Homework # 2

TJ Bai

October 22, 2023

1. (a) We first decompose $\mathbb{E}_S\mathbb{E}_{XY}\left[(y-\hat{h}_S(x))^2\right] = \mathbb{E}_S\mathbb{E}_{XY}\left[(\bar{h}(x)-\hat{h}_S(x))^2\right] + \mathbb{E}_{XY}\left[(y-\bar{h}(x))^2\right]$

   $\mathbb{E}_S\mathbb{E}_{XY}\left[(y-\bar{h}(x)+\bar{h}(x)-\hat{h}_S(x))^2\right]$
   $= \mathbb{E}_S\mathbb{E}_{XY}\left[(\bar{h}(x)-\hat{h}_S(x))^2\right] + 2\mathbb{E}_S\mathbb{E}_{XY}\left[(y-\bar{h}(x))(\bar{h}(x)-\hat{h}_S(x))\right] + \mathbb{E}_{XY}\left[(y-\bar{h}(x))^2\right]$
   $= \mathbb{E}_S\mathbb{E}_{XY}\left[(\bar{h}(x)-\hat{h}_S(x))^2\right] + 2\mathbb{E}_{XY}\left[(y-\bar{h}(x))\mathbb{E}_S\left[\bar{h}(x)-\hat{h}_S(x)\right]\right] + \mathbb{E}_{XY}\left[(y-\bar{h}(x))^2\right]$

   By definition, $\mathbb{E}_S\left[\bar{h}(x)-\hat{h}_S(x)\right] = 0\ldots$
   $\implies \mathbb{E}_S\mathbb{E}_{XY}\left[(\bar{h}(x)-\hat{h}_S(x))^2\right] + 2\mathbb{E}_{XY}\left[(y-\bar{h}(x))(0)\right] + \mathbb{E}_{XY}\left[(y-\bar{h}(x))^2\right]$
   $= \mathbb{E}_S\mathbb{E}_{XY}\left[(\bar{h}(x)-\hat{h}_S(x))^2\right] + 2\mathbb{E}_{XY}[0] + \mathbb{E}_{XY}\left[(y-\bar{h}(x))^2\right]$
   $= \mathbb{E}_S\mathbb{E}_{XY}\left[(\bar{h}(x)-\hat{h}_S(x))^2\right] + \mathbb{E}_{XY}\left[(y-\bar{h}(x))^2\right]$ ■

   It is natural to describe the first term in this decomposition as variance, as it gives the difference in prediction between the expected predictor, $\bar{h}(x)$, and the actual predictor produced by the learning algorithm $\hat{h}_S(x)$.

   (b) $\mathbb{E}_{XY}\left[(y-\bar{h}(x))^2\right] = \mathbb{E}_{XY}\left[y-h^*(x)+h^*(x)-\bar{h}(x)\right]$
   $= \mathbb{E}_{XY}[(y-h^*(x))^2] + 2\mathbb{E}_X\mathbb{E}_{Y|X}[(y-h^*(x))(h^*(x)-\bar{h}(x))] + \mathbb{E}_{XY}[(h^*(x)-\bar{h}(x))^2]$

   Focusing on the cross-term $\mathbb{E}_X\mathbb{E}_{Y|X}[(y-h^*(x))(h^*(x)-\bar{h}(x))]\ldots$
   For a fixed x, $h^*(x)-\bar{h}(x)$ is constant $\implies \mathbb{E}_X\left[(h^*(x)-\bar{h}(x))\mathbb{E}_{Y|X}[y-h^*(x)]\right]$
   By definition of the Bayes optimal rule, $\mathbb{E}_{Y|X}[y-h^*(x)] = 0 \implies \mathbb{E}_X[(h^*(x)-\bar{h}(x))(0)] = 0$

   Therefore, $\mathbb{E}_{XY}\left[(y-\bar{h}(x))^2\right] = \mathbb{E}_{XY}[(y-h^*(x))^2] + \mathbb{E}_{XY}[(h^*(x)-\bar{h}(x))^2]$ ■

   The first term may be referred to as "noise" as it provides the inherent loss that will be present in the given distribution even for the optimal predictor. Bias refers to the average difference between our expected output and the true output, which is captured by $\bar{h}$ and h^*, respectively. Therefore, the second term is the bias-squared because it squares the difference.

