα εκφράζεται ισοδύναμα ως εξής: $W^* = arg\max_W l(W) = arg\max_W tr(W^T S_B W) - \lambda \cdot tr(W^T S_W W)$

Συνεπώς ο βέλτιστος μετασχηματισμός από τα k ιδιοδιανίσματα που αντιστοιχούν στις k μεγαλύτερες ιδιοτιμές του πίνακα: $S_W^{-1} S_B$

Απόδειξη:

Ως πίνακες συνδιακύμανσης ισχύει ότι $S_B = S_B^T$ και $S_W = S_W^T$

$l(W) = tr(W^T S_B W) - \lambda \cdot tr(W^T S_W W)$

$\frac{\partial l(W)}{\partial W} = S_B W + S_B^T W - \lambda \cdot (S_W W + S_W^T W) = 2 \cdot S_B W - 2 \cdot \lambda \cdot S_W W$

$\frac{\partial l(W)}{\partial W} = 0 \Rightarrow 2 \cdot S_B W - 2 \cdot \lambda \cdot S_W W = 0 \Leftrightarrow S_B W = \lambda \cdot S_W W \Leftrightarrow S_W^{-1} S_B W = \lambda \cdot W$

Συνεπώς βλέπουμε πως προκύπτει ότι ο βέλτιστος μετασχηματισμός σχετίζεται με τα ιδιοδιάνυσματα του $S_W^{-1} S_B$

L09 – Support Vector Machines

Γραμμικός ταξινομητής: 2 κλάσεις

Εάν τα δεδομένα X διαχωρίζονται γραμμικά σε 2 κλάσεις τότε ο ταξινομητής αποτελεί ένα υπερεπίπεδο που ορίζει το σύνορο απόφασης (decision boundary): $f(x) = w^T x + b$
$f(x) > 0 \; iff \; x \; is \; of \; class \; A$
$f(x) < 0 \; iff \; x \; is \; of \; class \; B$
Το σύνορο (decision boundary) που διαχωρίζει τις κλάσεις (κατηγορίες), πρέπει να είναι όσο το δυνατόν πιο μακριά από τα δεδομένα εκπαίδευσης.

Κεντρική ιδέα: Ένας γραμμικός ταξινομητής ορίζεται ως υπερεπίπεδο στις d διαστάσεις.
Οποιοδήποτε υπερεπίπεδο εκφράζεται ως το σύνολο των σημείων x τα οποία ικανοποιούν την παρακάτω εξίσωση
$< w, x > + b = w^T x + b$
Το διάνυσμα w είναι κανονικό διάνυσμα (normal vector), δηλαδή είναι κάθετο στο υπερεπίπεδο. Η παράμετρος b καθορίζει τη μετατόπιση του υπερεπιπέδου από την αρχή των αξόνων κατά μήκος του διανύσματος w.
$dist \; to \; origin = \frac{|b|}{\|w\|_2}$

Η έννοια του περιθωρίου

Έστω d^+ η απόσταση του πλησιέστερου στο υπερεπίπεδο θετικού δείγματος και d^- η απόσταση του πλησιέστερου στο υπερεπίπεδο αρνητικού δείγματος.
Περιθώριο (margin) ορίζεται η απόσταση: $m = d^+ + d^-$
Στόχος μας είναι να βρούμε τον ταξινομητή $f(x)$ για τον οποίο το margin είναι μέγιστο.

Η έννοια των διανυσμάτων στήριξης (support vectors)

Υπάρχουν w και b για τα οποία ισχύει: $d^+ = d^- = \frac{1}{\|w\|_2} \Rightarrow m = \frac{2}{\|w\|_2}$
Τα σημεία δεδομένων που βρίσκονται σε απόσταση $\frac{1}{\|w\|_2}$ από το υπερεπίπεδο ονομάζονται διανύσματα στήριξης (support vectors)
Τα διανύσματα στήριξης ορίζουν δύο επίπεδα παράλληλα στον υπερεπίπεδο του ταξινομητή $f(x) = 0$
Επομένως ο στόχος μας είναι να μάθουμε τις w και b του γραμμικού ταξινομητή $f(x)$ ο οποίος μεγιστοποιεί το περιθώριο m. Αυτός ο ταξινομητής ονομάζεται support vector machine.

