Homework 2

Hoang Chu

February 12, 2024

Problem 1.

Solution:

1. Let $f(x) = 1 + e^{-x}$ and $A = \sigma(x)$. We have:

$$\begin{aligned}
\frac{dA}{dx} &= \frac{dA}{d[f(x)]} \cdot \frac{d[f(x)]}{dx} \\
&= \left(\frac{-1}{f(x)^2}\right) \cdot \left((-1)(e^{-x})\right) \\
&= \frac{e^{-x}}{(1+e^{-x})^2} \\
&= \frac{(1+e^{-x}) - 1}{(1+e^{-x})^2} \\
&= \frac{1}{1+e^{-x}} - \frac{1}{(1+e^{-x})^2} \\
&= \frac{1}{1+e^{-x}}(1 - \frac{1}{1+e^{-x}}) \\
&= A(1-A) = \sigma(x)(1-\sigma(x))
\end{aligned}$$

2. The negative log likelihood equation for logistic regression is:

$$L(\boldsymbol{\theta}) = -\sum_i y_i \log \sigma\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right) + (1-y_i)\log\left(1 - \sigma\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right)\right)$$

where I defined $\boldsymbol{\theta}$ being a column vector of weights, and $\mathbf{x}_i$ being a column vector. Taking the gradient of L with respect to $\boldsymbol{\theta}$:

$$\nabla_{\boldsymbol{\theta}^L} = -\sum_i y_i \frac{1}{\sigma\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right)} \sigma'\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right) + (1-y_i)\frac{1}{1-\sigma\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right)}\left(-\sigma'\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right)\right)$$

Applying (1) to σ:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}^L} &= -\sum_i y_i \left(1 - \sigma\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right)\right)\mathbf{x}_i - (1-y_i)\,\sigma\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right)\mathbf{x}_i \\
&= -\sum_i y_i\mathbf{x}_i - y_i\sigma\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right)\mathbf{x}_i - \sigma\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right)\mathbf{x}_i + y_i\sigma\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right)\mathbf{x}_i \\
&= \sum_i \left(\sigma\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right) - y_i\right)\mathbf{x}_i = \sum_i \left(\sigma\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right) - y_i\right)\mathbf{x}_i
\end{aligned}$$

The summation $= X^T(\boldsymbol{\sigma} - \mathbf{y})$, where $X = [\mathbf{x}_i \ldots]$ and $\boldsymbol{\sigma} - \mathbf{y}$ is a column vector of $\sigma\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right) - y_i$.

3. The Hessian Matrix is a matrix storing all second-order derivatives of the loss function with respect to pair-wise weights. Hence, we have:

$$\begin{aligned}
\mathbf{H}_{\boldsymbol{\theta}} &= \nabla_{\boldsymbol{\theta}} \left(\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})\right)^\top = \nabla_{\boldsymbol{\theta}} \left[X^\top(\boldsymbol{\mu} - \mathbf{y})\right]^\top \\
&= \nabla_{\boldsymbol{\theta}} \left(\boldsymbol{\mu}^\top X - \mathbf{y}^\top X\right) \\
&= \nabla_{\boldsymbol{\theta}} \boldsymbol{\mu}^\top X = \nabla_{\boldsymbol{\theta}} \sigma(X\boldsymbol{\theta})^\top X \\
&= X^\top \operatorname{diag}(\boldsymbol{\mu}(1-\boldsymbol{\mu})X \\
&= X^\top S X
\end{aligned}$$

To show that $\mathbf{H}_{\boldsymbol{\theta}}$ is positive semi-definite, it's equivalent to show that $\mathbf{S} = \operatorname{diag}(\boldsymbol{\mu}(1-\boldsymbol{\mu}))$ is positive semi-definite, which is equivalent to showing the diagonal entries of $\mathbf{S}$ are ≥ 0 (because all other non-diagonal entries are 0). Since $0 \leq \sigma\left(\boldsymbol{\theta}^\top\right) \leq 1$ by definition of a logistic classifier, $0 \leq 1 - \sigma\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right) \leq 1$. Hence:

$$\boldsymbol{\mu}_i \left(1 - \boldsymbol{\mu}_i\right) = \sigma\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right)\left(1 - \sigma\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right)\right) \geq 0$$

meaning the diagonal entries of $\mathbf{S}$ are ≥ 0.

Problem 2.

