Exercise 1

1/ given $Pr(Y=k|X)$: $P(D, \mu_i) = \prod \frac{\mu_i^{y_i}}{y_i!} \exp(-\mu_i)$

$$\log P(D, \mu_i) = \sum_{i=1}^{n} \left(y_i \log \mu_i - \mu_i - \log(y_i!) \right)$$

2/ let $\mu_i = \exp(w^T x_i + b)$

$$\log P(D, w) = \sum_{i=1}^{n} \left(y_i (w^T x_i + b) - \exp(w^T x_i + b) - \log(y_i!) \right)$$

if $w^T x_i + b > 0$, then we have $\exp(w^T x_i + b) > 1 \Rightarrow \mu_i > 1$

if $w^T x_i + b < 0$, then we have $\exp(w^T x_i + b) < 1 \Rightarrow \mu_i < 1$

3/ $$\langle \hat{w}, \hat{b} \rangle = \arg\max_{w,b} \sum_{i=1}^{n} \left(y_i (w^T x_i + b) - \exp(w^T x_i + b) - \log(y_i!) \right)$$

optimization variables: w and b, we are trying to maximize

4/ For simplicity we will have $w = \begin{bmatrix} w \\ b \end{bmatrix}$ $x_i = \begin{bmatrix} x_i \\ 1 \end{bmatrix}$ so we can use only ∇w like how lectures have been doing before

ie $\langle w, x_i \rangle = \begin{bmatrix} w \\ b \end{bmatrix}^T \begin{bmatrix} x_i \\ 1 \end{bmatrix}$

Thus we take gradient over w

$$\frac{\partial l(D, w)}{\partial w} = \sum_{i=1}^{n} \left(y_i x_i - x_i \exp(\langle w, x_i \rangle) \right)$$

$$= \sum_{i=1}^{n} x_i (y_i - \mu_i)$$ This is very similar to the one in lecture

Gradient over b

$$\frac{\partial l(D, w)}{\partial b} = \sum_{i=1}^{n} \left(y_i - \exp(\langle w, x_i \rangle) \right)$$

$$= \sum_{i=1}^{n} (y_i - \mu_i)$$

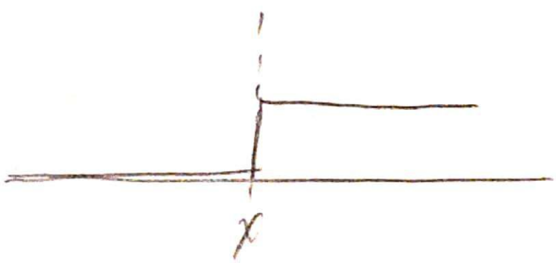
Thus we claim that is has no closed form for $\hat{w}$

We can also use the same gradient descent algorithm used in Lecture with updating w and b as follows.

$$w_t \leftarrow w_{t-1} - \eta \sum_{i=1}^{n} x_i y_i - x_i \exp(w_{t-1}^T x_i + b_{t-1}); \quad b_t \leftarrow b_{t-1} - \eta \sum_{i=1}^{n} y_i - \exp(w_{t-1}^T x_i + b_{t-1})$$

Exercise 2

1/ Logit fails to work with the given dataset as it responds with "Perfect separation detected, results not available". This happens because there exists a value "x" that perfectly predicts y. That means there is a value x, where depending on what the value of the input is, it can predict it with certainty. The graph looks like the following in 1D



Since logistic regression wants to find a model that fits this, it is impossible for it to regress to this ideal function and converge as it will continually get closer but will never converge to the target function.

