# Exercise Sheet 11

## 1 Activation Maximization

a) $\frac{\partial}{\partial x} \max_x \; w^T x + b - \lambda \|x\|^2 \overset{!}{=} 0$

$\Leftrightarrow \quad w^T - 2\lambda x \overset{!}{=} 0 \qquad \Leftrightarrow \quad x^* = \frac{w^T}{2\lambda}$

b) $\max_x \; w^T x + b + \log(p(x))$

$= \max_x \; w^T x + b + \log\left( \frac{\exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)}{(2\pi)^{\frac{d}{2}} |\Sigma|} \right)$

$= \max_x \; w^T x + b + \log\left(\exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)\right) - \log\left((2\pi)^{\frac{d}{2}} |\Sigma|\right)$

$\Leftrightarrow \quad \frac{\partial}{\partial x} w^T x - \frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \overset{!}{=} 0$

$\Leftrightarrow \quad w^T - \Sigma^{-1}(x-\mu) = 0$

$\Leftrightarrow \quad w^T - \Sigma^{-1} x + \Sigma^{-1}\mu = 0$

$\Leftrightarrow \quad x = \Sigma\left(\Sigma^{-1}\mu + w^T\right)$

$\Leftrightarrow \quad x^* = \mu + \Sigma w^T$

c) $\max_z \; w^T(Az + c) + b - \lambda \|z\|^2$

$\Rightarrow \quad \frac{\partial}{\partial z} w^T(Az+c) + b - \lambda\|z\|^2 \overset{!}{=} 0$

$\Leftrightarrow \quad w^T A - 2\lambda z = 0$

$\Leftrightarrow \quad z = \frac{w^T A}{2\lambda} \qquad | \; x = Az + c \Leftrightarrow z = \frac{x-c}{A}$

$\Leftrightarrow \quad \frac{x-c}{A} = \frac{w^T A}{2\lambda}$

$\Leftrightarrow \quad x = \frac{w^T A A}{2\lambda} + c$

## 2 Layer-Wise Relevance Propagation

a) $y = a_3 - a_4$

$a_1 \quad \max(0, a_1 - a_2)$

ReLU has no effect bc. $a_1$ is positive

if $a_1 \geq a_2$ : $y = a_1 - (a_1 - a_2) = a_2$

if $a_1 < a_2$ : $y = a_1 - 0 = a_1$

b) i) $R_4 = a$ ii) $R_4 = a$

$R_3 = a$ $R_3 = 0$

$R_1 = a$ $R_1 = 0$

## 3 Generalization

a) $$o(x) = -\frac{1}{\beta} \log \left( \sum_{i=1}^{M} a_i e^{-\beta \|x - \mu_i\|^2} \right)$$

$$= -\frac{1}{\beta} \log \left( \sum_{i=1}^{M} e^{-\beta \|x - \mu_i\|^2 - \frac{1}{\beta} \log(a_i)} \right)$$

b) $\min_{i=1}^{M} \{h_i\} = m$

$$o(x) - m = -\frac{1}{\beta} \log \left( \sum_{i=1}^{M} e^{-\beta h_i} \right) + \frac{1}{\beta} \log e^{-\beta m}$$

$$= -\frac{1}{\beta} \log \left( \sum_{i=1}^{M} e^{-\beta h_i} e^{\beta m} \right)$$

$$= -\frac{1}{\beta} \log \left( \sum_{i=1}^{M} e^{-\beta (h_i - m)} \right)$$

$$= -\frac{1}{\beta} \log e^{-\beta \cdot 0} + \sum_{i} e^{-\beta (h_i - m)}$$

$$= -\frac{1}{\beta} \log \left( 1 + \sum_{i} e^{-\beta (h_i - m)} \right)$$

$o(x) - m = 0$ for $\beta \to \infty$

$\Leftrightarrow o(x) = m \rightarrow$ min-pooling