Optimizing model likelihood using gradient descent

0.1 Linear regression

Input Data:

$$\text{Independent:} \quad \vec{x}^{(i)} \tag{1}$$
$$\text{Dependant:} \quad y^{(i)} \tag{2}$$

Parameters to learn

$$\text{Weights:} \quad \vec{w} \tag{3}$$
$$\text{Bias:} \quad b \tag{4}$$

Model

$$\widehat{y}^{(i)} \sim \mathcal{N}(\vec{w}^T \vec{x}^{(i)} + b, \sigma) \tag{5}$$

Error to minimize is negative log likelihood (assuming Gaussian)

$$E^{(i)} = -log\ p(y^{(i)}|\vec{x}^{(i)}) \tag{6}$$

$$E^{(i)} = -log\ \mathcal{N}(y^{(i)} \mid \vec{w}^T\vec{x}^{(i)} + b, \sigma) \tag{7}$$

$$E = -\sum_{i=1}^{n} log\ p(y^{(i)}|\vec{x}^{(i)}) \tag{8}$$

$$\vec{w}, b = \arg\min_{\vec{w},b} -\sum_{i=1}^{n} log\ \mathcal{N}(y^{(i)} \mid \vec{w}^T\vec{x}^{(i)} + b, \sigma) \tag{9}$$

$$\vec{w}, b = \arg\min_{\vec{w},b} -\sum_{i=1}^{n} log\ \frac{1}{\sqrt{2\sigma^2\pi}} \exp -\frac{(y^{(i)} - \vec{w}^T\vec{x}^{(i)} - b)^2}{\sigma^2} \tag{10}$$

$$\vec{w}, b = \arg\min_{\vec{w},b} \sum_{i=1}^{n} (y^{(i)} - \vec{w}^T\vec{x}^{(i)} - b)^2 \tag{11}$$

$$\frac{\partial E}{\partial \vec{w}} = \frac{2}{n}\sum_{i=1}^{n}(\vec{w}^T\vec{x}^{(i)} + b - y^{(i)})\vec{x}^{(i)} \tag{12}$$

$$\frac{\partial E}{\partial \vec{w}} = \frac{2}{n}\sum_{i=1}^{n}(\widehat{y}^{(i)} - y^{(i)})\vec{x}^{(i)} \tag{13}$$

$$\frac{\partial E}{\partial b} = \frac{2}{n}\sum_{i=1}^{n}(\widehat{y}^{(i)} - y^{(i)}) \tag{14}$$

0.2 Logistic regression

Input Data:

$$\text{Features:} \quad \vec{x}^{(i)} \tag{15}$$

$$\text{Labels:} \quad y^{(i)} \in \{1, 0\} \tag{16}$$

Parameters to learn

$$\text{Weights:} \quad \vec{w} \tag{17}$$

$$\text{Bias:} \quad b \tag{18}$$

Derived Features

$$z^{(i)} = \vec{w}^T \vec{x}^{(i)} + b \tag{19}$$
$$p^{(i)} = \sigma(z^{(i)}) \tag{20}$$
$$p^{(i)} = p(y^{(i)} = 1|\vec{x}^{(i)}) = \sigma(z^{(i)}) \tag{21}$$
$$\sigma(a) = \frac{1}{1 + e^{-a}} \tag{22}$$

Likelihood of y comes from a Bernouli process parameterized by $p^{(i)}$

$$p(y^{(i)}|\vec{x}^{(i)}) = (p^{(i)})^{y^{(i)}}(1 - p^{(i)})^{(1-y^{(i)})} \tag{23}$$
$$p(y|\vec{x}) = \prod_{i=1}^{n} p(y^{(i)}|\vec{x}^{(i)}) \tag{24}$$

Error to minimize is negative log likelihood

$$E^{(i)} = -(y^i log\ p^{(i)} + (1 - y^i) log(1 - p^{(i)})) \tag{25}$$

Use the chain rule to optimize $\vec{w}$ and b

$$\frac{\partial E^{(i)}}{\partial \vec{w}} = \frac{\partial E^{(i)}}{\partial p^{(i)}} \frac{\partial p^{(i)}}{\partial z^{(i)}} \frac{\partial z^{(i)}}{\partial \vec{w}} \tag{26}$$
$$\frac{\partial E^{(i)}}{\partial p^{(i)}} = -\frac{y^{(i)}}{p^{(i)}} + \frac{(1 - y^{(i)})}{(1 - p^{(i)})} \tag{27}$$
$$\frac{\partial p^{(i)}}{\partial z^{(i)}} = p^{(i)}(1 - p^{(i)}) = \sigma(z^{(i)})(1 - \sigma(z^{(i)})) \tag{28}$$
$$\frac{\partial z^{(i)}}{\partial \vec{w}} = \vec{x}^{(i)} \tag{29}$$
$$\frac{\partial E^{(i)}}{\partial \vec{w}} = (p^{(i)} - y^{(i)})\vec{x}^{(i)} \tag{30}$$
$$\frac{\partial E^{(i)}}{\partial b} = \frac{\partial E^{(i)}}{\partial p^{(i)}} \frac{\partial p^{(i)}}{\partial z^{(i)}} \frac{\partial z^{(i)}}{\partial b} \tag{31}$$
$$\frac{\partial z^{(i)}}{\partial b} = 1 \tag{32}$$
$$\frac{\partial E^{(i)}}{\partial b} = (p^{(i)} - y^{(i)}) \tag{33}$$

In gradient descent for batch B our update rule is

$$\vec{w}' = \vec{w} - \epsilon \frac{1}{|B|} \sum_{i \in B} \frac{\partial E^{(i)}}{\partial \vec{w}} \tag{34}$$

$$b' = b - \epsilon \frac{1}{|B|} \sum_{i \in B} \frac{\partial E^{(i)}}{\partial b} \tag{35}$$