2/

Exercise 3

1/ $\frac{1}{2}\|w\|_2^2 = \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$

$\sum_i C\gamma + \alpha_i(|y_i - (w^T x + b)| - \epsilon - \gamma_i) - \cancel{\beta_i \gamma_i}$

$= \sum_i \alpha_i |y_i - (w^T x + b)| - \alpha_i \epsilon$

$$\min_{C \geq \alpha \geq 0} \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum_i \alpha_i |y_i - (w^T x + b)| - \alpha_i \epsilon$$

2/ $\frac{\partial}{\partial w} = \frac{-x(y - (w^T x + b))}{|y - (w^T x + b)|}$

if $y - (w^T x + b) \geq 0$ ← arbitrarily chosen at 0

$\frac{\partial}{\partial w} = -x$

if $y - (w^T x + b) < 0$

$\frac{\partial}{\partial w} = x$

$\nabla_w l(x, y) = C \sum_{i=1}^{n} \frac{\partial}{\partial w}$

$\frac{\partial}{\partial b} = \frac{-(y - (w^T x + b))}{|y - (w^T x + b)|}$

if $y - (w^T x + b) \geq 0$

$\frac{\partial}{\partial b} = -1$

if $y - (w^T x + b) < 0$

$\frac{\partial}{\partial w} = 1$

$\nabla_b l(x, y) = C \sum_{i=1}^{n} \frac{\partial}{\partial b}$

3/ $\frac{1}{2\eta}\|z - w\|_2^2 + \frac{1}{2}\|z\|_2^2$

$= \frac{1}{2\eta}(z - w)^T(z - w) + \frac{1}{2} z^T z$

$= \frac{1}{2\eta}(z^T z - 2 z^T w - w^T w) + \frac{1}{2} z^T z$

$= \frac{1 - \eta z^T z}{2\eta} - \frac{1}{\eta} z^T w + \frac{1}{2\eta} w^T w$

Take gradient, set to 0

$\nabla_z \frac{1}{2\eta}\|z - w\|_2^2 + \frac{1}{2}\|z\|_2^2 = 0$

$\Rightarrow \frac{1+\eta}{\eta} z - \frac{1}{\eta} w = 0$

$\Rightarrow z = \frac{1}{1+\eta} w$

4/ training error: 610.2629, training loss: 610.7878, test error: 776.5091 a122rgug 20126636

Exercise 4

1/ Let $k(x,y) = \exp(-\alpha(x-y)^2)$

$$= e^{-\alpha x^2 + 2\alpha xy - \alpha y^2}$$

$$= e^{-\alpha(x^2+y^2)} \left(\sum_{k=0}^{\infty} \frac{(2\alpha xy)^k}{k!} \right)$$

$$= e^{-\alpha(x^2+y^2)} \left(\sum_{k=0}^{\infty} \frac{(2\alpha)^{\frac{k}{2}}}{k!} x^k \cdot \frac{(2\alpha)^{\frac{k}{2}}}{k!} y^k \right)$$

$$= e^{-\alpha x^2} \left(\sum_{k=0}^{\infty} \frac{(2\alpha)^{\frac{k}{2}}}{k!} x^k \right) \cdot e^{-\alpha y^2} \left(\sum_{k=0}^{\infty} \frac{(2\alpha)^{\frac{k}{2}}}{k!} y^k \right)$$

$$= \phi(x)^T \phi(y)$$

$$\phi(x) = e^{-\alpha x^2} \left[1, \frac{\sqrt{2\alpha}}{1!} x, \frac{\sqrt{(2\alpha)^2}}{2!} x^2, \dots \right]^T$$

It would be better to solve the dual representation as $\phi(x)$ would require an infinite computation time due to the infinite vector space of $\phi(x)$. By using the dual representation, $\langle \phi(x_i), \phi(x_j) \rangle$ is calculable in $O(d)$ using the given kernel function.

2/ The function is a valid kernel

$$k(x,y) = \frac{1}{1-xy}$$

$$= \sum_{k=0}^{\infty} (x.y)^k$$

$$= \sum_{k=0}^{\infty} x^k \sum_{k=0}^{\infty} y^k$$

$$\phi(x) = [1, x, x^2, \dots]^T$$