   (c) The above decomposition allows us to understand that the error in the output of the learning algorithm is a combination of variance (how far our output is from the average), bias (how far our average is from the true value), and noise (variability inherent to the joint distribution). Noise especially represents irreducible error in the data, and can not be eliminated. Meanwhile, we can make decisions in our algorithm design to tradeoff between minimizing bias and minimizing variance. One such technique is regularization, which lowers variance at the cost of increasing bias. Cross-validation techniques can also often assist in setting hyperparameters or developing learning algorithms that are able to find a "sweet-spot" for both, as a model with high variance will tend to overfit to a specific training set, whereas a model with high bias will have high error on average across each validation fold.

2. For each of the following parts, we make use of the following lemmas.

   Maximizing a is equivalent to minimizing $-a$, and vice versa.

   Maximizing ca is equivalent to maximizing a for $c \in \mathbb{R}$, and same for minimization.

   Maximizing $c + a$ is equivalent to maximizing a for $c \in \mathbb{R}$, and same for minimization.

   Maximizing a and maximizing b is equivalent to maximizing $a + b$, and same for minimization.

   (a) We show that maximizing $\hat{w}_{\text{MAP}} = \underset{w \in \mathbb{R}^{d+1}}{\text{argmax}} \left\{ \sum_{i=1}^{n} \log p(y_i | x_i, w) + \log p(w) \right\}$

   for $y = w^\top x + v$, $v \sim N(0, \sigma^2)$ and $w \sim N\left(0, \frac{I}{\mu}\right)$ is equivalent to ridge regression.

   First, $\sum_{i=1}^{n} \log p(y_i | x_i, w) = \sum_{i=1}^{n} \log \left(\frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{1}{2} \left(\frac{y_i - w^\top x_i}{\sigma} \right)^2 \right)$

   $= \log \frac{n}{\sigma\sqrt{2\pi}} - \frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - w^\top x_i)^2}{\sigma^2} \implies$ equivalent to maximizing $-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - w^\top x_i)^2$

   $\implies$ equivalent to minimizing $\sum_{i=1}^{n} (y_i - w^\top x_i)^2$ which is the first term in ridge regression.

   Now we consider $\log p(w) = \log \left(\frac{\mu}{I\sqrt{2\pi}} \exp \left(-\frac{1}{2} \left(\frac{\sqrt{\mu} w}{\sqrt{I}} \right)^2 \right) \right) = \log \frac{\mu}{I\sqrt{2\pi}} - \frac{\mu}{2I} w^\top w$

   $\implies$ equivalent to maximizing $-\frac{\mu}{2I} w^\top w = -\frac{\mu\sigma^2}{2} w^\top w$

   $\implies$ equivalent to minimizing $\mu\sigma^2 \sum_{i=1}^{d} w_i^2$ which is the second term in ridge regression.

   Thus, $\underset{w \in \mathbb{R}^{d+1}}{\text{argmax}} \left\{ \sum_{i=1}^{n} \log p(y_i | x_i, w) + \log p(w) \right\} = \underset{w \in \mathbb{R}^{d+1}}{\text{argmin}} \left\{ \sum_{i=1}^{n} (y_i - w^\top x_i)^2 + \mu\sigma^2 \sum_{i=1}^{d} w_i^2 \right\}$ ■

   (b) We seek to minimize $(y - Xw)^\top (y - Xw) + \mu\sigma^2 w^\top w$.