Πρόβλημα βελτιστοποίησης: $\max_{w,b} \frac{1}{\|w\|_2} \; s.t. \; y_i(w^T x_i + b) \geq 1, \forall \{x_i, y_i\} \in D$
Το παραπάνω πρόβλημα αποτελεί τετραγωνικό πρόβλημα βελτιστοποίησης (quadratic problem) για το οποίο υπάρχουν αρκετοί γνωστοί αλγόριθμοι.

Η μεγιστοποίηση του περιθωρίου ισοδυναμεί με το πρόβλημα ελαχιστοποίησης:
$\min_{w,b} \frac{1}{2}\|w\|_2 - \sum_{i=1}^{N} a_i y_i \left(x_i^T \cdot w + b\right) + \sum_{i=1}^{N} a_i$, όπου a_i οι πολλαπλασιαστές Lagrnage για κάθε δεδομένο του συνόλου εκπαίδευσης.
Οι άγνωστοι παράμετροι του SVM υπολογίζονται μόνο με τα S σε πλήθος support vectors: $w = \sum_{j=1}^{S} a_{t_j} y_{t_j} x_j$

Ταξινόμηση με SVM

Όταν έρθει ένα νέο δείγμα z (εκτός του συνόλου εκπαίδευσης) αυτό ταξινομείται στην κατηγορία 1 εάν $f(z) > 0$ και στην κατηγορία -1 εάν $f(z) < 0$.

Η συνάρτηση απόφασης είναι ένα linear discriminant: $f = w^T z + b$
Υπολογίζεται χρησιμοποιώντας μόνο τα support vectors: $f = w^T z + b = \sum_{j=1}^{S} a_{t_j} y_{t_j} w_{t_j}^T z + b$

SVMs και δεδομένα με θόρυβο

Τα δεδομένα συχνά περιέχουν θόρυβο ή/και ακραίες τιμές, γεγονός που μπορεί να οδηγήσει σε μη-ακριβή εκτίμηση των παραμέτρων. Για να αντιμετωπιστεί αυτό το πρόβλημα μεγιστοποιούμε το λεγόμενο soft-margin.

Επιτρέπουμε μερικά δεδομένα να πέσουν μέσα στο περιθώριο, αλλά τα τιμωρούμε. Αυτό επιτυγχάνεται με την εισαγωγή μεταβλητών χαλάρωσης (slack variables), μία για κάθε δεδομένο του συνόλου εκπαίδευσης.

Οι slack variables ξ εκφράζουν την απόσταση των δεδομένων τα οποία δεν ταξινομούνται σωστά από το υπερεπίπεδο μέγιστου περιθωρίου, το οποίο είναι γνωστό ως hard margin.

Θέλουμε να μεγιστοποιήσουμε το περιθώριο και ταυτόχρονα να ελαχιστοποιήσουμε τις αποστάσεις ξ_i

$\min \frac{1}{2}\|w\|_2 + C \cdot \sum_i \xi_i \ \ s.t. y_i(w \cdot x_i + b) \geq 1 - \xi_i, \forall x_i \ and \ \xi_i \geq 0$

Το υπερεπίπεδο που προκύπτει ως λύση του παραπάνω προβλήματος βελτιστοποίησης. Το C αποτελεί υπερπαράμετρο. Όταν τείνει στο άπειρο το soft margin τείνει στο hard margin. ονομάζεται soft margin.

Στη πράξη βελτιστοποιούμε το dual πρόβλημα:

$\max_a \sum_{i=1}^N a_i - \frac{1}{2}\sum_{i=1,j=1}^N a_i a_j y_i y_j x_i^T x_j \ \ s.t. C \geq a_i \geq 0, \sum_{i=1}^N a_i y_i = 0$

Το ενδιαφέρον στη παραπάνω διατύπωση είναι ότι εμφανίζονται εσωτερικά γινόμενα των δεδομένων γεγονός χρήσιμο για την μη-γραμμική επέκταση των svms.

Μη-γραμμικά διαχωρίσιμες κλάσεις

Γενική ιδέα: Μπορούμε να απεικονίσουμε τα δεδομένα σε ένα χώρο χαρακτηριστικών (feature space) μεγαλύτερης διάστασης όπου τα δεδομένα εκπαίδευσης είναι γραμμικά διαχωρίσιμα.
Εάν μετασχηματίσουμε μη-γραμμικά τα δεδομένα σε νέο χώρο χαρακτηριστικών το παραπάνω πρόβλημα τροποποιείται ελάχιστα ως $\max_a \sum_{i=1}^N a_i - \frac{1}{2}\sum_{i=1,j=1}^N a_i a_j y_i y_j \varphi(x_i)^T \varphi(x_j)$

