Logistic Regression

given $\{(x_n, y_n)\}_{n=1}^{N}$, $x_n \in \mathbb{R}^{D_x}$, $y_n \in \mathbb{Z}^{D_y}$

$$f(x) = \sigma(\underbrace{xw + b}_{\text{linear model}}) = \hat{y}$$

activation function for discrete real values

Logistic sigmoid activation Func

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

Sigmoid squashes data between 0 to 1

Bernouli prob density funct = 0 or 1

Liklihood based on Bernouli

$$P(Y|X,w) = \prod P(Y|x_n)^{y_n}(1-P(Y|x_n))^{1-y_n}$$

Pred, actual

$$= \prod f(x_n)^{y_n}(1-f(x_n))^{1-y_n}$$

Optimizing involves max liklihood of $P(Y|X,w)$

which is the same as Min the neg log liklihood (Min is always easier)

recall: $\frac{\partial}{\partial w}(-\log(P(Y|X,w))) = 0$

Sub in the ~~ER formula~~ Bernouli liklihood formula Distribute & apply logarithm rules

the result is a function of a function ... which also contains the w we want to Diff

Solution is to open up the Derivative Function

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial f(\Theta)} \cdot \frac{\partial f(\Theta)}{\partial \Theta} \cdot \frac{\partial \Theta}{\partial w}$$

" Applying rule of derivative of a function of a function"
This is the chain rule

- Expand & simplify
- Cancel common term to num & Den
- Divide by -1

Final state,

$$\sum x_n (f(x_n) - y_n) = 0$$

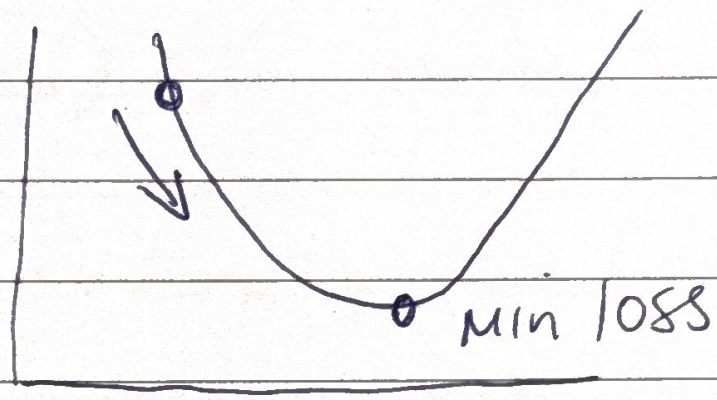
- note we cannot make the form = w
- no closed solution
- cannot be solved analytically to find optimal w

The Alternative solution is a ~~inte~~ iterative solution using gradient descent algorithm

Gradient Descent Algorithm

Optimal parameters are at:

$$\frac{\partial L}{\partial w} = 0$$



↓

Analytical Method to optimize solution

what we do instead is go in the dir of 0

Descend iteratively

iters

(0) init w get loss

(1) update w $\quad w_{new} = w_{old} - \lambda \frac{dL(w)}{dw}$

λ = learning rate

(2) iterate

(3) Stop when $w_{new} \approx w_{old}$

Log Reg Address Overfit w/ regularization

cross entropy loss

$$L(w) = \sum -y_n \log f(x_n) - (1 - y_n) \log(1 - f(x)) + \frac{a\|w\|^2}{2}$$

regularization term L2

Add reg parameter

Summary

1. log reg is classifier
2. loss func = cross entropy loss
3. Optimized by using gradient Descent which has a learning rate hyperparam
4. Overfit Dealt w/ L2 reg