   $\frac{\partial}{\partial w} (y - Xw)^\top (y - Xw) + \frac{\partial}{\partial w} \mu\sigma^2 w^\top w$

   $= \frac{\partial}{\partial w} (y^\top - w^\top X^\top)(y - Xw) + 2\mu\sigma^2 w = \frac{\partial}{\partial w} (y^\top y - y^\top Xw - w^\top X^\top y + w^\top X^\top Xw) + 2\mu\sigma^2 w$

   $= -X^\top y - X^\top y + 2X^\top Xw + 2\mu\sigma^2 w = -2(X^\top y - X^\top Xw - \mu\sigma^2 w) = 0$

   $\implies X^\top y = (X^\top X + \mu\sigma^2 I) w \implies \boxed{w = (X^\top X + \mu\sigma^2 I)^{-1} X^\top y}$ ■

   (c) We show that maximizing $\hat{w}_{\text{MAP}} = \underset{w \in \mathbb{R}^{d+1}}{\text{argmax}} \left\{ \sum_{i=1}^{n} \log p(y_i = c | x_i, w) + \log p(w) \right\}$

   for a logistic regression model where $w_i \sim \mathcal{L}(0, b)$ leads to L1 regularization.

   $\sum_{i=1}^{n} \log p(y_i = c | x_i, w) = \sum_{i=1}^{n} \log \left(\sigma(w^\top x_i)^{y_i} (1 - \sigma(w^\top x_i))^{1 - y_i} \right)$

   $= \sum_{i=1}^{n} \left(y_i \sigma(w^\top x_i) + (1 - y_i)(1 - \sigma(w^\top x_i)) \right)$

And maximizing this is equivalent to minimizing $-\sum_{i=1}^{n}\big(y_i\sigma(w^\intercal x_i)+(1-y_i)(1-\sigma(w^\intercal x_i))\big)$

Now consider $p(w)=\prod_{i=1}^{d+1}\frac{1}{2b}\exp\left(-\frac{|w_i|}{b}\right) \implies \log p(w)=\frac{d+1}{2b}+\sum_{i=1}^{d+1}\frac{-|w_i|}{b}$

Maximizing this is equivalent to minimizing $\frac{1}{b}\sum_{i=1}^{d+1}|w_i|=\frac{\|w\|_1}{b} \implies \lambda=\frac{1}{b}$

Thus, $\underset{w\in\mathbb{R}^{d+1}}{\text{argmax}}\left\{\sum_{i=1}^{n}\log p(y_i=c|x_i,w)+\log p(w)\right\}$

$=\underset{w\in\mathbb{R}^{d+1}}{\text{argmin}}\left\{-\sum_{i=1}^{n}\big(y_i\sigma(w^\intercal x_i)+(1-y_i)(1-\sigma(w^\intercal x_i))\big)+\lambda\|w\|_1\right\}$ for $\lambda=\frac{1}{b}$ ■

3. We show that $\forall q, \int_x\sum_{c=1}^{C}\sum_{c'=1}^{C}L(c',c)q(c_r=c'|x)p(x,y=c)dx \geq \int_x\sum_{c=1}^{C}L(h^*(x),c)p(x,y=c)dx$

   It suffices to demonstrate that the conditional risk $R(h_r|x)\geq R(h^*|x)$.

   $R(h_r|x)=\sum_{c=1}^{C}\left(\sum_{c'=1}^{C}L(c',c)q(c_r=c'|x)\right)p(y=c|x)=\sum_{c=1}^{C}(1-q(c_r=c|x))p(y=c|x)$

   $=\sum_{c=1}^{C}p(y=c|x)-\sum_{c=1}^{C}q(c_r=c|x)p(y=c|x)=1-\sum_{c=1}^{C}q(c_r=c|x)p(y=c|x)$

   Recall that $R(h^*(x)|x)=1-p(y=h^*(x)|x)$, as shown in lecture.

