ΠΟΛΥΤΕΧΝΕΙΟ
ΚΡΗΤΗΣ

$2^{η}$ σειρά ασκήσεων

Σωτήριος Μιχαήλ
2015030140

Στατιστική μοντελοποίηση
και
αναγνώριση προτύπων

Θέμα 1ο – Λογιστική παλινδρόμηση: αναλυτική εύρεση κλίσης (gradient)

α. Αναλυτική εύρεση της κλίσης της συνάρτησης κόστους

$$h_\theta(x) = f(\theta_\chi^\tau) = \frac{1}{1+e^{-\theta_\chi^\tau}}\ ,\ \theta=[\theta_1, \theta_2, \dots, \theta_n]^\tau$$

$$x=[x_1, x_2, \dots, x_n]^\tau$$

Υπολογίζουμε τη παράγωγο $\frac{d}{d\theta_j}(h_\theta(x))$, η οποία θα χρειαστεί παρακάτω:

$$\frac{d}{d\theta_j}(h_\theta(x)) =$$

$$\frac{d}{d\theta_j}\left(\frac{1}{1+e^{-\theta_\chi^\tau}}\right) =$$

$$-\frac{1}{(1+e^{-\theta_\chi^\tau})^2} \cdot \frac{d}{d\theta_j}(1+e^{-\theta_\chi^\tau}) =$$

$$-\frac{1}{(1+e^{-\theta_\chi^\tau})^2} \cdot \frac{d}{d\theta_j}(e^{-\theta_\chi^\tau}) =$$

$$-\frac{e^{-\theta_\chi^\tau}}{(1+e^{-\theta_\chi^\tau})^2} \cdot \frac{d}{d\theta_j}(-\theta_\chi^\tau) =$$

$$-\frac{e^{-\theta_\chi^\tau}}{(1+e^{-\theta_\chi^\tau})^2} \cdot \frac{d}{d\theta_j}\left(\sum_{k=1}^{n} \theta_k \cdot x_k\right) =$$

$$\frac{x_j \cdot e^{-\theta_\chi^\tau}}{(1+e^{-\theta_\chi^\tau})^2}$$

Υπολογίζουμε και τη ποσότητα $1-h_\theta(x)$:

$$1-h_\theta(x) = 1-\frac{1}{1+e^{-\theta_\chi^\tau}} = \frac{e^{-\theta_\chi^\tau}}{1+e^{-\theta_\chi^\tau}}$$

Για το ζητούμενο:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left(-y^{(i)}\ln(h_\theta(\chi^{(i)}))-(1-y^{(i)})\ln(1-h_\theta(\chi^{(1)}))\right)\Rightarrow$$

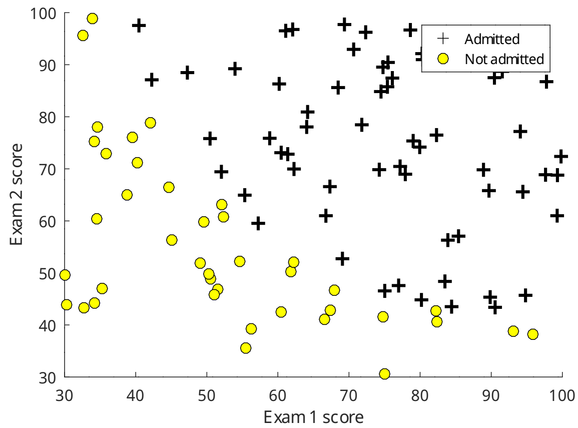
$$\frac{d}{d\theta_j}J(\theta) = \frac{1}{m}\sum_{i=1}^{m} -y^{(i)}\cdot\frac{1}{h_\theta(x^{(i)})}\cdot\frac{d}{d\theta_j}h_\theta(x^{(i)})+(1-y^{(i)})\cdot\frac{1}{1-h_\theta(x^{(i)})}\cdot\frac{d}{d\theta_j}h_\theta(x^{(i)})$$

