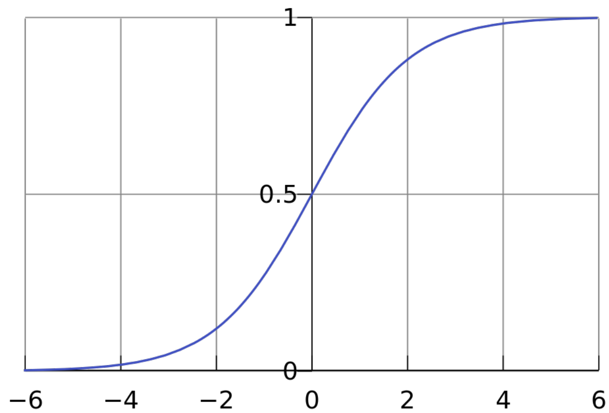
Logistic regression

[redacted]



Question 1

The sigmoid function is defined as

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

Prove that

1. $\sigma(x) + \sigma(-x) = 1$
2. $\frac{d}{dx}\sigma(x) = \sigma'(x) = \sigma(x)(1 - \sigma(x))$
3. $\sigma(x) = \frac{d}{dx}\text{softPlus(x)} = \frac{\mathrm{d}}{\mathrm{dx}} \log(1 + \mathrm{e}^{\mathrm{x}})$

Answer

1. This property can be seen graphically from the function plot and proven using the definition of σ given above

$$\begin{aligned}
\sigma(x) + \sigma(-x) &= \frac{1}{1+e^{-x}} + \frac{1}{1+e^{x}} && (1)\\
&= \frac{1}{1+e^{-x}} + \frac{e^{-x}}{1+e^{-x}} && (2)\\
&= \frac{1+e^{-x}}{1+e^{-x}} && (3)\\
&= 1 && (4)
\end{aligned}$$

2. Using the standard rules for the derivation of a composite function, $(g(f(x)))' = g'(f(x))f'(x)$, we have

$$\begin{aligned}
\frac{d}{dx}\sigma(x) &= \frac{d}{dx}\frac{1}{1+e^{-x}} && (5)\\
&= \frac{-1}{(1+e^{-x})^2}\frac{d}{dx}(1+e^{-x}) && (6)\\
&= \frac{e^{-x}}{(1+e^{-x})^2} && (7)\\
&= \sigma(x)\frac{1}{(1+e^{-x})} && (8)\\
&= \sigma(x)\sigma(-x) && (9)\\
&= \sigma(x)(1-\sigma(x)) && (10)
\end{aligned}$$

where in the last step we use $\sigma(x)+\sigma(-x)=1$.

3. The derivative of softPlus(x) is also obtained form standard derivation rules as follows

$$\begin{aligned}
\frac{d}{dx}\text{softPlus}(x) &= \frac{d}{dx}\log(1+e^{x}) && (11)\\
&= \frac{1}{(1+e^{x})}\frac{d}{dx}(1+e^{x}) && (12)\\
&= \frac{e^{x}}{(1+e^{x})} && (13)\\
&= \sigma(x) && (14)
\end{aligned}$$

Question 2

Consider a binary-classification task with label $Y \in \{0,1\}$. You are given a dataset

$$\mathcal{D} = \{(x_n, y_n) \in \mathcal{X} \times \{0,1\}\}_{n=1}^{N}$$

for training a *logistic regression* learning machine $F : \mathcal{X} \times \Lambda \to [0,1]$ defined as

$$\text{Prob}(Y=1) = F(X,\lambda) = \sigma([1, X^T]\lambda)$$

1. Compute the gradient of the negative log-likelihood of the model on $\mathcal{D}$ i.e. the gradient of

$$\ell = -\sum_{i=1}^{N}\log\left(y_n F([1,x_n^T]^T,\lambda) + (1-y_n)(1-F([1,x_n^T]^T,\lambda))\right)$$

as a function of λ.

