KDAG tasks

Anmol Kumar (20HS20010)

July 24, 2021

1 First subtask

Problem: Find the Hessian matrix H of the empirical loss function with respect to θ, and show that the Hessian H is positive semi-definite in nature.

Solution: The empirical loss function is given as,

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)}\log(h_\theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\theta(x^{(i)})) \tag{1}$$

Now defining $J(\theta)\cdot m$ as $L(\theta)$ and writing the loss expression for a single entry i, we get,

$$L_{(i)}(\theta) = -[y^{(i)}\log(h_\theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))] \tag{2}$$

The Hessian matrix H is composed of double derivatives of each of the elements, thus we find the double derivatives of the general element $L_{(i)}(\theta)$ we just calculated above. Note that we first differentiate w.r.t θ^T and then θ to get our desired value $\nabla_{\theta\theta^T} L_{(i)}(\theta)$

Now before we proceed further, we need to keep in mind an important relation, i.e,

$$\frac{\partial\sigma(z)}{\partial z} = \sigma(z)(1-\sigma(z)) \tag{3}$$

where $\sigma(z)$ is defined as,

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

In eq(2), our points of interest are $log(h_\theta(x^{(i)}))$ and $log(1 - h_\theta(x^{(i)}))$, because the rest of eq(2) are constants. Labeling them as L and M respectively, we first find the gradient of these two functions.

$$\begin{aligned}
\frac{\partial L}{\partial \theta^T} &= \frac{\partial \log h_\theta(x^{(i)})}{\partial \theta^T} \\
&= \frac{\partial \log \sigma(\theta^T x^{(i)})}{\partial \theta^T} \\
&= \frac{\partial \log \sigma(\theta^T x^{(i)})}{\partial \sigma(\theta^T x^{(i)})} \cdot \frac{\partial \sigma(\theta^T x^{(i)})}{\partial (\theta^T x^{(i)})} \cdot \frac{\partial (\theta^T x^{(i)})}{\partial \theta^T} \\
&= \frac{1}{\cancel{\sigma(\theta^T x^{(i)})}} \cdot \cancel{\sigma(\theta^T x^{(i)})}(1 - \sigma(\theta^T x^{(i)})) \cdot x^{(i)} \\
&= (1 - \sigma(\theta^T x^{(i)}))x^{(i)}
\end{aligned} \tag{5}$$

Similarly for M, we get,

$$\begin{aligned}
\frac{\partial M}{\partial \theta^T} &= \frac{\partial \log(1 - h_\theta(x^{(i)}))}{\partial \theta^T} \\
&= \frac{\partial \log(1 - \sigma(\theta^T x^{(i)}))}{\partial \theta^T} \\
&= \frac{\partial \log(1 - \sigma(\theta^T x^{(i)}))}{\partial \sigma(\theta^T x^{(i)})} \cdot \frac{\partial \sigma(\theta^T x^{(i)})}{\partial (\theta^T x^{(i)})} \cdot \frac{\partial (\theta^T x^{(i)})}{\partial \theta^T} \\
&= \frac{-1}{\cancel{1 - \sigma(\theta^T x^{(i)})}} \cdot \sigma(\theta^T x^{(i)})\cancel{(1 - \sigma(\theta^T x^{(i)}))} \cdot x^{(i)} \\
&= -\sigma(\theta^T x^{(i)})x^{(i)}
\end{aligned} \tag{6}$$

Putting all the values obtained, we get

$$\nabla_{\theta^T} L_{(i)}(\theta) = x^{(i)}(\sigma(\theta^T x^{(i)}) - y^{(i)}) \tag{7}$$

Evaluating further,

$$\begin{aligned}
\nabla_{\theta\theta^T} L_{(i)}(\theta) &= \frac{\partial^2 L_{(i)}(\theta)}{\partial \theta \partial \theta^T} \\
&= \frac{\partial \nabla_{\theta^T} L_{(i)}(\theta)}{\partial \theta} \\
&= \frac{\partial x^{(i)}(\sigma(\theta^T x^{(i)}) - y^{(i)})}{\partial \theta} \\
&= x^{(i)}[x^{(i)}]^T \sigma(\theta^T x^{(i)})(1 - \sigma(\theta^T x^{(i)}))
\end{aligned} \tag{8}$$

Thus, the Hessian matrix for $L_{(i)}(\theta)$ is given by the above expression. We can now find the Hessian matrix for our original empirical loss function $J_{(i)}(\theta)$

$$\begin{aligned} L_{(i)}(\theta) &= m \cdot J_{(i)}(\theta) \\ \Rightarrow \nabla^2 L_{(i)}(\theta) &= m \cdot \nabla^2 J_{(i)}(\theta) \\ \Rightarrow \nabla^2 J_{(i)}(\theta) &= \frac{1}{m} \cdot x^{(i)}[x^{(i)}]^T \sigma(\theta^T x^{(i)})(1 - \sigma(\theta^T x^{(i)})) \end{aligned} \tag{9}$$

Note that the quantity $\sigma(\theta^T x^{(i)})(1 - \sigma(\theta^T x^{(i)}))$ is *always* > 0 as $\sigma(z) \in (0, 1)$.