Solution: Since Gaussian random variable is a continuous random variable, it's CDF must summing up to 1 with bounds from $-\infty$ to ∞ i.e over the real number space. Therefore, we have:

$$\int_{\mathbb{R}} \mathbb{P}\left(x; \sigma^2\right) dx = \int_{\mathbb{R}} \frac{1}{Z} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx = \frac{1}{Z} \int_{\mathbb{R}} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx = 1$$

$$\leftrightarrow Z = \int_{\mathbb{R}} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx$$

Since x is just a random value, we can angular-ize with an introducing of y, without losing generality. Meaning:

$$\begin{aligned}
Z^2 &= \int_{\mathbb{R}} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx \int_{\mathbb{R}} \exp\left(-\frac{y^2}{2\sigma^2}\right) dy \\
&= \iint_{\mathbb{R}^2} \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) dxdy \\
&= \int_0^\infty \int_0^{2\pi} \exp\left(-\frac{r^2}{2\sigma^2}\right) r d\theta dr \\
&= 2\pi \int_0^\infty \exp\left(-\frac{r^2}{2\sigma^2}\right) r dr \\
&= 2\pi\left(-\sigma^2\right) \int_0^\infty \exp\left(-\frac{r^2}{2\sigma^2}\right)\left(-\frac{r}{\sigma^2}\right) dr \\
&= -\ 2\pi\sigma^2 \exp\left(-\frac{r^2}{2\sigma^2}\right)\Bigg|_0^\infty \\
&= -2\pi\sigma^2(0-1) = 2\pi\sigma^2
\end{aligned}$$

Thus, our constant Z is: $Z = \sqrt{2\pi\sigma^2} = \sqrt{2\pi}\sigma$

Problem 3.

Solution:

1. Given that:

$$max_w = \arg\max_{\mathbf{w}} \sum_{i=1}^{N} \log \mathcal{N}\left(y_i \mid w_0 + \mathbf{w}^\top \mathbf{x}_i, \sigma^2\right) + \sum_{j=1}^{D} \log \mathcal{N}\left(w_j \mid 0, \tau^2\right)$$

and we knew from problem 2 that:

$$\mathcal{N}(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Replacing μ with $w_0 + \mathbf{w}^\top \mathbf{x}_i$ and σ with τ, we have:

$$max_w = \arg\max_{\mathbf{w}} \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y_i - w_0 - \mathbf{w}^\top \mathbf{x}_i\right)^2}{2\sigma^2}\right) + \sum_{j=1}^{D} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{w_j^2}{2\tau^2}\right)$$

Since $log\left(\frac{X}{Y}\right) = log(X) - log(Y)$, we have:

$$\leftrightarrow max_w = \arg\max_{\mathbf{w}} \sum_{i=1}^{N} \left(-\frac{\left(y_i - w_0 - \mathbf{w}^\top \mathbf{x}_i\right)^2}{2\sigma^2} - \log\sqrt{2\pi}\sigma\right) + \sum_{j=1}^{D} \left(-\frac{w_j^2}{2\tau^2} - \log\sqrt{2\pi}\sigma\right)$$

$$\leftrightarrow max_w = \arg\max_{\mathbf{w}} -\left((N+D)\log\sqrt{2\pi}\sigma + \sum_{i=1}^{N} \frac{\left(y_i - w_0 - \mathbf{w}^\top \mathbf{x}_i\right)^2}{2\sigma^2} + \sum_{j=1}^{D} \frac{w_j^2}{2\tau^2}\right)$$

Because the constant $-(N+D)\log\sqrt{2\pi}\sigma$ does not affect or change our optimal value $\mathbf{w}^\star$, and we can similarly scale our problem by $2\sigma^2$ without changing $\mathbf{w}^\star$, we can ignore the constant and rescale the problem. In addition, since maximizing a function is equivalent to minimizing its negative, we now arrive at the equivalent optimization, we have:

$$max_w = \arg\min_{\mathbf{w}} \sum_{i=1}^{N} \left(y_i - w_0 - \mathbf{w}^\top \mathbf{x}_i\right)^2 + \frac{\sigma^2}{\tau^2} \sum_{j=1}^{D} w_j^2$$

Let $\lambda = \sigma^2/\tau^2$, we have:

$$max_w = \arg\min_{\mathbf{w}} \sum_{i=1}^{N} \left(y_i - w_0 - \mathbf{w}^\top \mathbf{x}_i\right)^2 + \lambda \sum_{j=1}^{D} w_j^2$$

$$\leftrightarrow max_w \arg\min_{\mathbf{w}} \sum_{i=1}^{N} \left(y_i - w_0 - \mathbf{w}^\top \mathbf{x}_i\right)^2 + \lambda \sum_{j=1}^{D} w_j^2$$

which is our desired ridge regression function.

2. To find the closed form solution $\mathbf{x}$ to the ridge regression problem:

$$\text{minimize} : f = \|A\mathbf{x} - \mathbf{b}\|_2^2 + \|\Gamma\mathbf{x}\|_2^2.$$

We want to find the gradient of f with respect to $\mathbf{x}$ and set it to 0:

$$\begin{aligned}
\nabla_{\mathbf{x}} f &= \nabla_{\mathbf{x}} \left[(A\mathbf{x} - \mathbf{b})^\top (A\mathbf{x} - \mathbf{b}) + (\Gamma\mathbf{x})^\top (\Gamma\mathbf{x})\right] \\
&= \nabla_{\mathbf{x}} \left[\left(\mathbf{x}^\top A^\top - \mathbf{b}^\top\right)(A\mathbf{x} - \mathbf{b}) + \mathbf{x}^\top \Gamma^\top \Gamma \mathbf{x}\right] \\
&= \nabla_{\mathbf{x}} \left[\mathbf{x}^\top A^\top A\mathbf{x} - 2\mathbf{x}^\top A^\top \mathbf{b} + \mathbf{b}^\top \mathbf{b} + \mathbf{x}^\top \Gamma^\top \Gamma \mathbf{x}\right] \\
&= 2A^\top A\mathbf{x} - 2A^\top \mathbf{b} + 2\Gamma^\top \Gamma \mathbf{x}
\end{aligned}$$

