Exercise 1

(a) Η maximum likelihood συνάρτηση είναι:

$$P(x_1,\dots,x_n;\theta) = \prod_{i=1}^{N}\left(P(\omega_1|x_i)^{y_i}\cdot P(\omega_2|x_i)^{1-y_i}\right)$$

$$= \prod_{n=1}^{N}\left(\sigma(\theta^T\cdot x_i)^{y_i}\cdot(1-\sigma(\theta^T\cdot x_i)^{1-y_i}\right)$$

Γιατί όπου: $P(\omega_1|x_i) = \sigma(\theta^T\cdot x_i)$ και $P(\omega_1|x_i)+P(\omega_2|x_i)=1$

Όπου $\sigma(\theta^T\cdot x) = \dfrac{1}{1+e^{-\theta^T\cdot x}}$

Και επειδή θέλουμε να κάνουμε minimize την Αρνητική log-likelihood:

$$\ln\left(L(\theta)\right) = -\sum_{i=1}^{N} y_i \ln\left(\sigma(\theta^T x) + (1-y_n)\cdot\ln\left(1-\sigma(\theta^T\cdot x)\right)\right)$$

Άρα $\dfrac{\partial \ln(L(\theta))}{\partial\theta} =$

$$-\sum_{i=1}^{N} y_i \frac{1}{\sigma(\theta^T x)}\left(\sigma(\theta^T x)\right)' + (1-y_i)\cdot\frac{1}{1-\sigma(\theta^T\cdot x)}\cdot\left((1-\sigma(\theta^T x)\right)'$$

$$= -\sum_{i=1}^{N}\left(y_i\frac{1}{\sigma(\theta^T x} + (1-y_i)\frac{1}{1-\sigma(\theta^T x)}\right)\left(\sigma(\theta^T x)\right)'$$

$$\begin{bmatrix}\frac{\partial\theta_1}{\partial\theta_i}\\ \vdots\\ \frac{\partial\theta_n}{\partial\theta_n}\end{bmatrix} = \begin{bmatrix}1\\ \vdots\\ 1\end{bmatrix}$$

$$= -\sum_{i=1}^{N}\left(y_i\frac{1}{\sigma(\theta^T x)} + (1-y_i)\frac{1}{1-\sigma(\theta^T x)}\right)\left(\sigma(\theta^T x)\left(1-\sigma(\theta^T x)\right)\frac{\partial\theta^T x}{\partial\theta}\right.$$

$$= -\sum_{i=1}^{N}\left(y\left(1-\sigma(\theta^T x) - (1-y_i)\cdot\sigma(\theta^T\cdot x)\right)X\right.$$

$$= \sum_{i=1}^{N}(S_i - y_i)\cdot x_i = X^T(S-y)$$

↘

Άρα Gradient Descent:

$$\theta^i = \theta^{i-1} - \mu_i \nabla L(\theta)|_{\theta=\theta^{i-1}} = \theta^{i-1} - \mu_i X^T (S - y)$$

Exercise 2

(a) .ipynb

(b) Θα χρησιμοποιήσουμε gradient descent για να κανουμε minimise την cost function:

$$J(\theta) = \sum_{i=1}^{N} (y_i - f(\theta^T x_i))^2$$

$$\boxed{f'(z) = af(z)(1 - f(z))}$$

Επίσης μπορούμε να πολλαπλασιάσουμε την cost function με το $\times \frac{1}{2}$ καθώς δεν θα επηρεαστεί το σημείο ελαχίστου που ψάχνουμε. (Για να φύγουν τα δύαρια)

Άρα $J(\theta) = \frac{1}{2} \sum_{i=1}^{N} (y_i - f(\theta^T x_i))^2$

$$\frac{\partial J(\theta)}{\partial \theta} = \sum_{i=1}^{N} (y_i - f(\theta^T x_i)) \cdot (-f(\theta^T x_i))'$$

$$= -\sum_{i=1}^{N} (y_i - f(\theta^T x_i)) \cdot f(\theta^T x_i) \cdot (1 - f(\theta^T x_i)) \cdot \frac{\partial \theta^T x_i}{\theta}$$

$$= -\sum_{i=1}^{N} \left((y_i - f(\theta^T x_i)) \cdot af(\theta^T x_i) \cdot (1 - f(\theta^T x_i)) \cdot x_i \right.$$

$$= -a(y - f) \cdot f \cdot (1 - f) \cdot X^T$$

όπου a = ελεύθερη παράμετρος

$$y = [y_1 \dots y_n]^T$$
$$X^T = [x_1 \dots x_n]^T$$
$$f = [f_1, \dots, f_n]$$

Άρα το gradient descent σχήμα που θα κάνει minimize την cost function θα είναι:

$$\theta^i = \theta^{i-1} - \mu_i \nabla J(\theta)\big|_{\theta=\theta^{i-1}} = \theta^{i-1} - \mu_i \cdot (-a)(y-f) \cdot f \cdot (1-f) \cdot X^T$$

(c) Το μοντέλο είναι αδύνατο να επιστρέψει ακριβώς τις τιμές 0 και 1 γιατί παίρνει αυτές τις τιμές στο $-+\infty$.
Συνεπώς για να επιστρέψει το 0 θα πρέπει $\theta^T x = -\infty$ κάτι το οποίο είναι πρακτικά αδύνατο.

$$\lim_{\theta x \to +\infty} \sigma(\theta x) = 1 \quad \text{και} \quad \lim_{\theta \cdot x \to -\infty} \sigma(\theta^T x) = 0$$

(d) Το αποτέλεσμα εκφράζει την πιθανότητα ενός συγκεκριμένου αποτελέσματος. (π.χ. να ανήκει στη κλάση 1)

(e) Όπως παρατηρούμε και από τα plots που δημιουργήσαμε στο ερώτημα (a), αλλάζοντας το a αλλάζει και η κλίση της συνάρτησης. Μικραίνοντας το a βλέπουμε ότι η σιγμοειδές συνάρτηση γίνεται πιο "smooth" και το αντίθετο όταν το αυξάνουμε. Για ένα πάρα πολύ μεγάλο a, η συνάρτηση θα έδινε τιμές 0 και 1.