   $\implies$ we now seek to show that $\sum_{c=1}^{C}q(c_r=c|x)p(y=c|x)\leq p(y=h^*(x)|x)$

   By definition, $h^*(x)=\underset{c}{\text{argmax}}\{p(y=c|x)\}$

   $\implies \sum_{c=1}^{C}q(c_r=c|x)p(y=c|x)\leq\sum_{c=1}^{C}q(c_r=c|x)p(y=h^*(x)|x)=p(y=h^*(x)|x)$ ■

4. (a) We show that the softmax regression model corresponds to modeling the log-odds between any two classes $c_1,c_2\in\{1,\dots,C\}$ by a linear function.

   For any $i\in\{1,\dots,C-1\}$ we can write $\log\left(\frac{p(y=c_i|x)}{p(y=c_C|x)}\right)=w_i\cdot x$

   $\implies p(y=c_i|x)=\exp(w_i\cdot x)p(y=c_C|x)$.

   It follows that $p(y=c_C|x)+\sum_{i=1}^{C-1}\exp(w_i\cdot x)p(y=c_C|x)=p(y=c_C|x)\big(1+\sum_{i=1}^{C-1}\exp(w_i\cdot x)\big)=1$

   $\implies p(y=c_C|x)=\frac{1}{1+\sum_{i=1}^{C-1}\exp(w_i\cdot x)}$ and $p(y=c_i|x)=\frac{\exp(w_i\cdot x)}{1+\sum_{i=1}^{C-1}\exp(w_i\cdot x)}$

   If we take $w_C=0 \implies \exp(w_C\cdot x)=1$ then we recover $p(y=c|x)=\frac{\exp(w_c\cdot x)}{\sum_{y=1}^{C}\exp(w_y\cdot x)}$ ■

(b) We show that the binary case of the softmax model is equivalent to logistic regression.

$$\frac{\exp(w_1 \cdot x)}{\exp(w_1 \cdot x) + \exp(w_2 \cdot x)} = \frac{1}{1 + \frac{\exp(w_2 \cdot x)}{\exp(w_1 \cdot x)}} = \frac{1}{1 + \exp((w_2 - w_1) \cdot x)} = \sigma((w_1 - w_2) \cdot x) \blacksquare$$

5. We first show that $\frac{\partial p_c}{\partial z_k}$ is positive if $k = c$. $\frac{\partial}{\partial z_c} \frac{\exp(z_c)}{\sum_{i=1}^{C} \exp(z_i)} = \frac{\left(\sum_{i=1}^{C} \exp(z_i)\right) \exp(z_c) - \exp(z_c) \exp(z_c)}{\left(\sum_{i=1}^{C} \exp(z_i)\right)^2}$

Note that $\left(\sum_{i=1}^{C} \exp(z_i)\right) \exp(z_c) = \left(\sum_{i \neq k}^{C} \exp(z_i)\right) \exp(z_c) + \exp(z_c) \exp(z_c) > \exp(z_k) \exp(z_k)$

It follows that the numerator is positive and $\left(\sum_{i=1}^{C} \exp(z_i)\right)^2$ is clearly also positive $\implies \frac{\partial p_c}{\partial z_k} > 0$

Now consider $\frac{\partial p_c}{\partial z_k}$ when $k \neq c$.

$$\frac{\partial}{\partial z_k} \frac{\exp(z_c)}{\sum_{i=1}^{C} \exp(z_k)} = \frac{\left(\sum_{i=1}^{C} \exp(z_i)\right)(0) - \exp(z_k) \exp(z_k)}{\left(\sum_{i=1}^{C} \exp(z_i)\right)^2} = \frac{-\exp(2z_k)}{\left(\sum_{i=1}^{C} \exp(z_i)\right)^2}$$

$\exp(2z_k) > 0 \implies -\exp(2z_k) < 0 \implies \frac{\partial p_c}{\partial z_k} = \frac{-\exp(2z_k)}{\left(\sum_{i=1}^{C} \exp(z_i)\right)^2} < 0 \blacksquare$

Because $\frac{\partial p_c}{\partial z_k} > 0$ when $k = c$, it follows that increasing z_k increases p_k.

