Honours Math for Machine Learning HW4

Saina Koukpari - McGill University

Problem 1.1

$$\begin{aligned}
KL(p,q) &= KL((0.6, 0.4, 0), (1/2, 1/2, 0)) \\
&= -\sum_{k=1}^{3} q_k \log(p_k/q_k) \\
&= -1/2(\log(0.5/0.6)) - 1/2(\log(0.5/0.4)) - 0(\log(0/0)) \\
&= -1/2\log(0.83) - 1/2\log(1.25) \\
&= 0.03959 - 0.048455 \\
&= -0.008864
\end{aligned}$$

$$\begin{aligned}
KL(p,q) &= KL((0.2, 0.1, 0.7), (1/2, 1/2, 0)) \\
&= -\sum_{k=1}^{3} q_k \log(p_k/q_k) \\
&= -1/2(\log(0.5/0.2)) - 1/2(\log(0.5/0.1)) - 0(\log(0/0.7)) \\
&= -1/2\log(2.5) - 1/2\log(5) \\
&= -0.19897 - 0.349485 \\
&= -0.548455
\end{aligned}$$

Problem 1.3

Suppose $X = \max\{x_1, ...x_n\}$, then $e^X \le \sum_{i=1}^n e^{x_i} \le ne^X$. So for $s \in \mathbb{R}^k$, let $M = \max_i(s)$ so that

$$e^M \le \sum_{i=1}^{K} e^{s_i} \le Ke^M$$

Taking the log of both sides we obtain,

$$\begin{aligned}
&\log(e^M) \le \log\left(\sum_{i=1}^{K} e^{s_i}\right) \le \log(Ke^M) \\
\implies &M \le LSE(s) \le \log(K) + \log(e^M) \\
\implies &\max_i(s) \le LSE(s) \le \max_i(s) + \log(K) \quad \blacksquare
\end{aligned}$$

Problem 1.4

$$\begin{aligned}
\lim_{t\to\infty} f(x,t) &= \lim_{t\to\infty} \frac{1}{t} LSE(tx) \\
&= \lim_{t\to\infty} \frac{1}{t} \log(e^{tx_1} + ... + e^{tx_K}) \\
&\le \lim_{t\to\infty} \frac{1}{t} \Big(\max_i (tx_i) + \log(K) \Big) \\
&= \lim_{t\to\infty} \frac{1}{t} \Big(t \max_i (x_i) + \log(K) \Big) \\
&= \max_i (x_i) + \lim_{t\to\infty} \frac{1}{t} \log(K) \\
&= \max_i (x_i) \quad \blacksquare
\end{aligned}$$

Problem 1.5

$$\begin{aligned}
\sigma_i(s) &= \nabla_{s_i} LSE(s) \\
&= \frac{\partial}{\partial s_i} \log(e^{s_1} + ... + e^{s_K}) \\
&= \frac{e^{s_i}}{e^{s_1} + ... + e^{s_K}} \\
&= \frac{1}{1 + \sum_{j=1}^{K} e^{s_j - s_i}} \quad i \neq j \quad \square
\end{aligned}$$

$$\sum_{i=1}^{K} \frac{e^{s_i}}{e^{s_1} + ... + e^{s_K}} = \frac{e^{s_1} + ... + e^{s_K}}{e^{s_1} + ... + e^{s_K}} = 1 \in \triangle \quad \blacksquare$$

Problem 1.6

i) For $x_i \le 0$, the maximum value is 0 and so we have,

$$LSE(x) \le \max_i (x) + \log(K) = 0 + \log(K) = \log(K) \quad \square$$

ii) If at least 2 components are non-negative, then the LSE depends more significantly on the positive components. Let x_1, x_2 be two positive component, then we have

$$LSE(x) = \log(e^{x_{-n}} + ... + e^{x_{-1}} + e^{x_1} + e^{x_2}) \ge \log(e^{x_1} + e^{x_2}) \ge \log(e^0 + e^0) = \log(2) \quad \blacksquare$$

Problem 2.1

$$\begin{aligned}
\ell_{margin-K}((s_1, s_2), y) &= \ell_{margin-2}((s, -s), y) \\
&= \max(0, 1 - \delta(s, y)) \\
&= \max(0, 1 - \frac{s_y - s_{noty}}{2}) \\
&= \max(0, 1 - \frac{2sgn(y)s}{2}) \\
&= \max(0, 1 - ys) \\
&= \ell_{margin}(s, y) \quad \blacksquare
\end{aligned}$$

