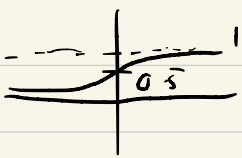
CS371N Lecture 4: Multiclass

Announcements

- A1 due in one week
- Social impact response on Tues, open for 1 week

Recap Logistic regression

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



Training: minimize the negative log likelihood

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

$$\bar{w} \leftarrow \bar{w} + \alpha f(\bar{x}^{(i)}) \quad \text{if } y^{(i)}=+1$$

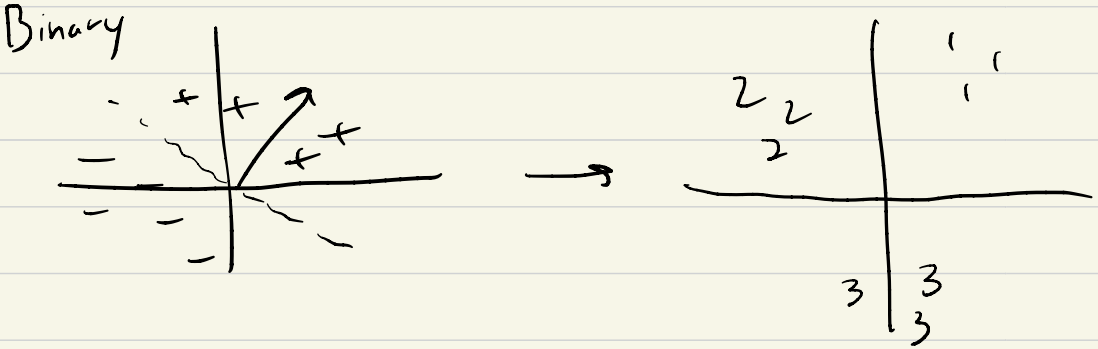
$$-\alpha f(\bar{x}^{(i)}) \quad \text{if } y^{(i)}=-1$$

Today – Sentiment analysis: see slides

– Multiclass: formulation, perc, LR

Multiclass basics

Binary



Output space $\mathcal{Y} = \{1, 2, 3\}$

1 vs (2 or 3)

↓

2 or 3?

1 vs (2 or 3)

2 vs (1 or 3)

3 vs (1 or 2)

Decision tree

$|\mathcal{Y}|$ classifiers

One-vs-all

take highest $w^T f$ as label

Multiclass solution: like one-vs-all, but more unified

Two ways of setting things up:

① Different weights per class

② Different features per class

① Different weights

$\bar{w}_1$ $\bar{w}_2$ $\bar{w}_3$ weight vector per class in $\mathcal{Y}$

$$y_{pred} = \underset{y \in \mathcal{Y}}{\text{argmax}} \; \bar{w}_y^T f(\bar{x})$$

Ex Headline Classification $y=1$ (health)

$\bar{x}$ = "too many drug trials, too few patients"

$y = \{\underset{1}{\text{health}}, \underset{2}{\text{sports}}, \underset{3}{\text{science}}\}$

$f(\bar{x}) = [\underset{\text{drug}}{1} \quad \underset{\text{patients}}{1} \quad \underset{\text{baseball}}{0}]$

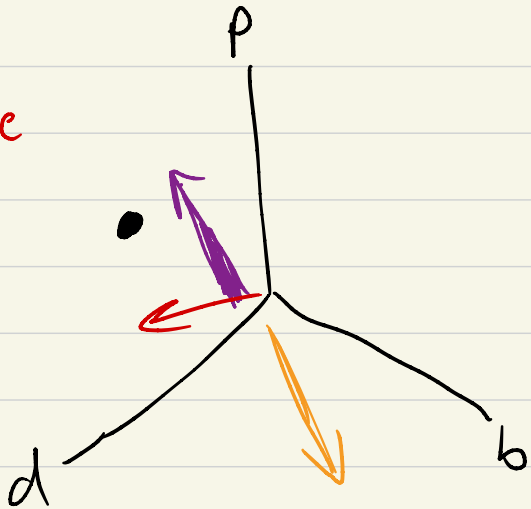
$\bar{w}_1 = [2, 5.6, -3] \rightarrow 7.6$

