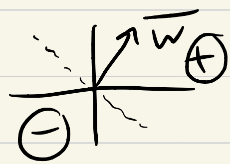
CS371N Lecture 3
Classification 2: Logistic Regression and Optimization

<u>Announcements</u>: A1 due next Thurs

<u>Recap</u> Linear binary classification $y \in \{-1,+1\}$

$$\bar{w}^T f(\bar{x}) \overset{?}{>} 0$$



Features f: Bag of words

$$f(\bar{x}) = \begin{array}{c} \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & \ldots & 1 & \ldots \end{bmatrix} \\ \begin{array}{cccccccc} \text{a} & \text{the} & \text{was} & \text{of} & \text{in} & \ldots & \text{movie} & \ldots \end{array} \end{array}$$

$\bar{x}$ = the movie was great

Perceptron: dataset $\{(\bar{x}^{(i)}, y^{(i)})\}_{i=1}^{D}$

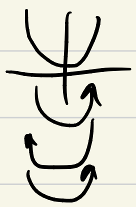
Init $\overline{w} = \overline{0}$

for t in range(0, epochs)

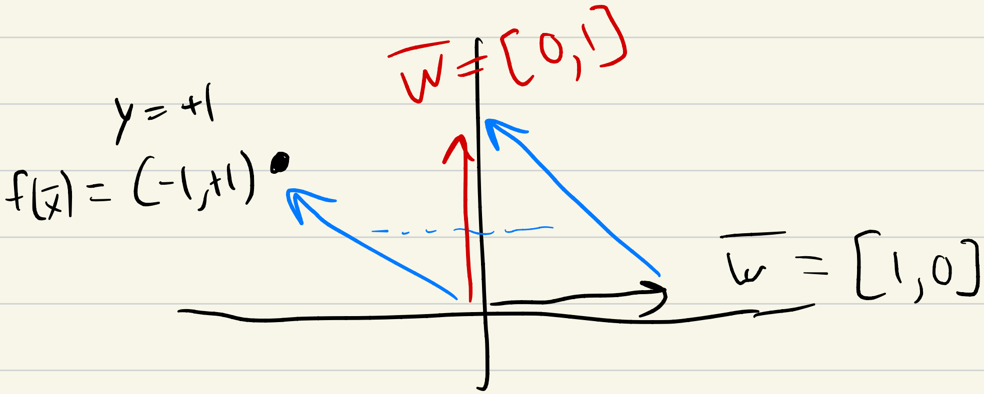
  for i in range(0, D)

$$y_{pred} \leftarrow \begin{cases} 1 & \text{if } \overline{w}^T f(\overline{x}^{(i)}) > 0 \\ -1 & \text{else} \end{cases}$$

$$\overline{w} \leftarrow \begin{cases} \overline{w} & \text{if } y_{pred} = y^{(i)}; \text{ else:} \\ \overline{w} + \alpha f(\overline{x}^{(i)}) & \text{if } y^{(i)} = +1 \\ \overline{w} - \alpha f(\overline{x}^{(i)}) & \text{if } y^{(i)} = -1 \end{cases}$$



$\alpha = \frac{1}{t}$



Ex $\bar{x}^{(1)}$: good $\quad y=+1$

$\bar{x}^{(2)}$: not good $\quad y=-1$

$\bar{x}^{(3)}$: bad $\quad y=-1$

$\bar{w}^T f(\bar{x}) > 0$

if $0 \Rightarrow -1$

1. Write the feature vectors
2. Execute one epoch of perceptron

Start with $\bar{w}=0$, go in order

Assume $\alpha=1$

$\bar{x}$	y	g	n	b	ng	nb
g	+1	[1	0	0]	0	0
ng	−1	[1	1	0] ☆	1	0
b	−1	[0	0	1]	0	0

($f(\bar{x})$ over columns g, n, b)

$\bar{w} = [0\ 0\ 0]$

case 2, add $f(\bar{x})$

Ex 1 $\quad y_{pred} = -1 \Rightarrow \bar{w} = [1\ 0\ 0]$ — ☆

Ex 2 $\quad [1\ 0\ 0]\begin{bmatrix}1\\1\\0\end{bmatrix} = 1,\ y_{pred}=1 \Rightarrow \bar{w} = [0\ -1\ 0]$

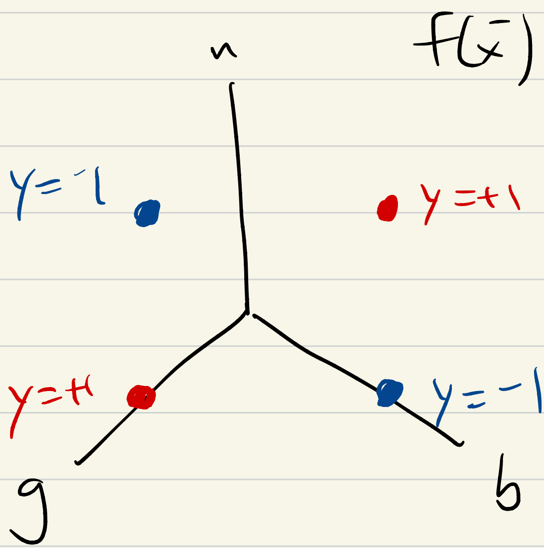
Ex 3 $\quad [0\ -1\ 0]\begin{bmatrix}0\\0\\1\end{bmatrix} = 0 \quad y_{pred} = -1 \Rightarrow$ no change