2. Let

$$\mathcal{D} = \{([1,1]^T,1), ([1,2]^T,0), ([2,2]^T,1), ([2,1]^T,1)\}$$

and compute $\ell = \ell(\mathcal{D}, \lambda^{(0)})$ for $\lambda^{(0)} = [\frac{1}{3}, \frac{1}{3}, -\frac{1}{3}]^T$.

3. Evaluate the gradient of ℓ at $\lambda^{(0)} = [\frac{1}{3}, \frac{1}{3}, -\frac{1}{3}]^T$.
4. Compute the first *gradient descent update* of $\lambda^{(0)}$ with learning rate $\eta = \frac{5}{60}$
5. Check that $\ell(\mathcal{D}, \lambda^{(0)}) \geq \ell(\mathcal{D}, \lambda^{(1)})$

You can use the approximations given in Appendix A.

Answer

1. The gradient of ℓ at λ is

$$\nabla_\lambda \ell = -\sum_{i=1}^{N} c_n \nabla_\lambda F_n \tag{15}$$

$$c_n = \frac{y_n - (1 - y_n)}{y_n F_n + (1 - y_n)(1 - F_n)} \tag{16}$$

$$\nabla_\lambda F_n = F_n(1 - F_n)[1, x_n^T]^T \tag{17}$$

$$F_n = F([1, x_n^T]^T, \lambda) = \lambda_0 + \sum_{i=1}^{d} x_{ni}\lambda_i \tag{18}$$

where d is the number of attributes, e.g. $d = 2$ in this case.

2. To compute $\ell(\mathcal{D}, \lambda)$ at $\lambda = \lambda^{(0)}$, we need

$$F_1 = \sigma(\frac{1}{3}) \approx 0.58 \tag{19}$$

$$F_2 = \sigma(0) = 1 - \sigma(0) = 0.5 \tag{20}$$

$$F_3 = \sigma(\frac{1}{3}) \approx 0.58 \tag{21}$$

$$F_4 = \sigma(\frac{2}{3}) \approx 0.66 \tag{22}$$

$$\tag{23}$$

and obtain

$$\ell(\mathcal{D}, \lambda^{(0)}) = -(\log F_1 + \log(1 - F_2) + \log F_3 + \log F + 4) \tag{24}$$

$$= -(2 \log 0.58 + \log 0.5 + \log 0.66) \tag{25}$$

$$= (2 * 0.54 + 0.69 + 0.41) \tag{26}$$

$$= 2.18 \tag{27}$$

3. The gradient of ℓ is the sum of four terms, one for each data point in $\mathcal{D}$.

Letting $\lambda = \lambda^{(0)}$, we have

$$-c_1 \nabla F_1 = -\frac{1}{\sigma(\frac{1}{3})} \sigma(\frac{1}{3})(1 - \sigma(\frac{1}{3})[1,1,1]^T \tag{28}$$

$$= -(1 - \sigma(\frac{1}{3})[1,1,1]^T \approx -0.42[1,1,1]^T \tag{29}$$

$$-c_2 \nabla F_2 = -\frac{1}{1 - \sigma(0)} \sigma(0)(1 - \sigma(0)[1,1,2]^T \tag{30}$$

$$= \sigma(0)[1,1,2]^T = 0.5[1,1,2]^T \tag{31}$$

$$-c_3 \nabla F_3 = -\frac{1}{\sigma(\frac{1}{3})} \sigma(\frac{1}{3})(1 - \sigma(\frac{1}{3})[1,2,2]^T \tag{32}$$

$$= -(1 - \sigma(\frac{1}{3})[1,2,2]^T \approx -0.42[1,2,2]^T \tag{33}$$

$$-c_4 \nabla F_4 = -\frac{1}{\sigma(\frac{2}{3})} \sigma(\frac{2}{3})(1 - \sigma(\frac{2}{3})[1,2,1]^T \tag{34}$$

$$= -(1 - \sigma(\frac{2}{3})[1,2,1]^T \approx -0.34[1,2,1]^T \tag{35}$$

We obtain $\nabla \ell = [\frac{\partial \ell}{\partial \lambda_0}, \frac{\partial \ell}{\partial \lambda_1}, \frac{\partial \ell}{\partial \lambda_2}]^T$ by summing over all terms. In particular, we have