Similarly, $\frac{\partial p_c}{\partial z_k} < 0$ when $k \neq c$ implies that increasing z_k decreases p_c. This is intuitive as increasing z_k would increase the denominator while keeping the value in the numerator the same.

6. We first consider when $z_j < M$.

$q_j(s) = \frac{\exp(s \cdot z_j)}{\sum_{i=1}^{K} \exp(s \cdot z_i)} = \frac{1}{\sum_{i=1}^{K} \exp(s \cdot (z_i - z_j))}$

Because $z_j < M$, there exists some $k \in \{1, \ldots, K\}$ such that $z_k > z_j$.

$\implies q_j(s) = \frac{1}{\exp(s \cdot (z_k - z_j)) + \sum_{i \neq k} \exp(s \cdot (z_i - z_j))}$

As $s \to \infty$, $\exp(s \cdot (z_k - z_j)) \to \infty$ because $z_k - z_j > 0$.

Every term in $\sum_{i \neq k} \exp(s \cdot (z_i - z_j))$ either tends to 0 or $\infty \implies q_j(s) \to 0 \blacksquare$

We now consider when $z_j = M$.

$q_j(s) = \frac{1}{\sum_{i=1}^{C} \exp(s \cdot (z_i - z_j))} = \frac{1}{\sum_{z_i = M} \exp(s \cdot (z_i - M)) + \sum_{z_i \neq M} \exp(s \cdot (z_i - M))}$

$$\sum_{z_i = M} \exp(s \cdot (z_i - z_j)) = \sum_{i=1}^{K} \exp(s \cdot (M - M)) = \sum_{i=1}^{K} 1 = K$$

Every term in $\sum_{z_i \neq M} \exp(s \cdot (z_i - M))$ approaches 0 as $s \to \infty$ because $z_i - M < 0$.

Therefore, the denominator of $q_j(s)$ approaches $K \implies q_j(s) \to \frac{1}{K} \blacksquare$

7. The log-loss contribution of a single training pair (x, y) is $-\sum_{i=1}^{N} \log \hat{p}(y = c|x, W, b) + \frac{\lambda}{N}\|W\|^2$

This is also equivalent to writing $-\sum_{i=1}^{N} \log \hat{p}(y = c|x, W, b) + \lambda'\|W\|^2$ for some $\lambda' = \frac{\lambda}{N}$

Thus, for the time-being let us denote the log-loss $l(x, y)$

$= -\left(w_c^\intercal x + b_c - \log \sum_{i=1}^{N} \exp\left(w_i^\intercal x + b_i\right)\right) + \lambda\|W\|^2 = -w_c^\intercal x - b_c + \log \sum_{i=1}^{N} \exp\left(w_i^\intercal x + b_i\right) + \lambda\|W\|^2$

We can compute the gradient with respect to W column-wise.

Let A_i represent the ith column of any matrix A.

$\frac{\partial l(x,y)}{\partial W} = \left(\frac{\partial l(x,y)}{\partial w_1}, \ldots, \frac{\partial l(x,y)}{\partial w_c}, \ldots, \frac{\partial l(x,y)}{\partial w_C}\right)$, where $\left(\frac{\partial l(x,y)}{\partial W}\right)_i = \frac{\partial l(x,y)}{\partial w_i}$.

We now consider the cases where $i = c$ and $i \neq c$.

