CS7643: Deep Learning
Fall 2018
HW0 Solutions

Niranjan Thakurdesai

September 5, 2018

Note to instructor and TAs: The solutions were discussed with Rahul Duggal and Sreenivasan AC.

1 Probability and Statistics

1.

$$\begin{aligned}\mathbb{E}[payout] &= 1 \times P(payout = 1) - 1/4 \times P(payout = -1/4) \\ &= 1 \times P(x = 1) - 1/4 \times P(x \neq 1) \\ &= 1 \times 1/6 - 1/4 \times 5/6 \\ &= -1/24\end{aligned}$$

We are expected to lose -\$1/24. Hence, this is not a good bet.

2.

$$C(x) = Pr(X \leq x) = \int_0^x p(x)dx$$

Case 1: $-\infty < x \leq 0$

$$C(x) = 0$$

Case 2: $0 \leq x \leq 1/2$

$$\begin{aligned} C(x) &= \int_0^x 4xdx \\ &= 4[x^2/2]_0^x \\ &= 2x^2 \end{aligned}$$

Case 3: $1/2 \leq x \leq 1$

$$\begin{aligned} C(x) &= \int_0^{1/2} 4xdx + \int_{1/2}^x (-4x+4)dx \\ &= 4[x^2/2]_0^{1/2} - 4[x^2/2]_{1/2}^x + 4[x]_{1/2}^x \\ &= 1 - 2x^2 + 4x - 2 \\ &= -2x^2 + 4x - 1 \end{aligned}$$

Case 4: $x \geq 1$

$$C(x) = 1$$

Thus,

$$C(x) = \begin{cases} 0 & -\infty < x \leq 0 \\ 2x^2 & 0 \leq x \leq 1/2 \\ -2x^2 + 4x - 1 & 1/2 \leq x \leq 1 \\ 1 & x \geq 1 \end{cases}$$

3.

$$E[(X-\mu)^2] = E[X^2 - 2X\mu + \mu^2]$$

By the property of linearity and the expected value of a constant,

$$\begin{aligned} E[(X-\mu)^2] &= E[X^2] - 2\mu E[X] + \mu^2 \\ E[(X-\mu)^2] &= E[X^2] - 2E[X].E[X] + (E[X])^2 \\ &= E[X^2] - (E[X])^2 \end{aligned}$$

4.

$$\int_{-\infty}^{\infty} p(x)(ax^2 + bx + c)dx = a\int_{-\infty}^{\infty} x^2 p(x)dx + b\int_{-\infty}^{\infty} xp(x)dx + c\int_{-\infty}^{\infty} p(x)dx$$

From Q3,

$$\int_{-\infty}^{\infty} p(x)(ax^2 + bx + c)dx = a(Var[x] + (E[x])^2) + bE[x] + c$$

As $E[x] = 0$ and $Var[x] = 1$ for a standard normal distribution,

$$\int_{-\infty}^{\infty} p(x)(ax^2 + bx + c)dx = a + c$$

2 Proving stuff

5. Let $f(x) = \log_e x - (x-1)$. The first derivative of $f(x)$ is $f'(x) = \frac{1}{x} - 1$ and the second derivative is $f''(x) = -\frac{1}{x^2}$. Observe that $f''(x) < 0, \forall x > 0$. Thus, $f(x)$ is concave $\forall x > 0$. Moreover, $f'(1) = 0$ which means $f(x)$ has a global maximum at $x = 1$. Hence, $f(x) = \log_e x - (x-1) \leq f(1) = 0$ or $\log_e x \leq x - 1, \forall x > 0$ with equality if and only if $x = 1$.

6. (a)

$$
\begin{aligned}
KL(p,q) &= \sum_{i=1}^{k} p_i \log\left(\frac{p_i}{q_i}\right) \\
&= -\sum_{i=1}^{k} p_i \log\left(\frac{q_i}{p_i}\right) \\
&\geq \sum_{i=1}^{k} p_i \left(1 - \frac{q_i}{p_i}\right), \text{using the identity in Q5} \\
&= \sum_{i=1}^{k} p_i - q_i \\
&= 1 - 1 \\
&= 0 \\
\implies KL(p,q) &\geq 0
\end{aligned}
$$

Thus, $KL(p,q)$ is always positive (or more correctly, always non-negative).

(b) $KL(p, q) = 0$ when $\frac{p_i}{q_i} = 1$ or $p_i = q_i$, i.e. the two probability distributions are identical.