If we start epoch 2:

update $\Rightarrow$ $[1 \quad -1 \quad 0]$ no more updates

<u>Ex</u> Add the example "not bad"

$$\begin{array}{lllcc} & & & ng & nb \\ nb & +1 & [0 \quad 1 \quad 1] & 0 & 1 \end{array}$$



Counter and Indexer

Indexer : $\left\{ \begin{array}{l} 0 \Leftrightarrow \text{"the"} \\ 1 \Leftrightarrow \text{"a"} \\ \vdots \\ 147 \Leftrightarrow \text{"not good"} \end{array} \right\}$

Counter : sparse vector

$\left\{ \begin{array}{l} 147 \Rightarrow 1 \\ 192 \Rightarrow 2 \end{array} \right\}$ feature vector

Logistic Regression

$\frac{e^z}{1+e^z}$

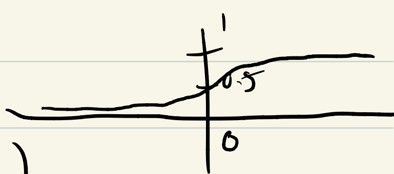
Discriminative probabilistic model

$\frac{1}{1+e^{-z}}$

$P(y \mid \bar{x})$

logistic function

$$P(y=+1 \mid \bar{x}) = \frac{e^{\bar{w}^T f(\bar{x})}}{1+e^{\bar{w}^T f(\bar{x})}}$$



maps $z \in \mathbb{R} \Rightarrow (0,1)$

$$\overline{w}^T f(\overline{x}) \overset{?}{>} 0 \Leftrightarrow y=+1$$

$$\Updownarrow$$

$$P(y=+1 \mid \overline{x}) \overset{?}{>} 0.5$$

<u>Learning</u> Maximize <u>data</u> <u>likelihood</u>

Likelihood $\mathcal{L} = \prod_{i=1}^{D} P(y=y^{(i)} \mid \overline{x}^{(i)})$

Transform: likelihood $\Rightarrow$ log likelihood

$$LL = \sum_{i=1}^{D} \log P(y=y^{(i)} \mid \overline{x}^{(i)})$$

$$\underset{\overline{w}}{\operatorname{argmax}}\ \mathcal{L}(\overline{w}) = \underset{\overline{w}}{\operatorname{argmax}}\ LL(\overline{w})$$

Actually: minimize negative LL

$$\underset{\overline{w}}{\operatorname{argmin}} \sum_{i=1}^{D} \underbrace{-\log\left(P(y=y^{(i)} \mid \overline{x}^{(i)}; \overline{w}\right)}_{\text{log loss}})$$

SGD: $\frac{\partial}{\partial \bar{w}} \text{loss}(\bar{x}^{(i)}, y^{(i)}, \bar{w})$

This gradient $\approx$ perceptron update

$\boxed{\text{Assume } y^{(i)} = +1}$

$$\frac{\partial}{\partial \bar{w}} -\log P(y=+1 \mid \bar{x})$$

$$= \frac{\partial}{\partial \bar{w}} -\log \left[\frac{e^{\bar{w}^T f(\bar{x})}}{1+e^{\bar{w}^T f(\bar{x})}} \right]$$

$$= \frac{\partial}{\partial \bar{w}} \left[-\bar{w}^T f(\bar{x}) + \log\left(1+e^{\bar{w}^T f(\bar{x})}\right) \right]$$

$$= -f(\bar{x}) + \frac{1}{1+e^{\bar{w}^T f(\bar{x})}} \cdot e^{\bar{w}^T f(\bar{x})} \cdot f(\bar{x})$$

$$= f(\bar{x}) \left(-1 + \frac{e^{\bar{w}^T f(\bar{x})}}{1+e^{\bar{w}^T f(\bar{x})}} \right)$$

$$= f(\bar{x}) \left(-1 + P(y=+1 \mid \bar{x}) \right)$$

$$-\frac{\partial}{\partial \bar{w}} = f(\bar{x})\left(1 - P(y=+1 \mid \bar{x})\right)$$

In SGD: we add $-\frac{\partial}{\partial \bar{w}}$ to $\bar{w}$

If $P(y=+1 \mid \bar{x}) \simeq 1$, what happens? $\simeq$ no change

$P(y=+1 \mid \bar{x}) \simeq 0$? $+f(\bar{x})$

If $P(y=+1 \mid \bar{x}) = 0.5$ "half an update"

$\Downarrow$

differs from perceptron

<u>Update</u>

$$y^{(i)} = +1: \bar{w} \leftarrow \bar{w} + \alpha f(\bar{x}^{(i)})\left(1 - P(y=+1 \mid \bar{x}^{(i)})\right)$$

$$y^{(i)} = -1: \bar{w} \leftarrow \bar{w} - \alpha f(\bar{x}^{(i)})\left(\underbrace{1 - P(y=-1 \mid \bar{x}^{(i)})}_{\substack{1-P(y=-1) \\ = P(y=+1)}}\right)$$

Step size:

$\frac{1}{t}$, $\frac{1}{\sqrt{t}}$ ... others possible

SGD = first-order