Considering each entry to be composed of n features, we take X as a matrix of dimensions $n \times m$, where every column represents $x^{(i)}$, which is the vector corresponding to a given entry, and every row represents a feature of that entry. Formally, $\sum_{i=1}^{m} x^{(i)}[x^{(i)}]^T = XX^T$ The number of columns is m, as its the number of entries for the particular data-set. For the factor of probability, we define a diagonal matrix D of size $m \times m$, with D_{ii} as $\frac{1}{m}\sigma(\theta^T x^{(i)})(1 - \sigma(\theta^T x^{(i)}))$ for each set of inputs.

Therefore, using X and D, we finally define our Hessian H as,

$$H(\theta) = XDX^T \tag{10}$$

To prove that H is a positive semi-definite matrix, we need to show that the quantity $z^T Hz$, a scalar, is positive, where z is any arbitrary matrix of dimensions $1 \times n$, where n is the number of features.

$$z^T Hz = z^T XDX^T z \quad = (z^T X)D(z^T X)^T \tag{11}$$

Since D is a positive contributing entity and $z^T X$ is being multiplied with itself, the whole scalar turns out to be non-negative.

Hence, the Hessian matrix H has been proved to be positive semi-definite in nature.

2 Third subtask

Problem: In order to show that Gaussian Discriminant Analysis results in a classifier that has a linear decision boundary, show that the following expression is true.

$$p(y=1|x;\phi,\mu_0,\mu_1,\Sigma)=\frac{1}{1+\exp(-(\theta^T x+\theta_0))} \tag{12}$$

where $\theta \in \Re^n$ and $\theta_0 \in \Re$ are appropriate functions of ϕ, μ_0, μ_1 and Σ.

Solution: To show that the classifier has a linear decision boundary, we need to show that it has behavior like that of a logistic regression, as LR makes use of a straight line to divide the whole dataset into two classes.

Before we begin, we list the formulas that we have be given,

$$p(y)=\begin{cases}\phi & \text{if } y=1\\ 1-\phi & \text{if } y=0\end{cases} \tag{13}$$

$$p(x|y=0)=\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right) \tag{14}$$

$$p(x|y=1)=\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right) \tag{15}$$

To keep the calculations straightforward, we denote $p(y=1|x;\phi,\mu_0,\mu_1,\Sigma)$ as $p(y=1|x)$. From Bayes theorem we know that,

$$p(y|x)=\frac{p(x|y)p(y)}{p(x)} \tag{16}$$

where, $p(x)=p(x|y=1)p(y=1)+p(x|y=0)p(y=0)$.

Putting the value of $p(x)$ and using eq(16) for finding $p(y=1|x)$, we get,

$$\begin{aligned}p(y=1|x)&=\frac{p(x|y=1)p(y=1)}{p(x|y=1)p(y=1)+p(x|y=0)p(y=0)}\\&=\frac{1}{1+\frac{p(x|y=0)}{p(x|y=1)}\frac{p(y=0)}{p(y=1)}}\end{aligned} \tag{17}$$

Using the values of $p(x|y=0)$, $p(x|y=1)$ and $p(y)$ from eq(13) to eq(15) we get,

$$=\frac{1}{1+\exp(Z)(\frac{1-\phi}{\phi})}=\frac{1}{1+\exp(Z')} \tag{18}$$

where Z' is defined as,

$$\begin{aligned}
Z' &= -\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0) + \frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) + \log(\frac{1-\phi}{\phi}) \\
&= -\frac{1}{2}(x^T-\mu_0^T)\Sigma^{-1}(x-\mu_0) + \frac{1}{2}(x^T-\mu_1^T)\Sigma^{-1}(x-\mu_1) + \log(\frac{1-\phi}{\phi})
\end{aligned} \tag{19}$$

As μ_0, μ_1, x have the same dimensions i.e $n \times 1$ and the covariance matrix Σ has dimensions $n \times n$, $(\mu_0-\mu_1)^T\Sigma^{-1}x = x^T\Sigma^{-1}(\mu_0-\mu_1)$. Using this expression and simplifying, we finally get,

$$\begin{aligned}
Z' &= -(\mu_1-\mu_0)^T\Sigma^{-1}x + [\log(\frac{1-\phi}{\phi}) + \frac{1}{2}(\mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0)] \\
&= -(\theta^T x + \theta_0)
\end{aligned} \tag{20}$$

Note that the terms in the square bracket simplifies to a scalar $-\theta_0$ and $\theta^T = (\mu_1-\mu_0)^T\Sigma^{-1}x$. Putting this in eq(18), we finally get our desired expression,

$$p(y=1|x;\phi,\mu_0,\mu_1,\Sigma) = \frac{1}{1+\exp\left(-(\theta^T x + \theta_0)\right)} \tag{21}$$

As discussed earlier, this proves that GDA takes the form of logistic regression and hence, has a linear decision boundary.