When $i = c \ldots$

$$\frac{\partial}{\partial w_c}\left(-w_c^\intercal x - b_c + \log \sum_{i=1}^{N} \exp\left(w_j^\intercal x + b_j\right) + \lambda\|W\|^2\right)$$

$$= -x + \frac{\frac{\partial}{\partial w_c}\sum_{i=1}^{N} \exp\left(w_i^\intercal x + b_i\right)}{\sum_{i=1}^{N} \exp\left(w_i^\intercal x + b_i\right)} + \lambda \sum_{i=1}^{C} \frac{\partial}{\partial w_c}\|w_i\|^2$$

$$= -x + \frac{x \exp\left(w_c^\intercal x + b_c\right)}{\sum_{i=1}^{N} \exp\left(w_i^\intercal x + b_i\right)} + 2\lambda w_c = x\left(\frac{\exp\left(w_c^\intercal x + b_c\right)}{\sum_{i=1}^{N} \exp\left(w_i^\intercal x + b_i\right)} - 1\right) + 2\lambda w_c$$

Once again, this is equivalent to $x\left(\frac{\exp\left(w_c^\intercal x + b_c\right)}{\sum_{i=1}^{N} \exp\left(w_i^\intercal x + b_i\right)} - 1\right) + \lambda' w_c$

Now, when $i \neq c \ldots$

$$\frac{\partial}{\partial w_i} l(x, y) = \frac{\partial}{\partial w_i}(-w_c^\intercal x - b_c) + \frac{\frac{\partial}{\partial w_i}\sum_{j=1}^{N} \exp\left(w_j^\intercal x + b_j\right)}{\sum_{j=1}^{N} \exp\left(w_j^\intercal x + b_j\right)} + \lambda' w_i = x\left(\frac{\exp\left(w_i^\intercal x + b_i\right)}{\sum_{j=1}^{N} \exp\left(w_j^\intercal x + b_j\right)}\right) + \lambda' w_i$$

Let z be the vector such that $z_i = \frac{\exp\left(w_i^\intercal x + b_i\right)}{\sum_{j=1}^{N} \exp\left(w_j^\intercal x + b_j\right)} = \hat{p}(y|x, W, b)$

Let t be the one-hot encoded vector corresponding to this training pair (x, y).

We can equivalently denote $\boxed{\frac{\partial l(x,y)}{\partial W} = x(z - t)^\intercal + \lambda W}$

The gradient with respect to b can be computed similarly...

$$\frac{\partial l(x, y)}{\partial b_c} = \frac{\partial}{\partial b_c}\left(-w_c^\intercal x - b_c + \log \sum_{i=1}^{N} \exp\left(w_j^\intercal x + b_j\right) + \lambda\|W\|^2\right)$$

$$= -1 + \frac{\frac{\partial}{\partial b_c}\sum_{i=1}^{N} \exp\left(w_j^\intercal x + b_j\right)}{\sum_{i=1}^{N} \exp\left(w_j^\intercal x + b_j\right)} + 0 = \frac{\exp\left(w_c^\intercal x + b_c\right)}{\sum_{i=1}^{N} \exp\left(w_j^\intercal x + b_j\right)} - 1$$

When $i \neq c$, $\dfrac{\partial l(x,y)}{\partial b_i} = \dfrac{\partial}{\partial b_i}(-w_c^\intercal x - b_c) + \dfrac{\frac{\partial}{\partial b_c}\sum_{i=1}^{N}\exp\left(w_j^\intercal x + b_j\right)}{\sum_{i=1}^{N}\exp\left(w_j^\intercal x + b_j\right)} = \dfrac{\exp\left(w_i^\intercal x + b_c\right)}{\sum_{i=1}^{N}\exp\left(w_j^\intercal x + b_j\right)}$

Thus, for the same one-hot encoded vector t and softmax vector z, $\boxed{\dfrac{\partial l(x,y)}{\partial b_c} = z - t}$

During SGD we seek to *minimize* the log-loss so we move in the direction of the negative gradient.

$$\boxed{W^{t+1} = W^t - \eta\frac{\partial l(x,y)}{\partial W} = W^t - \eta\left(x(z-t)^\intercal + \lambda W\right)}$$

$$\boxed{b^{t+1} = b^t - \eta\frac{\partial l(x,y)}{\partial b} = b^t - \eta(z-t)}$$