Kernel functions

Μια συνάρτηση πυρήνα (kernel function) ορίζεται ως $K(x_1, x_2) =< \Phi(x_1), \Phi(x_2) >= \varphi(x_1)^T \varphi(x_2)$
Περιγράφουν την ομοιότητα των δεδομένων και είναι εξαιρετικά χρήσιμοι στην μηχανική μάθηση

$\max_a \sum_{i=1}^N a_i - \frac{1}{2}\sum_{i=1,j=1}^N a_i a_j y_i y_j K(x_i, x_j)$

Χρησιμοποιώντας συναρτήσεις πυρήνα δεν χρειάζεται να υπολογίσουμε ξεχωριστά τις συναρτήσεις φ και συνεπώς έχουμε καλύτερη πολυπλοκότητα.

Παραδείγματα Συναρτήσεων Πυρήνα

Polynomial kernel with degree d: $K(x, y) = (x^T y + 1)^d$

Radical Basis Function kernel with width σ: $K(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$

Sigmoid with parameters κ and θ: $K(x, y) = \tanh(\kappa x^t y + \theta)$

Classifying with a Kernel

$w = \sum_{j=1}^S a_{t_j} y_{t_j} \varphi\left(x_{t_j}\right)$

$f =< w^T \varphi(z) > + b = \sum_{j=1}^S a_{t_j} y_{t_j} K\left(x_{t_j}, z\right) + b$

L10 – Βαθιά Μάθηση (Deep Learning)

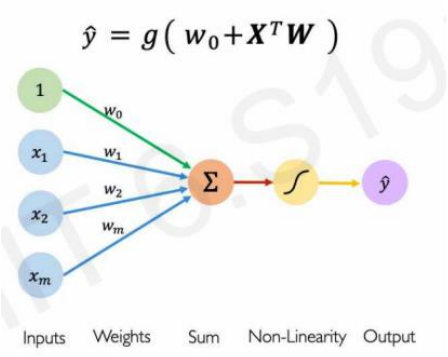
Τι είναι το deep learning

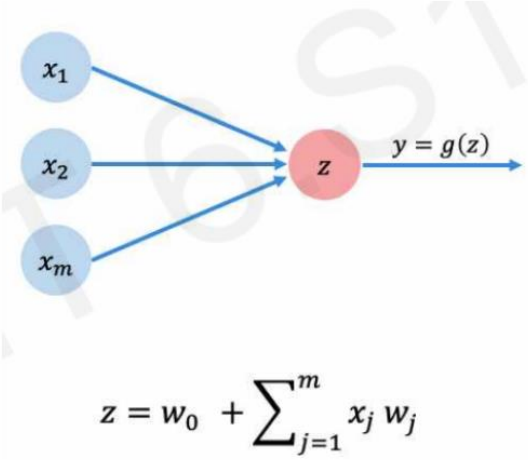
Τα δεδομένα στις εφαρμογές επιστήμης και τεχνολογίας είναι μη-γραμμικά.

Οι κλασικές μέθοδοι μηχανικής μάθησης αντιμετωπίζουν το πρόβλημα της μη-γραμμικότητας βρίσκοντας κατάλληλες αναπαραστάσεις των δεδομένα (πχ μέσω hand crafted features, τεχνικές μείωσης διαστάσεων, kernels) ή/και χρησιμοποιώντας μη-γραμμικές activation functions (πχ λογιστική παλινδρόμηση).

Με τα νευρωτικά δίκτυα βάθους (deep neural networks) μπορούμε να μάθουμε τις αναπαραστάσεις και τον ταξινομητή/regressor απευθείας από τα δεδομένα σε ένα βήμα.

Νευρώνας perceptron





Η g είναι η μη-γραμμική Activation Function και το w_0 το bias

$$\hat{y} = g\left(w_0 + \sum_{j=1}^{m} x_j w_j\right) = g(w_0 + X^T W), X = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}, W = \begin{bmatrix} w_1 \\ \vdots \\ w_m \end{bmatrix}$$

Παραδείγματα μη-γραμμικών Activation Function

Sigmoid: $\sigma(x) = \frac{1}{1+e^{-x}}$

tanh: $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{e^{2x}-1}{e^{2x}+1}$