Από τα παραπάνω:

$$= \frac{1}{m}\sum_{i=1}^{m} -y^{(i)}\cdot\frac{x_j^{(i)}\cdot e^{-\theta_\chi^\tau(i)}}{(1+e^{-\theta_\chi^\tau})} + \frac{(1-y^{(i)})x_j^{(i)}}{1+e^{\theta_\chi^\tau(i)}}$$

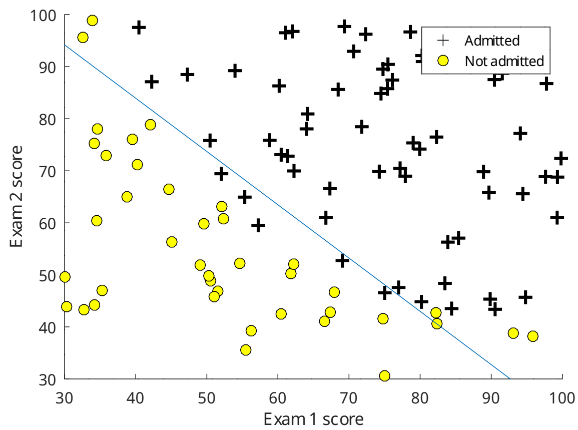
$$= \frac{1}{m}\sum_{i=1}^{n} x_j(h_\theta(x^{(i)})-y^{(i)})$$

β. Στο δεύτερο κομμάτι, εφαρμόζουμε τη λογιστική παλινδρόμηση σε ένα σετ δεδομένων με βαθμούς φοιτητών, με σκοπό τη πρόβλεψη της πιθανότητας να γίνουν δεκτοί σε ένα πανεπιστήμιο. Το σετ δεδομένων το οποίο επεξεργάσθηκε είναι το εξής:



Εικόνα 1: *Απεικόνιση σετ δεδομένων*

Ελαχιστοποιώντας τη συνάρτηση κόστους cross entropy, υπολογίζουμε τις παραμέτρους θ και έπειτα σχεδιάζουμε την ευθεία παλινδρόμησης που ορίζουν τα δεδομένα. Χρησιμοποιώντας την συνάρτηση σίγμα και το προαναφερθέν θ, προβλέπουμε εάν ένας φοιτητής θα γίνει αποδεκτός με βάση τους βαθμούς του.



Εικόνα 2: *Διάγραμμα ορίου απόφασης*

Το μοντέλο αυτό επιτυγχάνει ποσοστό σωστής ταξινόμησης 89%, κάτι το οποίο είναι αναμενόμενο, καθώς τα δεδομένα που βασίζεται η πρόβλεψη είναι γραμμικά διαχωρίσιμα. Το όριο απόφασης μας δείχνει με ικανοποιητική ακρίβεια τη γραμμή που διαχωρίζει τους αποδεκτούς και μη φοιτητές.

Θέμα 2ο – Λογιστική παλινδρόμηση με ομαλοποίηση

α. Αναλυτική εύρεση της κλίσης της συνάρτησης κόστους

Αν θ_j και $x_j^{(i)}$ είναι η j-οστή συνιστώσα των διανυσμάτων $\theta = [\theta_1, \theta_2, \ldots, \theta_n]^T$ και $x = [x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}]^T$, τότε:

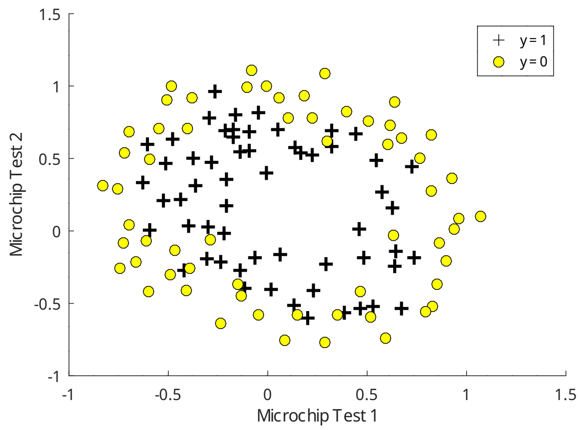
$$J(\theta) = \frac{1}{m}\sum_{i=1}^{k} -y^{(i)}\ln\left(h_\theta\left(x^{(i)}\right)\right) - \left(1-y^{(i)}\right)\ln\left(1-h_\theta\left(x^{(i)}\right)\right) + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2 \Rightarrow$$