(c) Let p and q be as follows:

$$p = \begin{cases} 1/2 & k = 0 \\ 1/2 & k = 1 \end{cases}$$

$$q = \begin{cases} 1/4 & k = 0 \\ 3/4 & k = 1 \end{cases}$$

$$\begin{aligned} KL(p, q) &= \sum_i p_i \log\left(\frac{p_i}{q_i}\right) \\ &= \frac{1}{2}\log 2 + \frac{1}{2}\log\left(\frac{2}{3}\right) \\ &= \frac{1}{2}\log\left(\frac{4}{3}\right) \\ &= 0.14 \\ KL(q, p) &= \sum_i q_i \log\left(\frac{q_i}{p_i}\right) \\ &= \frac{1}{4}\log\left(\frac{1}{2}\right) + \frac{3}{4}\log\left(\frac{3}{2}\right) \\ &= 0.13 \end{aligned}$$

$$\therefore KL(p, q) \neq KL(q, p)$$

3 Calculus

7. Let

$$f_1(x) = \log x + \frac{1}{2}, \frac{df_1}{dx} = \frac{1}{x}$$
$$f_2(x) = 5x, \frac{df_2}{dx} = 5$$
$$f_3(\mathbf{x}) = \max\{x_1, x_2\}.\frac{x_3}{x_4} - (x_5 + x_6)$$

When $x_1 > x_2$,

$$\frac{\partial f_3}{\partial x_1} = \frac{x_3}{x_4}, \frac{\partial f_3}{\partial x_2} = 0, \frac{\partial f_3}{\partial x_3} = \frac{x_1}{x_4}, \frac{\partial f_3}{\partial x_4} = -\frac{x_1 x_3}{x_4^2}, \frac{\partial f_3}{\partial x_5} = -1, \frac{\partial f_3}{\partial x_6} = -1$$
$$\frac{d\sigma}{dx} = \frac{-1}{(1+e^{-x})^2}.-e^{-x} = \frac{e^{-x}}{(1+e^{-x})^2} = \sigma(x)(1-\sigma(x))$$

Thus,

$$f(\mathbf{x}) = \sigma\left(f_1\left(f_2\left(f_3\left(\mathbf{x}\right)\right)\right)\right)$$

At $\hat{\mathbf{x}}$,

$$\begin{aligned}
f_3(\hat{\mathbf{x}}) &= \frac{1}{2} \\
f_2(f_3(\hat{\mathbf{x}})) &= \frac{5}{2} \\
f_1(f_2(f_3(\hat{\mathbf{x}}))) &= \log\left(\frac{5}{2}\right) + \frac{1}{2} \\
\sigma(f_1(f_2(f_3(\hat{\mathbf{x}})))) &= \frac{1}{1+e^{-\left(\log\left(\frac{5}{2}\right)+\frac{1}{2}\right)}} \\
&= \frac{1}{1+\frac{2}{5}e^{-\frac{1}{2}}} \\
&= 0.805
\end{aligned}$$

Using the chain rule,

$$\begin{aligned}
\frac{\partial f}{\partial x_1} &= \frac{d\sigma(f_1(f_2(f_3(\hat{\mathbf{x}}))))}{dx} \times \frac{df_1(f_2(f_3(\hat{\mathbf{x}})))}{dx} \times \frac{df_2(f_3(\hat{\mathbf{x}}))}{dx} \times \frac{\partial f_3(\hat{\mathbf{x}})}{\partial x_1} \\
&= 0.805(1-0.805) \times \frac{2}{5} \times 5 \times \frac{6}{12} \\
&= 0.157
\end{aligned}$$

$$\begin{aligned}
\frac{\partial f}{\partial x_2} &= \frac{d\sigma(f_1(f_2(f_3(\hat{\mathbf{x}}))))}{dx} \times \frac{df_1(f_2(f_3(\hat{\mathbf{x}})))}{dx} \times \frac{df_2(f_3(\hat{\mathbf{x}}))}{dx} \times \frac{\partial f_3(\hat{\mathbf{x}})}{\partial x_2} \\
&= 0.805(1-0.805) \times \frac{2}{5} \times 5 \times 0 \\
&= 0
\end{aligned}$$

$$\begin{aligned}\frac{\partial f}{\partial x_3} &= \frac{d\sigma(f_1(f_2(f_3(\hat{\mathbf{x}}))))}{dx} \times \frac{df_1(f_2(f_3(\hat{\mathbf{x}})))}{dx} \times \frac{df_2(f_3(\hat{\mathbf{x}}))}{dx} \times \frac{\partial f_3(\hat{\mathbf{x}})}{\partial x_3} \\ &= 0.805(1-0.805) \times \frac{2}{5} \times 5 \times \frac{5}{12} \\ &= 0.131\end{aligned}$$