ReLU: $\max\{0, x\}$

Leaky ReLU: $\max\{0.1 \cdot x, x\}$

Max-Out: $\max\{w_1^T x + b_1, w_2^T x + b_2\}$

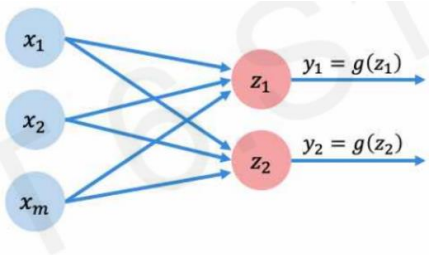
ELU: $\begin{cases} x & , x \geq 0 \\ \alpha \cdot (e^x - 1) & , x < 0 \end{cases}$

Perceptron: Forward pass

Forward pass είναι η ο υπολογισμός του $\hat{y}$ εφαρμόζοντας τον τύπο του. Δηλαδή βρίσκοντας το άθροισμα των weights με τα features και προσθέτοντας το bias και τέλος περνώντας το από την activation function.

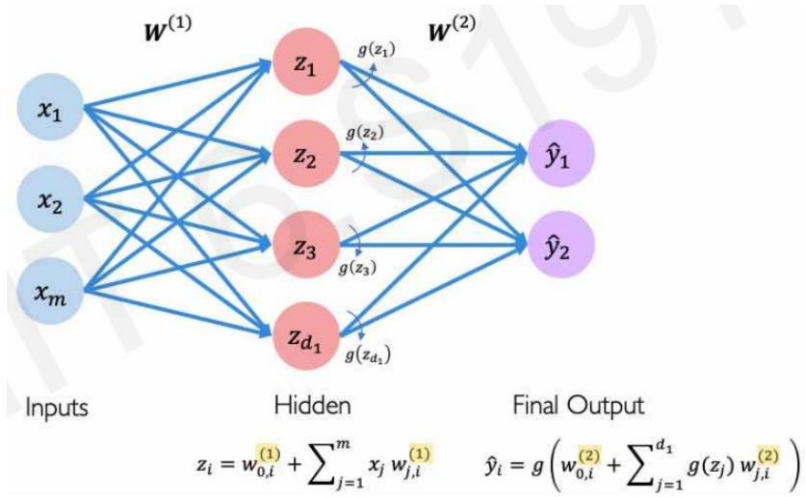
Perceptron με πολλαπλές εξόδους

Καθώς όλα τα στοιχεία εισόδου είναι συνδεδεμένα με όλα τα στοιχεία εξόδου του δικτύου, το layer ονομάζεται dense layer.



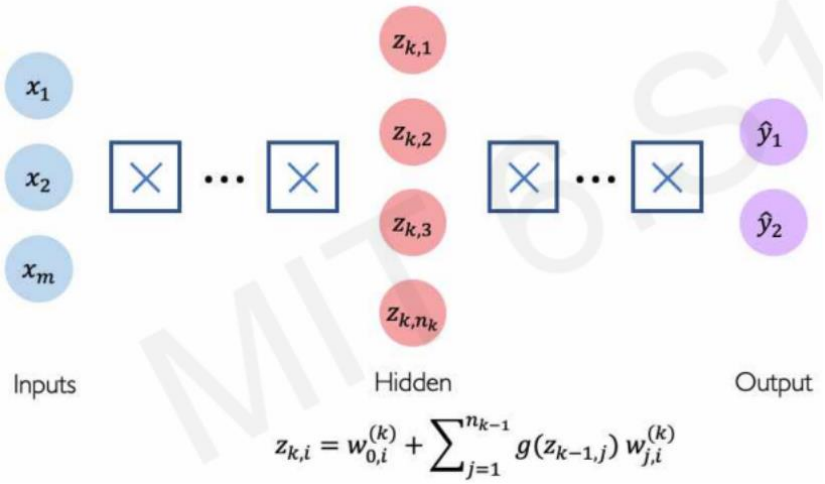
$$z_i = w_{0,i} + \sum_{j=1}^{m} x_j \, w_{j,i}$$

Δίκτυο ενός επιπέδου (shallow neural network)



$$z_i = w_{0,i}^{(1)} + \sum_{j=1}^{m} x_j \, w_{j,i}^{(1)} \qquad \hat{y}_i = g\left(w_{0,i}^{(2)} + \sum_{j=1}^{d_1} g(z_j) \, w_{j,i}^{(2)}\right)$$

Δίκτυο πολλών επιπέδων (deep neural network)



$$z_{k,i} = w_{0,i}^{(k)} + \sum_{j=1}^{n_{k-1}} g(z_{k-1,j}) \, w_{j,i}^{(k)}$$

Εκπαίδευση Νευρωνικών Δικτύων

Empirical Loss

$J(W) = \frac{1}{n}\sum_{i=1}^{n} L\left(f\left(x^{(i)}; W\right), y^{(i)}\right)$

