Problem 1. Calculate $\mathbb{E}_{X\sim Bernoulli(\pi)}[X]$.

Solution:

Bernoulli random variable will take value 1 or 0, and assuming that π is the probability that X takes the value 1, then

$$E_{X\sim Bernoulli(\pi)}[X] = 1 * Pr(X = 1) + 0 * Pr(X = 0) \tag{1}$$
$$= 1 * \pi = \pi \tag{2}$$

Note, if π is the probability that X takes value 0, then the answer would be $1 - \pi$. The usual convention is that we have $Bernouli(\pi)$ to mean that π is the probability that X takes value of 1.

Problem 2. Define a sigmoid function as $\sigma(z) = \frac{1}{1+e^{-z}}$. Prove that $\sigma(z) = 1 - \sigma(-z)$.

Solution:

The key is to recognize that $e^{-z}e^{z} = e^{-z+z} = e^{0} = 1$. Starting from the definition of sigmoid function that's given,

$$\sigma(z) = \frac{1}{1+e^{-z}} = \frac{1}{\frac{e^z}{e^z} + e^{-z}} \tag{3}$$
$$= \frac{1}{\frac{e^z}{e^z} + \frac{e^{-z}e^{z}}{e^z}} = \frac{1}{\frac{e^z + e^{-z+z}}{e^z}} \tag{4}$$
$$= \frac{e^z}{e^z+1} = 1 - \frac{1}{1+e^z}. \tag{5}$$

Problem 3. Let $t \in \{0, 1\}$ be the target of a binary classification problem and a logistic regression model be $y = \sigma(\boldsymbol{w}^\top \boldsymbol{x} + b)$. We can define two distributions $\boldsymbol{t}$ and $\boldsymbol{y}$ by

$$\boldsymbol{t} = \begin{pmatrix} 1-t \\ t \end{pmatrix}, \quad \boldsymbol{y} = \begin{pmatrix} 1-y \\ y \end{pmatrix}$$

Show that minimizing the cross-entropy loss $-t \log y - (1-t)\log(1-y)$ wrt $\boldsymbol{w}, b$ is equivalent to minimizing the Kullback--Leibler (KL) divergence $\mathrm{KL}(\boldsymbol{t}||\boldsymbol{y})$ wrt $\boldsymbol{w}, b$, where the KL divergence is defined as

$$\mathrm{KL}(\boldsymbol{t}||\boldsymbol{y}) = \sum_i t_i \log \frac{t_i}{y_i}$$

Solution:

Considering the cross-entropy loss $-t\log y - (1-t)\log(1-y)$, where $t \in \{0, 1\}$. Then, we can rewrite the cross-entropy loss as follows,

$$t \log \frac{1}{y} + (1-t)\log\frac{1}{1-y} \tag{6}$$

$$= t \log \frac{t}{y} + (1-t)\log\frac{1-t}{1-y}. \tag{7}$$

One can check that eq. (6) equals eq. (7) is because t is a constant (not a random variable) and can only take a value of 0 or 1. Also note that we take $t\log(t/y)$ to be 0 if $t = 0$ and likewise $(1-t)\log((1-t)/(1-y))$ to be 0 if $t = 1$. We also notice that eq. (7) is the same as the KL divergence between two Bernoulli distributions $\boldsymbol{t}$ and $\boldsymbol{y}$:

$$KL(\boldsymbol{t}||\boldsymbol{y}) = \sum_i t_i \log \frac{t_i}{y_i} = (1-t)\log\frac{1-t}{1-y} + t\log\frac{t}{y}. \tag{8}$$

Thus, minimizing the cross-entropy loss w.r.t $\boldsymbol{w}, b$ for a binary classification task is the same as minimizing the KL divergence w.r.t $\boldsymbol{w}, b$.

Problem 4. Prove that $[\sigma(z)]' = \sigma(z)(1-\sigma(z))$, where σ is the sigmoid function and $[\cdot]'$ represents the derivative.

<u>Solution:</u>

First note that

$$1-\sigma(z) = \frac{1+e^{-z}}{1+e^{-z}} - \frac{1}{1+e^{-z}} = \frac{e^{-z}}{1+e^{-z}}. \tag{9}$$

Taking the derivative of $\sigma(z)$ w.r.t z becomes

$$\begin{aligned} -(1+e^{-z})^{-2}[-e^{-z}] &= e^{-z}(1+e^{-z})^{-2} && (10)\\ &= \frac{1}{1+e^{-z}}\frac{e^{-z}}{1+e^{-z}} && (11)\\ &= \sigma(z)(1-\sigma(z)). && (12) \end{aligned}$$

END