CS 378 Lecture 3
Classification 2: Logistic Regression, Optimization

Announcements
- A1
- Videos / readings

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

Bag-of-words featurization:

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

$f(\bar{x}) = [\,0 \;\; 1 \;\; 1 \;\; 0 \;\; 0 \;\; \ldots \;\; 1 \;\; \ldots \;\; 1\,]$

a the was of in movie great

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

init $\bar{w} = \bar{0}$

for t in range(0, epochs)

  for i in range(0, D)

    $y_{pred} \leftarrow 1$ if $\bar{w}^T f(\bar{x}^{(i)}) > 0$

    else -1

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

Example

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

not good $y: -1$

bad $y: -1$

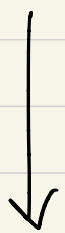
$\bar{w}^T f(\bar{x}) > 0$

if $= 0 \Rightarrow -1$

(1) Write the feature vector for each

(2) Execute 1 epoch of perceptron

	y	feats g	b	n
g	+1	[1	0	0]
ng	-1	[1	0	1]
b	-1	[0	1	0]



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

Ex 1: $\overline{w}^T \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \Rightarrow 0$ (-1)

$\overline{w} \leftarrow \overline{w} + 1\,[1\ 0\ 0] \Rightarrow [1\ 0\ 0]$

Ex 2: $[1\ 0\ 0] \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \Rightarrow 1$ (+1)

$\overline{w} \leftarrow \overline{w} - 1\,[1\ 0\ 1] \Rightarrow [0\ 0\ -1]$

Ex 3: Correct, no Change

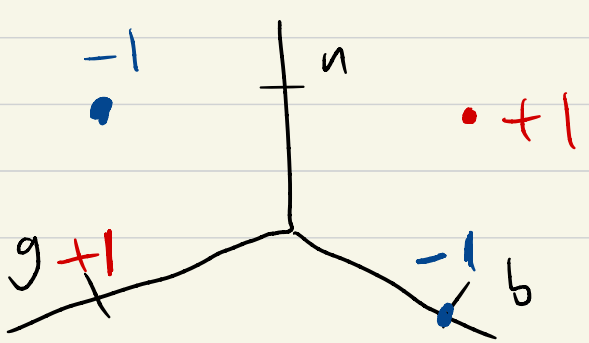
If we do ex 1 again:

$[1 \ 0 \ -1]$ no more updates

Example 2

		g	b	n	ng
good	+1	1	0	0	0
not good	−1	1	0	1	1
bad	−1	0	1	0	0
not bad	+1	0	1	1	0

$[1 \ 0 \ -1]$ $-1 \Rightarrow (-1)$ ypred



Logistic Regression

Discriminative probabilistic model

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

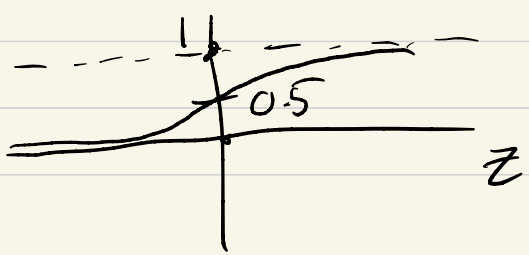
label | features / instance

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

Naive Bayes

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

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



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

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

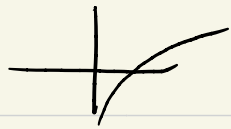
$$P(y=-1 \mid \bar{x}) \triangleq 1 - P(y=+1 \mid \bar{x})$$

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

Learning: Maximize the data likelihood

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

We want $\bar{w} = \underset{\bar{w}}{\text{argmax}}$ data likel.



$$\Rightarrow \underset{\bar{w}}{\text{argmax}} \log \text{data likelihood}$$

$$\underset{\bar{w}}{\text{argmax}} \underbrace{\sum_{i=1}^{D} \log P(y = y^{(i)} | \bar{x}^{(i)})}_{\log \text{likelihood}}$$

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

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

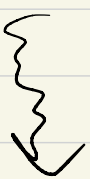
$$\text{SGD: } \frac{\partial}{\partial \bar{w}} \text{loss}(\bar{x}^{(i)}, y^{(i)}, \bar{w})$$

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

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

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

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



Update:

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

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

Update:

α

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

$y^{(i)} = -1: \quad \bar{w} \leftarrow \bar{w} - f(\bar{x}^{(i)})\left(1 - P(y=-1|\bar{x}\right.$

α

$y^{(i)} = +1:$

What happens if $P(y=+1|\bar{x})$ on this ex. is 1?

$\bar{w}$ Doesn't change

$P(y=+1|\bar{x})$ is 0?

$\bar{w} \leftarrow \bar{w} + \alpha f(\bar{x}^{(i)})$

LR never "converges" exactly

$P(y=+1|\bar{x})$ is 0.7? $+ 0.3\alpha f(\bar{x}^{(i)})$

Optimization

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

$\mathcal{L}(\bar{w})$ loss on the whole dataset

$\left(\sum \mathcal{L}(\bar{x}^{(i)}, y^{(i)}, \bar{w})\right)$

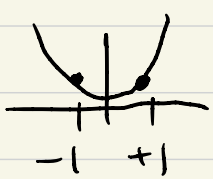
SGD: $\bar{w} \leftarrow \bar{w} - \alpha \frac{\partial}{\partial \bar{w}} \mathcal{L}(\bar{x}^{(i)}, y^{(i)}, \bar{w})$

step size

$\mathcal{L}(w) = w^2$

$w = -1$

$\frac{\partial}{\partial w} \mathcal{L} = 2w$



SGD:

$w = -1$

$w \leftarrow w - \alpha \cdot 2w$

loops infinitely

$\alpha = 1$: $w: -1 \rightarrow +1$

$\alpha = 0.5$: $w: -1 \rightarrow 0$ ✓

How to choose step size?

— Start with 1, decrease it

$1/t$ t : epoch

(may decrease too fast)

$\frac{1}{\sqrt{t}}$, e^{-t} is even faster

2nd deriv: Newton

quasi-Newton

Adagrad, Adam, AdamW

approximate 2nd-order