Cross entropy για προβλήματα ταξινόμησης

$J(W) = -\frac{1}{n}\sum_{i=1}^{n} y^{(i)} \log f\left(x^{(i)}; W\right) + \left(1 - y^{(i)}\right) \log\left(1 - f\left(x^{(i)}; W\right)\right)$

Μέσω τετραγωνικό σφάλμα για προβλήματα παλινδρόμησης

$J(W) = \frac{1}{n}\sum_{i=1}^{n} \left(y^{(i)} - f\left(x^{(i)}; W\right)\right)^2$

Βελτιστοποίηση

Πρόβλημα βελτιστοποίησης: $W^* = \arg\min_W \frac{1}{n}\sum_{i=1}^{n} L\left(f\left(x^{(i)}; W\right), y^{(i)}\right) = \arg\min_W J(W)$

Gradient descent

Algorithm

1. Initialize weights randomly $\sim N(0, \sigma^2)$
2. Loop until convergence
3. Compute gradient: $\frac{\partial J(W)}{\partial W}$
4. Update weights: $W \leftarrow W - \eta \cdot \frac{\partial J(W)}{\partial W}$
5. Return weights

Stochastic Gradient Descent

Η παράγωγος $\frac{\partial J(W)}{\partial W}$ είναι υπολογιστικά απαιτητική για να υπολογιστεί.

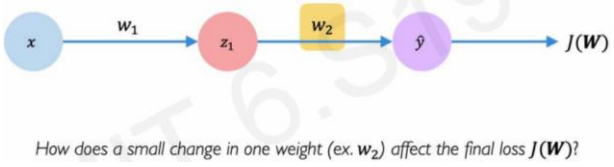
Μία λύση θα ήταν να υπολογίζαμε την $\frac{\partial J_i(W)}{\partial W}$, δηλαδή την παράγωγο σε ένα τυχαίο σημείο, που θα ήταν εύκολος ο υπολογισμός, ωστόσο είναι πολύ επιρρεπής στον θόρυβο.

Αντικαθιστώντας το $\frac{\partial J(W)}{\partial W}$ με $\frac{1}{B} \cdot \sum_{k=1}^{B} \frac{\partial J_k(W)}{\partial W}$ μπορούμε να έχουμε μία καλή εκτίμηση της πραγματικής παραγώγου που είναι πολύ ευκολότερος ο υπολογισμός της.

Algorithm

1. Initialize weights randomly $\sim N(0, \sigma^2)$
2. Loop until convergence
3. Pick batch of B data points
4. Compute gradient: $\frac{\partial J(W)}{\partial W} = \frac{1}{B} \cdot \sum_{k=1}^{B} \frac{\partial J_k(W)}{\partial W}$
5. Update weights: $W \leftarrow W - \eta \cdot \frac{\partial J(W)}{\partial W}$
6. Return weights

Υπολογισμός παραγώγων: Backpropagation



$\frac{\partial J(W)}{w_2} = \frac{\partial J(W)}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$

$\frac{\partial J(W)}{w_1} = \frac{\partial J(W)}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1} = \frac{\partial J(W)}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z_1} \cdot \frac{\partial z_1}{\partial w_1}$

Repeat this for every weight in the network using gradients from later layers

Επιλογή Learning Rate

Ιδέα 1: Δοκιμάζουμε πολλά διαφορετικά και επιλέγουμε αυτό που λειτουργεί “just right”
Ιδέα 2: Σχεδιάζουμε ένα προσαρμοστικό learning rate που προσαρμόζεται στην επιφάνεια

Προσαρμοστικά learning rates

Τα learning rates δεν είναι πλέον σταθερά. Μπορούν να μεταβληθούν ανάλογα με:

- Το μέγεθος της παραγώγου
- Το πόσο γρήγορα γίνεται η μάθηση
- Το μέγεθος συγκεκριμένων weights
- Και άλλα

Κανονικοποίηση

Dropout: Κατά την εκπαίδευση, τυχαία θέτουμε κάποια activations σε 0.
Συνήθως κάνουμε “drop” 50% των activations σε κάθε layer.
Αναγκάζει το δίκτυο να μην βασίζεται σε συγκεκριμένους κόμβους.

Early Stopping: Σταματάμε την εκπαίδευση πριν προλάβει να συμβεί overfitting.