$$[\nabla \ell]_0 = -(2 * 0.42 - 0.5 + 0.34) = -0.68 \tag{36}$$

$$[\nabla \ell]_1 = -(3 * 0.42 - 0.5 + 2 * 0.34) = -1.44 \tag{37}$$

$$[\nabla \ell]_2 = -(3 * 0.42 - 2 * 0.5 + 0.34) = -0.60 \tag{38}$$

or, equivalently,

$$\nabla_\lambda \ell(\mathcal{D}, \lambda^{(0)}) = -[0.68, 1.44, 0.60]^T$$

4. The first gradient descent update, $\lambda^{(1)}$, is

$$\lambda^{(1)} = \lambda^{(0)} - \eta \nabla_\lambda \ell(\mathcal{D}, \lambda^{(0)}) \tag{39}$$

$$\approx [\frac{1}{3}, \frac{1}{3}, -\frac{1}{3}]^T + \frac{1}{6}[0.7, 1.4, 0.6]^T \tag{40}$$

$$\approx \frac{1}{30}[10 + 4, 10 + 7, -10 + 3]^T \tag{41}$$

$$= \frac{1}{30}[14, 17, -7]^T \tag{42}$$

$$\approx [\frac{1}{3}, \frac{2}{3}, -\frac{1}{3}] \tag{43}$$

where we have set $\eta = \frac{1}{6}$.

5. Plugging $\lambda^{(1)}$ into the learning machine we obtain the updated prediction

probabilities, i.e.

$$F_1 = \sigma(\frac{2}{3}) \approx 0.66 \tag{44}$$

$$F_2 = \sigma(\frac{1}{3}) \approx 0.58 \tag{45}$$

$$F_3 = \sigma(1) \approx 0.73 \tag{46}$$

$$F_4 = \sigma(\frac{4}{3}) \approx 0.79 \tag{47}$$

$$\tag{48}$$

Finally, we obtain

$$\begin{aligned}
\ell(\mathcal{D}, \lambda^{(0)}) &= -\left(\log F_1 + \log(1 - F_2) + \log F_3 + \log F_4\right) && (49)\\
&= -\left(\log 0.66 + \log 0.42 + \log 0.73 + \log 0.79\right) && (50)\\
&= (0.41 + 0.86 + 0.31 + 0.23) && (51)\\
&= 1.81 && (52)
\end{aligned}$$

that shows $\ell(\mathcal{D}, \ell^{(1)}) \leq \ell(\mathcal{D}, \ell^{(0)})$.

A Approximations

$$\sigma(0) = 0.5 \tag{53}$$
$$\sigma(\frac{1}{6}) \approx 0.54 \tag{54}$$
$$\sigma(\frac{1}{3}) \approx 0.58 \tag{55}$$
$$\sigma(\frac{1}{2}) \approx 0.62 \tag{56}$$
$$\sigma(\frac{2}{3}) \approx 0.66 \tag{57}$$
$$\sigma(0.8) \approx 0.68 \tag{58}$$
$$\sigma(0.9) \approx 0.71 \tag{59}$$
$$\sigma(1) \approx 0.73 \tag{60}$$
$$\sigma(\frac{4}{3}) \approx 0.79 \tag{61}$$
$$\sigma(1.3) \approx 0.78 \tag{62}$$
$$\sigma(1.4) \approx 0.80 \tag{63}$$
$$\log(0.32) \approx -1.13 \tag{64}$$
$$\log(0.5) \approx -0.69 \tag{65}$$
$$\log(0.54) \approx -0.61 \tag{66}$$
$$\log(0.58) \approx -0.54 \tag{67}$$
$$\log(0.62) \approx -0.47 \tag{68}$$
$$\log(0.66) \approx -0.41 \tag{69}$$
$$\log(0.71) \approx -0.34 \tag{70}$$
$$\log(0.73) \approx -0.31 \tag{71}$$
$$\log(0.78) \approx -0.24 \tag{72}$$
$$\log(0.79) \approx -0.23 \tag{73}$$
$$\log(0.80) \approx -0.22 \tag{74}$$