Problem 2.2

$$\begin{aligned}
\nabla_s \ell_{log-K}(s,y) &= < \frac{\partial}{\partial s_1} \log(e^{m_1(s,y)} + ... + e^{m_K(s,y)}), ..., \frac{\partial}{\partial s_K} \log(e^{m_1(s,y)} + ... + e^{m_K(s,y)} > \\
&= < \frac{\partial}{\partial s_1} \log(e^{s_1 - s_y} + ... + e^{s_K - s_y}), ..., \frac{\partial}{\partial s_K} \log(e^{s_1 - s_y} + ... + e^{s_K - s_y}) > \\
&= < \frac{e^{s_1 - s_y}}{e^{s_1 - s_y} + ... + e^{s_K - s_y}}, ..., \frac{e^{s_K - s_y}}{e^{s_1 - s_y} + ... + e^{s_K - s_y}} > \\
&= \sigma(m(s,y)) \quad \square
\end{aligned}$$

$$\frac{e^{s_1 - s_y}}{e^{s_1 - s_y} + ... + e^{s_K - s_y}} + ... + \frac{e^{s_K - s_y}}{e^{s_1 - s_y} + ... + e^{s_K - s_y}} = \frac{e^{s_1 - s_y} + ... e^{s_K - s_y}}{e^{s_1 - s_y} + ... + e^{s_K - s_y}} = 1 \in \triangle \quad \blacksquare$$

Problem 2.3

If $c(s) = y$ then $s = s_y$ and we have,

$$\begin{aligned}
\ell_{\log -K}(s,y) &= LSE(m(s,y)) \\
&\leq \max_i(m(s,y)) + \log(K) \\
&= \max_i(s - s_y) + \log(K) \\
&= \max_i(0) + \log(K) \\
&= \log(K) \quad \square
\end{aligned}$$

If $c(s) \neq y$ then $s \neq s_y$ and the maximum s_i gives 1, i.e. there will always be e^1 as one of the components. Further there will never be e^0. Without loss of generality, let $\max_i s_i = s_K$, then

$$\begin{aligned}
\ell_{\log -K}(s,y) &= LSE(m(s,y)) \\
&= \log(e^{s_1 - s_y} + ... + e^{s_K - s_y}) \\
&= \log(e^{s_1 - s_y} + ... + e^1) \\
&\geq \log(2) \quad \square
\end{aligned}$$

Problem 2.4

For margin loss, we have

$$\begin{aligned}
\ell_{margin}(s+t,y) &= \max(0, 1 - \delta(s+t,y)) \\
&= \max(0, 1 - s_y - t + \max_{j \neq y}(s_j + t)) \\
&= \max(0, 1 - s_y - t + \max_{j \neq y}(s_j) + t) \\
&= \max(0, 1 - s_y + \max_{j \neq y}(s_j)) \\
&= \max(0, 1 - \delta(s,y)) \\
&= \ell_{margin}(s,y) \quad \square
\end{aligned}$$

$$\begin{aligned}
\ell_{margin}(s,y) &= \max(0, 1 - \delta(s,y)) \\
&= \max(0, 1 - s_y + \max_{j \neq y}(s_j)) \\
&= \max(0, 1 - s_y - s_y + s_y + \max_{j \neq y}(s_j)) \\
&= \max(0, 1 - s_y + \max_{j \neq y}(s_j)) \\
&= \max(0, 1 - (s_y - s_y) + \max_{j \neq y}(s_j - s_y)) \\
&= \max(0, 1 - m(s_y, y) + \max_{j \neq y}(m(s_j, y))) \\
&= \max(0, 1 - \delta(m(s,y), y)) \\
&= \ell_{margin}(m(s,y), y) \quad \square
\end{aligned}$$

For log loss, we know that the margin $(s_i - s_y)$ is shift invariant since for every shift in t, the vector component shifts as well. Let s'_y be the vector component of the shifted score, then

$$s_i + t - s'_y = s_i + t - (s_y + t) = s_i - s_y$$

So we have,

$$\begin{aligned}
\ell_{log}(s+t, y) &= LSE(m(s+t, y)) \\
&= \log(e^{m_1(s+t,y)} + \dots + e^{m_K(s+t,y)}) \\
&= \log(e^{m_1(s,y)} + \dots + e^{m_K(s,y)}) \\
&= LSE(m(s,y)) \\
&= \ell_{log}(s,y) \quad \square
\end{aligned}$$

And,

$$\begin{aligned}
\ell_{log}(s,y) &= LSE(m(s,y)) \\
&= \log(e^{m_1(s,y)} + \dots + e^{m_K(s,y)}) \\
&= \log(e^{s_1 - s_y} + \dots + e^{s_K - s_y}) \\
&= \log(e^{s_1 - s_y - s_y + s_y} + \dots + e^{s_K - s_y - s_y + s_y}) \\
&= \log(e^{s_y}) + \log(e^{s_1 - s_y - s_y} + \dots + e^{s_K - s_y - s_y}) \\
&= \log(1) + \log(e^{m_1(m_1(s,y),y)} + \dots + e^{m_K(m_K(s,y),y)}) \\
&= \log(e^{m_1(m_1(s,y),y)} + \dots + e^{m_K(m_K(s,y),y)}) \\
&= LSE(m(m(s,y),y)) \\
&= \ell_{log}(m(s,y),y) \quad \square
\end{aligned}$$