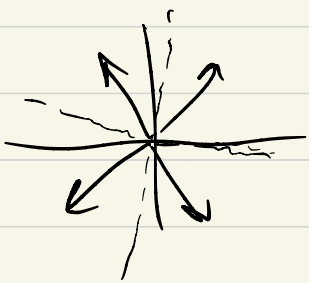
$\bar{w}_2 = [1.2, -3.1, 5.7] \rightarrow -1.9$

$\bar{w}_3 = [1, 1.2, -0.5] \rightarrow 2.2$

decision = health (class 1)

health
sports
science





Multiclass Perceptron

for t in epochs

  for i in data $(\bar{x}^{(i)}, y^{(i)})$

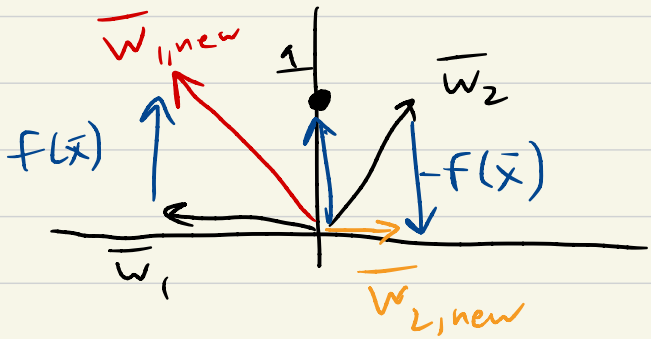
$$y_{pred} \leftarrow \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \; \bar{w}_y^T f(\bar{x}^{(i)})$$

  if $y_{pred} \neq y^{(i)}$:

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

$$\bar{w}_{y_{pred}} \leftarrow \bar{w}_{y_{pred}} - \alpha f(\bar{x}^{(i)})$$

$f(\bar{x}) = (0,1)$



$y_{pred} = 2$

$2 \neq 1$

Add $f(\bar{x})$ to $\bar{w}_1$

Subtract $f(\bar{x})$ from $\bar{w}_2$

Multiclass LR

$$P(y=\hat{y} \mid \bar{x}) = \frac{e^{\bar{w}_y^T f(\bar{x})}}{\sum_{y' \in Y} e^{\bar{w}_{y'}^T f(\bar{x})}}$$

distribution over $\hat{y} \in Y$

$$P(y = \text{class } 1 \mid \bar{x}) = \frac{e^{\bar{w}_1^T f(\bar{x})}}{e^{\bar{w}_1^T f(\bar{x})} + e^{\bar{w}_2^T f(\bar{x})} + e^{\bar{w}_3^T f(\bar{x})}}$$

Softmax operation

Update SGD of negative log likelihood

For $y^{(i)}$: the correct class

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

For all other y'

$$\bar{w}_{y'} \leftarrow \bar{w}_{y'} - \alpha f(\bar{x}^{(i)}) \left(P(y = y' | \bar{x})\right)$$

Classes w/probs $[0.1, 0.8, 0.1]$ No ypred

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

$1 - P(y = 1 | \bar{x}) = 1 - 0.1 = 0.9$

$\bar{w}_1$ add $0.9\, f(\bar{x})$

$\bar{w}_2$ sub. $0.8\, f(\bar{x})$

$\bar{w}_3$ sub. $0.1\, f(\bar{x})$

Suppose it's $[0.99 \quad 0.005 \quad 0.005]$

Ex 2 Suppose we have a single ex

$[1\ 1\ 0]$ label = 1 out of $y = [1\ 2\ 3]$

Initialize all $\vec{w}$ to $\vec{0}$

(1) What are the class probs we get?

(2) Do one step of the update. $\alpha = 1$

(3) What if we keep training on this ex?

$[1/3\ \ 1/3\ \ 1/3]$

$\vec{w}_1 = [2/3\ \ 2/3\ \ 0]$

$\vec{w}_2 = [-1/3\ \ -1/3\ \ 0]$

$\vec{w}_3 = [-1/3\ \ -1/3\ \ 0]$