CS371N Lecture 3

Classification 2: Logistic Regression and Optimization

Announcements – A1 due in 9 days

Recap Linear binary classifier: $\bar{w}^T f(\bar{x}) \overset{?}{>} 0$

Bag-of-words featurization:

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

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

(4 1s)

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

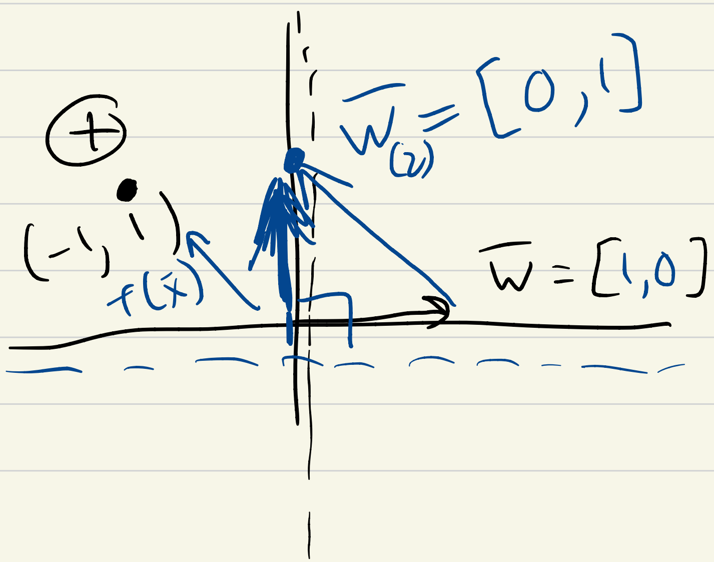
init $\bar{w} = 0$

for t in range(0, epochs)

for i in range(0, D)

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

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



Example

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

if $0 \Rightarrow -1$

$\bar{x}$: good $\quad y : +1$

not good $\quad y : -1$

bad $\quad y : -1$

$\alpha = 1$

① Write the feature vectors (x3)

② Execute one epoch of perceptron
Start with $\bar{w} = 0$, go in order
Give final weight vector

$$\begin{array}{cc|c|ccc}
 & y & \text{feats} & & & \\
 & & \begin{matrix} g & n & b \end{matrix} & & nb & ng \\
g & +1 & [1 \ \ 0 \ \ 0] & & 0 & 0 \\
ng & -1 & [1 \ \ 1 \ \ 0] & & 0 & 1 \\
b & -1 & [0 \ \ 0 \ \ 1] & & 0 & 0
\end{array}$$

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

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

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

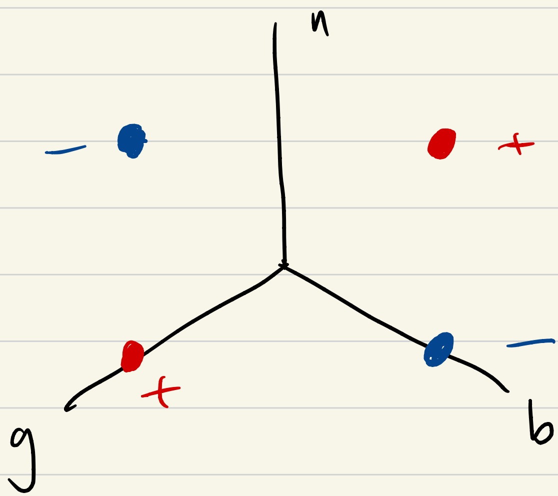
Ex 3 : $[0 \ -1 \ 0] \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = 0$ $\quad y_{pred} = -1$ no change

If we start epoch 2:

update ⇒ [1 -1 0] converged

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

$$\begin{array}{cccc} & & nb & ng \\ nb \; +1 & [0 \; 1 \; 1] & 1 & 0 \end{array}$$



Logistic Regression

Discriminative probabilistic model

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

label features/instance

(generative: $P(\bar{x}, y)$)

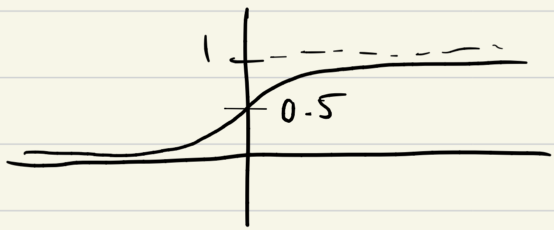
Naive Bayes

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

$\left(\frac{1}{1+e^{-z}}\right)$

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

logistic fcn



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

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

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

equivalent to our earlier decision rule

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

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

Learning: Maximize the data likelihood

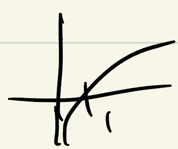
Likelihood $L = \prod_{i=1}^{D} P(y=y^{(i)} \mid \bar{x}^{(i)})$

We want $\bar{w}^* = \underset{\bar{w}}{\operatorname{argmax}}\ L(\bar{w})$

① Likelihood $\Rightarrow$ log likelihood (LL)

$$\underset{\bar{w}}{\operatorname{argmax}} \sum_{i=1}^{D} \log P(y=y^{(i)} \mid \bar{x}^{(i)})$$

log is monotonic



(NLL)

② Minimize the negative LL

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

$$\text{loss}(\bar{x}^{(i)}, y^{(i)}, \bar{w})$$

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

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

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

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

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

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

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

$$= f(\bar{x}) \underbrace{\left(-1 + P(y = +1 \mid \bar{x}) \right)}_{\text{scaling factor}}$$

What if $P(y=+1 \mid \bar{x}) \approx 1$ ("y_{pred}")

Little update (but still nonzero)

What if $P(y=+1 \mid \bar{x}) \approx 0$? Essentially perc update

$P(y=+1 \mid \bar{x}) = 0.50001$? $(-f(\bar{x}))$ update

Perc: no update

HPerc: update!

Update:

$$y^{(i)} = +1: \quad \bar{w} \leftarrow \bar{w} + \overset{\substack{\text{step size}\\ \downarrow}}{\alpha} f(\bar{x}^{(i)}) \left(1 - P(y=+1 \mid \bar{x}^{(i)})\right)$$

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

Still want a step size.

Consider constant, $\frac{1}{t}$, $\frac{1}{\sqrt{t}}$, ... $\quad$ t epoch number

SGD is first-order optimization

Newton's method is second-order

second deriv: Hessian $n \times n$ matrix

$$\frac{\partial^2}{\partial w_i w_j}$$