Sheet 02

Exercise 4: Logistic Regression

1)

$y \in \{0,1\}$, $p = p(y=1|x) = \sigma(w^T x) = \mu_1$ $\qquad y \sim Ber(p)$

$$\Rightarrow \quad p(y=k|x) = \mu_k^k (1-\mu_k)^{1-k} \overset{*}{=} \sigma(w^T x)^k \, \sigma(-w^T x)^{1-k}$$

$$* \; 1-\sigma(z) = \frac{\cancel{1}+e^{-z}-\cancel{1}}{1+e^{-z}} = \frac{1}{1+e^{z}} = \sigma(-z)$$

Log likelihood:

$$\ell(w) = \sum_{i=1}^{N} \log(p(y_i|x_i)) = \sum_{i=1}^{N} \log\left(\sigma(w^T x_i)^{y_i} \sigma(-w^T x_i)^{1-y_i}\right)$$
$$= \sum_{i=1}^{N} y_i \log(\sigma(w^T x_i)) + (1-y_i)\log(\sigma(-w^T x_i))$$

2) a)

Convexity of function ensures, that following the direction indicated by gradient descent you will find its global minimum, since any local minimum is the global minimum of a convex function.

To proof: $\quad -\ell(w) = -\sum_{i=1}^{N} k \log(\sigma(w^T x_i)) + (1-k)\log(\sigma(-w^T x_i))$ is convex

Convexity: $f: D \in \mathbb{R} \to \mathbb{R}$ convex iff $f(tx_1 + (1-t)x_2) \leq t f(x_1) + (1-t) f(x_2) \quad \forall x_1, x_2 \in D, \; 0 \leq t \leq 1$

Equivalently: f convex $\Leftrightarrow$ $f'' > 0 \quad \forall x \in D$

i) First consider $-\log(\sigma(x))$:

$$\partial_x - \log(\sigma(x)) = \partial_x \log(1+e^{-x}) = \frac{-e^{-x}}{1+e^{-x}} = -\sigma(x)$$

$$\partial_x^2 - \log(\sigma(x)) = \partial_x \sigma(x) = \sigma(x)\sigma(-x) > 0 \quad \forall x \qquad \Rightarrow -\log(\sigma(x)) \text{ is convex}$$

ii) Secondly consider: $-\log(\sigma(-x))$:

$$-\partial_x \log(\sigma(-x)) = \partial_x \log(1+e^{x}) = \frac{1}{1+e^{x}} e^{x} = \sigma(x)$$

$$\partial_x^2 - \log(\sigma(-x)) = \partial_x \sigma(x) = \sigma(x)\sigma(-x) > 0 \quad \forall x \qquad \Rightarrow -\log(\sigma(-x)) \text{ is convex}$$

Since $-\ell(w)$ is a sum of convex functions of the form i) and ii) it is convex itself. $\blacksquare$

b)

$$\mathcal{L}_{BCE}(w) = -N\,\ell(w)$$

$\Rightarrow$ maximizing log-likelihood $\Leftrightarrow$ minimizing average binary cross-entropy

3) a)

Suppose we have found w s.t. $w^T x = 0$ defines the decision boundary between two linearly separable classes.

That would lead to a loss:

$$\mathcal{L}(w) = -\ell(w) = -\sum_{i=1}^{N} y_i \log(\sigma(w^T x_i)) + (1-y_i)\log(\sigma(-w^T x_i))$$

where the arguments of all σ-functions contributing to it will be positive, ie. $\sigma \in (0{,}5, 1]$, because the model performs "correct" classifications.

The only way for the model to optimize, i.e. minimize the loss, is now to have not only correct classifications, but also "confident" classifications, meaning $\sigma(\pm w^T x_i) \sim 1$ and thus $\log(\sigma(\pm w^T x_i)) \sim 0$. For this we need: $\|w^T x_i\| \longrightarrow \infty$, which the model will try to ensure by taking $|w|$ to infinity.

b)

This can be mitigated implementing a so-called L2-regularization in the Loss:

$$\tilde{\mathcal{L}}(w) = \mathcal{L}(w) + \lambda \|w\|^2$$

which penalizes high w-magnitudes and prevents the divergence.