$$\frac{dJ(\theta)}{d\theta_j} = \frac{1}{m}\frac{d}{d\theta_j}\left(\sum -y^{(i)}\ln\left(h_\theta\left(x^{(i)}\right)\right) - \left(1-y^{(i)}\right)\ln\left(1-h_\theta\left(x^{(i)}\right)\right)\right) + \frac{d}{d\theta_j}\left(\frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2\right) =$$

$$\frac{1}{m}\sum_{k=1}^{m} x_j^{(i)}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) + \frac{\lambda}{2m}\sum_{j=1}^{n}\frac{d}{d\theta_j}\theta_j^2 =$$

$$\frac{1}{m}\sum_{k=1}^{m} x_j^{(i)}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) + \frac{\lambda}{m}\theta_j$$

β. Εφαρμόζουμε τη λογιστική παλινδρόμηση με ομαλοποίηση σε ένα σετ δεδομένων με 2 τεστ από μικροτσίπ με σκοπό να πραγματοποιηθεί μία ανάλυση της ποιότητας των παραγόμενων τσιπ, δηλαδή της ορθής λειτουργίας τους.



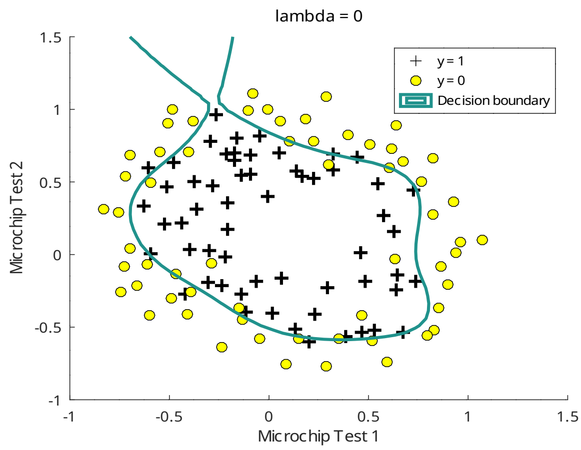
***Εικόνα 3:** Απεικόνιση σετ δεδομένων*

Επειδή τα δεδομένα αυτά δεν είναι γραμμικώς διαχωρίσιμα, θα χρειαστεί να προβληθούν σε χώρο μεγαλύτερων διαστάσεων έτσι ώστε να υπολογιστεί το όριο απόφασης. Αυτό θα επιτευχθεί με το να προβληθούν τα δεδομένα στον χώρο που ορίζουν όλοι οι όροι των πολυωνύμων x_1 και x_2 βαθμού μέχρι και 6. Οι σχέσεις που χρησιμοποιήθηκαν είναι:

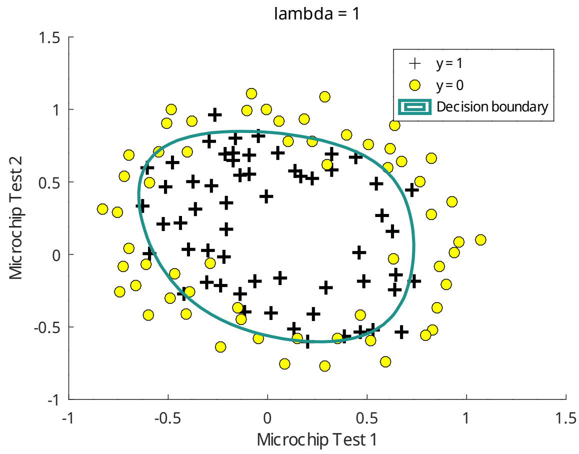
$$P(x_1, x_2) = \sum_{i=0}^{6}\sum_{j=0}^{i} x_1^{i-j}x_2^{j}$$

$$\mathit{mapfeature}(x) = \left[1, x_1, x_2, x_1^2, x_1x_2, x_2^2, x_1^3, \ldots, x_1x_2^5, x_2^6\right]^T$$