Set $\nabla_{\mathbf{x}} f = 0$ gives us:

$$\left(A^\top A + \Gamma^\top \Gamma\right)\mathbf{x} = A^\top \mathbf{b}$$

Therefore, the closed form solution is:

$$\mathbf{x} = \left(A^\top A + \Gamma^\top \Gamma\right)^{-1} A^\top \mathbf{b}$$

If we let $\Gamma = \sqrt{\lambda}\mathbf{I}$, then we can see this gives an objective of the form:

$$\text{minimize} : f = \|A\mathbf{x} - \mathbf{b}\|_2^2 + \lambda \mathbf{x}^\top \mathbf{x}$$

with the closed form optimal solution:

$$\mathbf{x} = \left(A^\top A + \lambda\mathbf{I}\right)^{-1} A^\top \mathbf{b}$$

3. See code and images in github repo.

4. The objective function is:

$$f = \|A\mathbf{x} + b\mathbf{1} - \mathbf{y}\|_2^2 + \|\Gamma\mathbf{x}\|_2^2$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{1}$ is vector of all ones and $\mathbf{y} \in \mathbf{R}^n$.

$$\begin{aligned}
\leftrightarrow f &= \|A\mathbf{x} + b\mathbf{1} - \mathbf{y}\|_2^2 + \|\Gamma\mathbf{x}\|_2^2 \\
&= (A\mathbf{x} + b\mathbf{1} - \mathbf{y})^\top (A\mathbf{x} + b\mathbf{1} - \mathbf{y}) + (\Gamma\mathbf{x})^\top (\Gamma\mathbf{x}) \\
&= \left(\mathbf{x}^\top A\top + b\mathbf{1}^\top - \mathbf{y}^\top\right)(A\mathbf{x} + b\mathbf{1} - \mathbf{y}) + \mathbf{x}^\top \Gamma^\top \Gamma \mathbf{x} \\
&= \mathbf{x}^\top A\top A\mathbf{x} + 2b\mathbf{1}^\top A\mathbf{x} - 2\mathbf{y}^\top A\mathbf{x} - 2b\mathbf{1}^\top \mathbf{y} + b^2 n + \mathbf{y}^\top \mathbf{y} + \mathbf{x}^\top \Gamma^\top \Gamma \mathbf{x}
\end{aligned}$$

To find the arguments that minimize the output, we take the gradients w.r.t to the interested arguments and set them to 0:

$$\nabla_{\mathbf{x}} f = 2A^\top A\mathbf{x} + 2bA^\top \mathbf{1} - 2A^\top \mathbf{y} + 2\Gamma^\top \Gamma \mathbf{x} = 0 \quad \nabla_b f = 2\mathbf{1}^\top A\mathbf{x} - 2\mathbf{1}^\top \mathbf{y} + 2bn = 0$$

Solving for b gives us:

$$b = \frac{\mathbf{1}^\top (\mathbf{y} - A\mathbf{x})}{n}$$

Plugging the result of b back to equation to solve for $\mathbf{x}$, we have:

$$
\begin{aligned}
\left(A^\top A + \Gamma^\top \Gamma\right) \mathbf{x} + \left(\frac{\mathbf{1}^\top(\mathbf{y} - A\mathbf{x})}{n}\right) A^\top \mathbf{1} - A^\top \mathbf{y} = 0 \\
\left(A^\top A + \Gamma^\top \Gamma\right) \mathbf{x} + \frac{1}{n} A^\top \mathbf{1}\mathbf{1}^\top \mathbf{y} - \frac{1}{n} A^\top \mathbf{1}\mathbf{1}^\top A\mathbf{x} - A^\top \mathbf{y} = 0 \\
\left[A^\top A + \Gamma^\top \Gamma - \frac{1}{n} A^\top \mathbf{1}\mathbf{1}^\top A\right] \mathbf{x} = A^\top \mathbf{y} - \frac{1}{n} A^\top \mathbf{1}\mathbf{1}^\top \mathbf{y} \\
\left[A^\top \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right) A + \Gamma^\top \Gamma\right] \mathbf{x} = A^\top \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right) \mathbf{y} \\
\mathbf{x} = \left[A^\top \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right) A + \Gamma^\top \Gamma\right]^{-1} A^\top \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right) \mathbf{y}
\end{aligned}
$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{1}$ is vector of all ones and $\mathbf{y} \in \mathbf{R}^n$.

5. See code and images in github repo.