$$\begin{aligned}\frac{\partial f}{\partial x_4} &= \frac{d\sigma(f_1(f_2(f_3(\hat{\mathbf{x}}))))}{dx} \times \frac{df_1(f_2(f_3(\hat{\mathbf{x}})))}{dx} \times \frac{df_2(f_3(\hat{\mathbf{x}}))}{dx} \times \frac{\partial f_3(\hat{\mathbf{x}})}{\partial x_4} \\ &= 0.805(1-0.805) \times \frac{2}{5} \times 5 \times -\frac{5 \times 6}{12^2} \\ &= -0.065\end{aligned}$$

$$\begin{aligned}\frac{\partial f}{\partial x_5} &= \frac{d\sigma(f_1(f_2(f_3(\hat{\mathbf{x}}))))}{dx} \times \frac{df_1(f_2(f_3(\hat{\mathbf{x}})))}{dx} \times \frac{df_2(f_3(\hat{\mathbf{x}}))}{dx} \times \frac{\partial f_3(\hat{\mathbf{x}})}{\partial x_5} \\ &= 0.805(1-0.805) \times \frac{2}{5} \times 5 \times -1 \\ &= -0.314\end{aligned}$$

$$\begin{aligned}\frac{\partial f}{\partial x_6} &= \frac{d\sigma(f_1(f_2(f_3(\hat{\mathbf{x}}))))}{dx} \times \frac{df_1(f_2(f_3(\hat{\mathbf{x}})))}{dx} \times \frac{df_2(f_3(\hat{\mathbf{x}}))}{dx} \times \frac{\partial f_3(\hat{\mathbf{x}})}{\partial x_6} \\ &= 0.805(1-0.805) \times \frac{2}{5} \times 5 \times -1 \\ &= -0.314\end{aligned}$$

The above calculations were done analytically and verified by numerical computation using self-written Python code.

8.

```
vectorized loss: 2.360132e+00 computed in 1.231000s
loss: 2.360132
sanity check: 2.302585
numerical: -0.542295 analytic: -0.542295, relative error: 4.279945e-08
numerical: 2.716988 analytic: 2.716987, relative error: 2.555772e-08
numerical: -1.597924 analytic: -1.597924, relative error: 3.561914e-08
numerical: 1.944617 analytic: 1.944617, relative error: 4.721436e-09
numerical: -1.244018 analytic: -1.244018, relative error: 3.222809e-09
numerical: 0.790825 analytic: 0.790825, relative error: 9.420841e-08
numerical: -0.102579 analytic: -0.102579, relative error: 1.644851e-07
numerical: 0.869372 analytic: 0.869371, relative error: 8.167231e-08
numerical: -3.055064 analytic: -3.055065, relative error: 2.539041e-08
numerical: -1.228758 analytic: -1.228758, relative error: 1.193294e-08
```