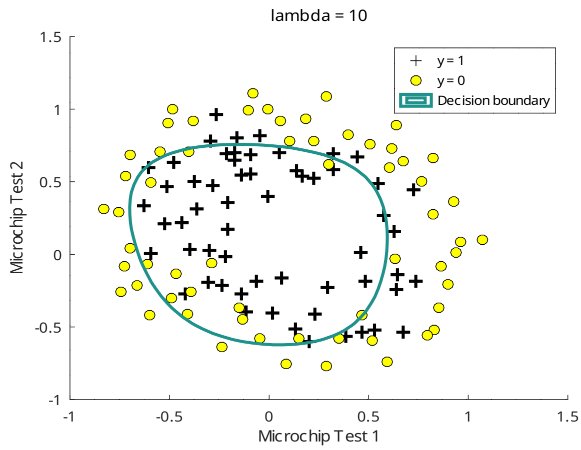
Με τη βελτιστοποίηση των παραγόντων *θ*, σχεδιάζονται τα όρια απόφασης για λ = 0, 1, 10 & 100:



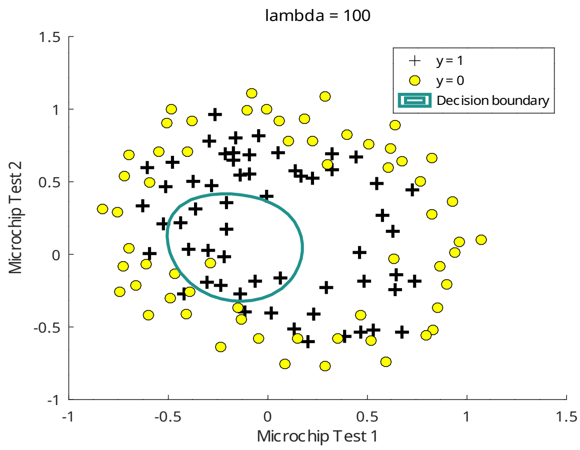
Εικόνα 4: *Ακρίβεια = 87.2%*



Εικόνα 5: *Ακρίβεια = 82.2%*



Εικόνα 6: *Ακρίβεια = 74.5%*



Εικόνα 7: *Ακρίβεια = 60.1%*

Παρατηρούμε πως για λ = 0, το μοντέλο πετυχαίνει το μεγαλύτερο ποσοστό ακρίβειας για το συγκεκριμένο σετ δεδομένων, ωστόσο είναι εμφανές πως έχει γίνει overfitting για τα συγκεκριμένα δεδομένα, με αποτέλεσμα τη κακή συμπεριφορά του μοντέλου σε γενικότερες περιπτώσεις, δηλαδή για δεδομένα εκτός του συγκεκριμένου σετ. Απεναντίας, για λ = 100, παρατηρείται το φαινόμενο του underfitting, με χαμηλότερο ποσοστό ακρίβειας για το συγκεκριμένο σετ δεδομένων, και καμία πιθανότητα γενίκευσης. Για λ = 1 και λ = 10, το όριο απόφασης είναι ικανοποιητικά γενικό με καλή ακρίβεια για τα συγκεκριμένα δεδομένα, που όμως μπορεί να γενικευθεί για δεδομένα εκτός του σετ αυτού.

Θέμα 3ο – Εκτίμηση παραμέτρων (maximum likelihood)

$$D=\{x_1, x_2, \ldots, x_n\}, \text{ανεξάρτητα από } p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Θέλω να εκτιμήσω το λ με MLE

$$L_D(\lambda) = P(D|\lambda) = \prod_{k=1}^{n} p(x_k|\lambda)$$

$$\lambda_{MLE} = \mathbf{argmax}_\lambda P(D|\lambda)$$

$$l_d(\lambda) = lnL_D(\lambda) = \sum_{k=1}^{n} \ln(p(x_k|\lambda))$$

$$\frac{d}{d\lambda} l_d(\lambda) = 0 \Rightarrow$$

$$\frac{d}{d\lambda}\left(\sum_{k=1}^{n} \ln(p(x_k|\lambda))\right) = 0 \Rightarrow$$