Figure 1: Output of cell 3

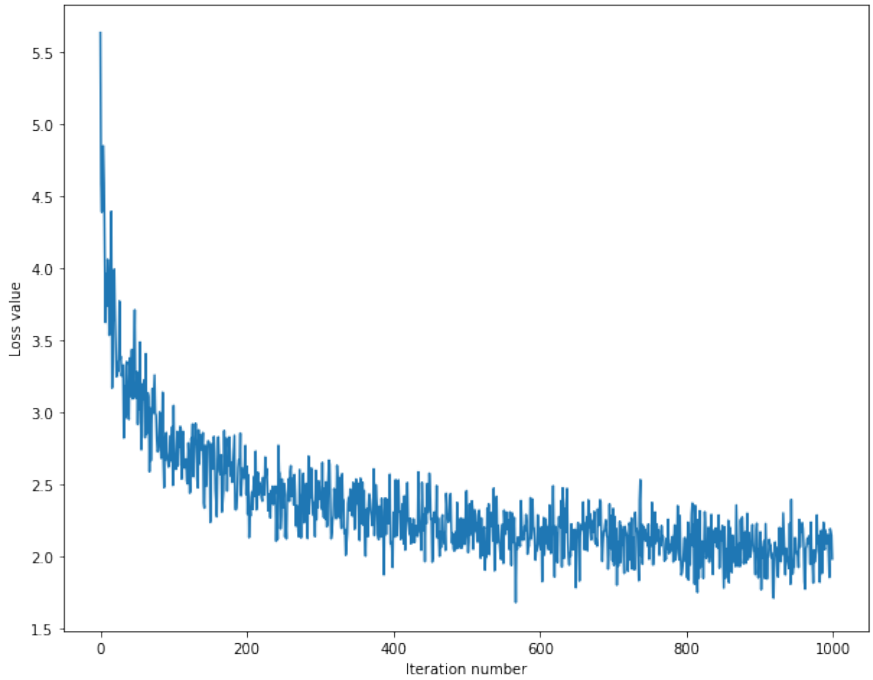

Figure 2: Training loss

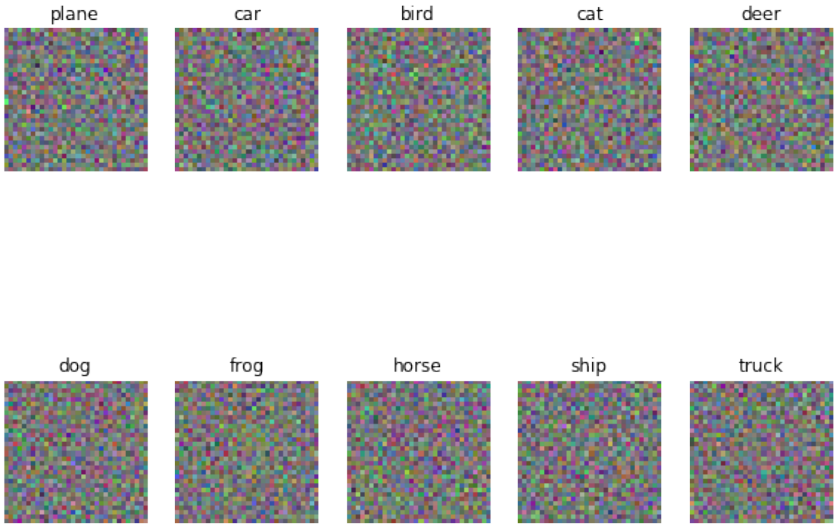

Figure 3: Weight visualizations

The weight visualizations do not correspond well with their respective classes. This indicates that the network hasn't learnt the representations of the classes.

9. We will prove this using the definition of convexity, i.e. $L(W)$ is convex if and only if $L(\lambda W_1 + (1-\lambda)W_2) \leq \lambda L(W_1) + (1-\lambda)L(W_2), \forall \lambda \in [0,1]$.

Let w_{1i} and w_{2i} denote the ith rows of W_1 and W_2 respectively.

$$L(\lambda W_1 + (1-\lambda)W_2) = -\log \frac{e^{(\lambda w_{1y} + (1-\lambda)w_{2y})x}}{\sum_k e^{(\lambda w_{1k} + (1-\lambda)w_{2k})x}} \tag{1}$$

$$\lambda L(W_1) + (1-\lambda)L(W_2) = -\lambda \log \frac{e^{w_{1y}x}}{\sum_k e^{w_{1k}x}} - (1-\lambda)\log \frac{e^{w_{2y}x}}{\sum_k e^{w_{2k}x}} \tag{2}$$

$$= -\log \frac{e^{\lambda w_{1y}x}}{(\sum_k e^{w_{1k}x})^{\lambda}} - \log \frac{e^{(1-\lambda)w_{2y}x}}{(\sum_k e^{w_{2k}x})^{(1-\lambda)}} \tag{3}$$

$$= -\log \frac{e^{(\lambda w_{1y} + (1-\lambda)w_{2y})x}}{(\sum_k e^{w_{1k}x})^{\lambda}(\sum_k e^{w_{2k}x})^{(1-\lambda)}} \tag{4}$$

As the numerator in (1) and (4) is the same, we just have to prove that $\log \sum_k e^{(\lambda w_{1k} + (1-\lambda)w_{2k})x} \leq \log(\sum_k e^{w_{1k}x})^{\lambda}(\sum_k e^{w_{2k}x})^{(1-\lambda)}$. Since $\log x$ is monotone increasing, it is enough to prove that $\sum_k e^{(\lambda w_{1k} + (1-\lambda)w_{2k})x} \leq (\sum_k e^{w_{1k}x})^{\lambda}(\sum_k e^{w_{2k}x})^{(1-\lambda)}$. Let $e^{w_{1k}x} = u_k$ and $e^{w_{2k}x} = v_k$. From Holder's inequality,

$$\sum_{i=1}^{n} x_i y_i \leq \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}} \cdot \left(\sum_{i=1}^{n} |y_i|^q\right)^{\frac{1}{q}} \tag{5}$$

$$\text{where } \frac{1}{p} + \frac{1}{q} = 1 \tag{6}$$

Applying this inequality to $\sum_k e^{(\lambda w_{1k} + (1-\lambda)w_{2k})x}$,

$$\sum_k u_k^{\lambda} v_k^{(1-\lambda)} \leq \left(\sum_k u_k^{\lambda \cdot \frac{1}{\lambda}}\right)^{\lambda} \left(\sum_k v_k^{(1-\lambda)\cdot \frac{1}{(1-\lambda)}}\right)^{(1-\lambda)} \tag{7}$$

$$\implies \sum_k e^{(\lambda w_{1k} + (1-\lambda)w_{2k})x} \leq (\sum_k e^{w_{1k}x})^{\lambda}(\sum_k e^{w_{2k}x})^{(1-\lambda)} \tag{8}$$

Hence, $L(.)$ is convex in W.