$$\sum_{k=1}^{n} \frac{d}{d\lambda}\left(\ln \frac{\lambda_k^x \cdot e^{-\lambda}}{x!}\right) = 0 \Rightarrow$$

$$\sum_{k=1}^{n} \frac{x!}{\lambda_k^x \cdot e^{-\lambda}} \frac{d}{d\lambda}\left(\frac{\lambda_k^x \cdot e^{-\lambda}}{x!}\right) = 0 \Rightarrow$$

$$\sum_{k=1}^{n} \frac{\cancel{x!}}{\lambda_k^x \cdot e^{-\lambda}} \cdot \frac{x_k \lambda^{x_k - 1} - \lambda_k^x \cdot e^{-\lambda}}{\cancel{x!}} = 0 \Rightarrow$$

$$\sum_{k=}^{n} \cancel{\lambda_k^x \cdot e^{-\lambda}} \frac{\left(\frac{x_k}{\lambda} - 1\right)}{\cancel{\lambda_k^x \cdot e^{-\lambda}}} = 0 \Rightarrow$$

$$\sum_{k=1}^{n} \frac{x_k}{\lambda} - n = 0 \Rightarrow$$

$$\frac{1}{\lambda} \sum_{k=1}^{n} x_k = n \Rightarrow$$

$$\lambda_{MLE} = \frac{1}{n} \sum_{k=1}^{n} x_k$$

Θέμα 4ο – Support vector machines (αναλυτική βελτιστοποίηση με KKT)

$$l(\omega, \omega_0, \lambda) = \frac{1}{2}(\omega_1^2+\omega_2^2) - \sum \lambda_i [y_i(\omega^T x_i + \omega_0) - 1] =$$

$$\frac{(\omega_1^2+\omega_2^2)}{2} - \lambda_1(-(-10\,\omega_1 - 10\,\omega_2 + \omega_0) - 1) - \lambda_2(-(-7\,\omega_1 - 7\,\omega_2 + \omega_0) - 1)$$

$$Από\ \frac{d}{d\omega} l(\omega, \omega_0, \lambda),\ έχουμε:\ \frac{d}{d\omega_1} l = 0 \Rightarrow \omega_1 = 10\lambda_1 - 7\lambda_2$$

$$\frac{d}{d\omega_2} l = 0 \Rightarrow \omega_2 = 10\lambda_1 - 7\lambda_2$$

$$Άρα:\ \omega_1 = \omega_2$$

$$\lambda_1\left(-\left((\omega_1, \omega_2)\begin{pmatrix}-10\\-10\end{pmatrix} + \omega_0\right) - 1\right) = 0$$

$$\lambda_2\left((\omega_1, \omega_2)\begin{pmatrix}-7\\-7\end{pmatrix} + \omega_0 - 1\right) = 0$$

$$Από\ τα\ παραπάνω\ ,\ προκύπτουν:$$

$$10\,\omega_1 + 10\,\omega_2 - \omega_0 - 1 = 0\ \ και$$

$$-7\,\omega_1 - 7\,\omega_2 + \omega_0 - 1 = 0$$

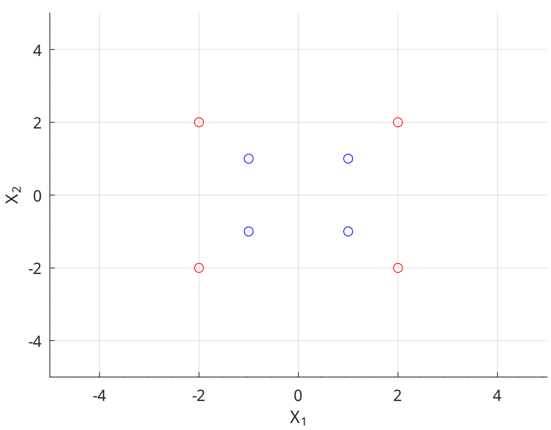
$$Καθώς\ \omega_1 = \omega_2,\ τότε\ ,\ από\ τις\ παραπάνω\ σχέσεις:\ \omega_1 = \omega_2 = \frac{1}{3}$$

$$Άρα\ \omega_0 = \frac{17}{3}$$

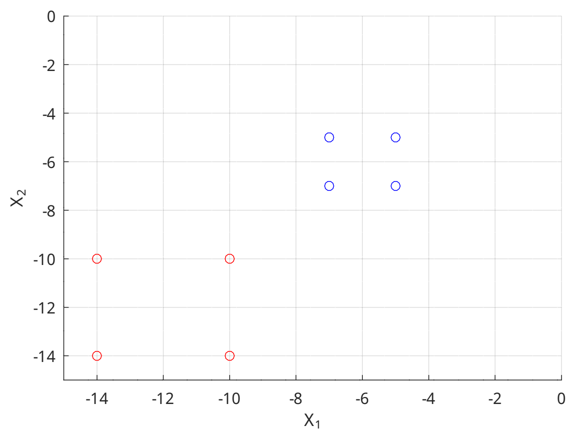
$$\omega_x^T + \omega_0 = 0 \Rightarrow \left(\frac{1}{3}\ \frac{1}{3}\right)\begin{pmatrix}x_1\\x_2\end{pmatrix} + \frac{17}{3} = 0 \Rightarrow$$

$$\frac{x_1}{3} + \frac{x_2}{3} + \frac{17}{3} = 0 \Rightarrow x_1 = -x_2 - 17$$

Τα δείγματα πριν και μετά τον μετασχηματισμό (μέσω MATLAB):



Εικόνα 8: *Πριν τον μετασχηματισμό*



Εικόνα 9: *Μετά τον μετασχηματισμό*

Πριν τον μετασχηματισμό, δεν είναι δυνατή η διαισθητική προσέγγιση ενός γραμμικού επιπέδου διαχωρισμού των δεδομένων (όπως θα ήταν δυνατό, για παράδειγμα, σε γραμμικά διαχωρίσιμα δεδομένα). Μετά τον μετασχηματισμό, το βέλτιστο υπερεπίπεδο διαχωρισμού είναι εμφανές.

Θέμα 5ο – Υλοποίηση ενός απλού νευρωνικού δικτύου

α.

$$\textit{Έχουμε}\ \ J\left(y^{(i)},\ \hat{y}^{(i)};\ w,\ b\right)=-y^{(i)}\ln\hat{y}^{(i)}-\left(1-y^{(i)}\right)\ln\left(1-\hat{y}^{(i)}\right)\ (1)$$

$$\textit{και}\ \ J\left(Y,\ \hat{Y};W,\ b\right)=\frac{1}{B}\sum_{i}\left(-y^{(i)}\ln\hat{y}^{(i)}-\left(1-y^{(i)}\right)\ln\left(1-\hat{y}^{(i)}\right)\right)\ (2)$$

$$\textit{Χρησιμοποιώντας τη σχέση}\ (2),\ \textit{και θέτοντας}\ z^{(i)}=x^{(i)}W+b\ \textit{και}\ \hat{y}^{(i)}=f\left(z^{(i)}\right),\ \textit{παίρνουμε}:$$

$$J\left(Y,\ \hat{Y};W,\ b\right)=\frac{1}{B}\sum_{i}-y^{(i)}\ln\frac{1}{1+e^{\left(-z^{(i)}\right)}}-\left(1-y^{(i)}\right)\ln\frac{e^{\left(-z^{(i)}\right)}}{1+e^{\left(-z^{(i)}\right)}}=$$

$$\frac{1}{B}\sum_{i}\cancel{y^{(i)}\ln\left(1+e^{\left(-z^{(i)}\right)}\right)}-\ln e^{\left(-z^{(i)}\right)}+\ln\left(1+e^{\left(-z^{(i)}\right)}\right)+y^{(i)}\ln e^{\left(-z^{(i)}\right)}-\cancel{y^{(i)}\ln\left(1+e^{\left(-z^{(i)}\right)}\right)}=$$

$$\frac{1}{B}\sum_{i}z^{(i)}-y^{(i)}z^{(i)}+\ln\left(1+e^{\left(-z^{(i)}\right)}\right)$$

β.

$$\textit{Θα παραγωγίσουμε τη σχέση}\ (1),\ \textit{αφού πρώτα γραφεί συναρτήσει του}\ z^{(i)}$$

$$(1)\Rightarrow J\left(y^{(i)},\ \hat{y}^{(i)};\ w,\ b\right)=-y^{(i)}\ln\hat{y}^{(i)}-\left(1-y^{(i)}\right)\ln\left(1-\hat{y}^{(i)}\right)=$$

$$-y^{(i)}\ln\left(f\left(z^{(i)}\right)\right)-\left(1-y^{(i)}\right)\ln\left(1-f\left(z^{(i)}\right)\right)=$$

$$y^{(i)}\ln\frac{1}{1+e^{\left(-z^{(i)}\right)}}-\left(1-y^{(i)}\right)\ln\frac{e^{\left(-z^{(i)}\right)}}{1+e^{\left(-z^{(i)}\right)}}=$$

$$-y^{(i)}\ln\left(1+e^{\left(-z^{(i)}\right)}\right)-\left(1-y^{(i)}\right)\left(\ln\left(e^{\left(-z^{(i)}\right)}\right)-\ln\left(1+e^{\left(-z^{(i)}\right)}\right)\right)=$$

$$z^{(i)}-z^{(i)}y^{(i)}+\ln\left(1+e^{\left(-z^{(i)}\right)}\right)$$

$$\textit{Άρα}\ \ \frac{dJ\left(y^{(i)},\ \hat{y}^{(i)};\ w,\ b\right)}{dz^{(i)}}=\frac{d}{dz^{(i)}}\left(z^{(i)}-z^{(i)}y^{(i)}+\ln\left(1+e^{\left(-z^{(i)}\right)}\right)\right)=$$

$$1-y^{(i)}+\frac{\left(-e^{\left(-z^{(i)}\right)}\right)}{1+e^{\left(-z^{(i)}\right)}}=$$

$$\hat{y}^{(i)}-y^{(i)}$$

γ.

$$Για\ \frac{dJ}{dw}:\ \frac{dJ\left(y^{(i)},\ \hat{y}^{(i)};\ w,\ b\right)}{dw}=\frac{dJ\left(y^{(i)},\ \hat{y}^{(i)};\ w,\ b\right)}{dz^{(i)}}\cdot\frac{dz^{(i)}}{dw}=\left(\hat{y}^{(i)}-y^{(i)}\right)x^{(i)}$$

$$J\left(Y,\ \hat{Y};W,\ b\right)=\frac{1}{B}\sum_{i}J\left(y^{(i)},\ \hat{y}^{(i)};\ w,\ b\right)$$

$$\frac{dJ\left(Y,\ \hat{Y};W,\ b\right)}{dW}=\frac{1}{B}\sum_{i}\frac{dJ\left(y^{(i)},\ \hat{y}^{(i)};\ w,\ b\right)}{dw}=$$

$$\frac{1}{B}\sum_{i}\left(\hat{y}^{(i)}-y^{(i)}\right)x^{(i)}$$

$$Για\ \frac{dJ}{db}:\ \frac{dJ\left(y^{(i)},\ \hat{y}^{(i)};\ w,\ b\right)}{db}=\frac{dJ\left(y^{(i)},\ \hat{y}^{(i)};\ w,\ b\right)}{dz^{(i)}}\cdot\frac{dz^{(i)}}{db}=\hat{y}^{(i)}-y^{(i)}$$

$$J\left(Y,\ \hat{Y};W,\ b\right)=\frac{1}{B}\sum_{i}J\left(y^{(i)},\ \hat{y}^{(i)};\ w,\ b\right)$$

$$\frac{dJ\left(Y,\ \hat{Y};W,\ b\right)}{db}=\frac{1}{B}\sum_{i}\frac{dJ\left(y^{(i)},\ \hat{y}^{(i)};\ w,\ b\right)}{db}=$$

$$\frac{1}{B}\left(\hat{y}^{(i)}-y^{(i)}\right)$$