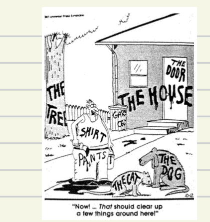
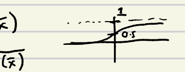
CS 378 Lecture
Classification 4 :
Multiclass



Announcements
- Pronouns on Zoom
- A1 due in a week
- Videos

Recap: Logistic regression

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



Train to maximize likelihood of data:

$$\max_{\bar{w}} \prod_{i=1}^{D} P(y=y^{(i)}|\bar{x}^{(i)}) \Rightarrow \max_{\bar{w}} \sum_{i=1}^{D} \log P(y=y^{(i)}|\bar{x}^{(i)})$$

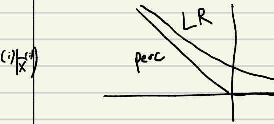
Update from SGD:

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

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

If $P(y^{(i)}=+1|\bar{x}^{(i)}) \approx 1$, prediction is correct
$\Rightarrow \bar{w} \leftarrow \bar{w} + \alpha \cdot 0$
If $P(\text{—}) \approx 0$, wrong
$\Rightarrow \bar{w} \leftarrow \bar{w} + \alpha f(\bar{x}^{(i)})$ perceptron

Define $z = \bar{w}^T f(\bar{x})$ $loss_{LR} = \log(1+e^{z}) - z$



Today Finish sentiment
Multiclass classification
Perc, LR

Sentiment Features

Unigrams: $[\underset{\text{the}}{1}\ \underset{\text{a}}{0}\ \underset{\text{of}}{0}\ 1 \ldots 0]$ 10K long

Bigrams: $[\underset{\text{the movie}}{1}\ \underset{\text{a movie}}{0}\ \ldots]$ 1M long

How to combine?

"Mixed" feature vector

$[\underset{\text{Unigram=the}}{1}\ \underset{\text{Bigram=the movie}}{1}\ \underset{\text{Unigram=movie}}{1} \ldots]$

10K+1M feats in the same vector
Maintain one feature index

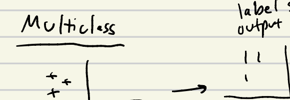
the movie was

Unigram = the
Unigram = movie
Unigram = was
Bigram = the m
Bigram = m was

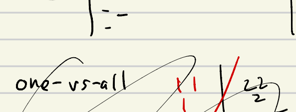
→ index, put counts in vector

Multiclass

label space / output space $\mathcal{Y} = \{1,2,3\}$



one-vs-all



Hacky!

Key idea Two ways to do multiclass:
one weight vector per class (different weights DW)
different feature vector per class, single weight vector (DF)

DW $\bar{w}_1, \bar{w}_2, \bar{w}_3$ for each class

predicted label $= \underset{y \in \mathcal{Y}}{\text{argmax}}\ \underbrace{\bar{w}_y^T f(\bar{x})}$

$f(\bar{x}) = [1,1,0]$
$\bar{w}_1 = [2, 5.6, -3]$ $\bar{w}_2 = [1.2, -3.1, 5.7]$
score for class 1 = 7.6 class 2 = -1.9
$\Rightarrow$ class 1 is highest $\Rightarrow$ return 1

DF One $\bar{w}$, features differ per class

predicted label $= \underset{y \in \mathcal{Y}}{\text{argmax}}\ \bar{w}^T \underline{\underline{f(\bar{x}, y)}}$

Ex $\bar{x}$ = too many drug trials, too few patients
$\mathcal{Y} = \{\text{health, sports, science}\}$
y = health

$f(\bar{x}) = [\text{Unigram=drug}, \text{Unigram=patients}, \text{Unigram=baseball}]$
$f(\bar{x}) = [1,1,0]$ copy-pasted $f(\bar{x})$ 3 times
$f(\bar{x},y) = [\underbrace{\_\ \_\ \_}_{\text{health}}, \underbrace{\_\ \_\ \_}_{\text{sports}}, \underbrace{\_\ \_\ \_}_{\text{science}}]$
$f(\bar{x}, y=\text{health}) = [1,1,0,0,0,0,0,0,0]$

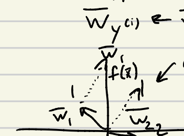
$\bar{w} = [\underbrace{2, 5.6, -3}_{\text{health}}, \underbrace{1.2, -3.1, 5.7}_{\text{sports}}, \ldots]$
$f(\bar{x}, y=\text{sports}) = [0,0,0,1,1,0,0,0,0]$
$\bar{w}^T f(\bar{x}, y=\text{sports}) = -1.9$
$\bar{w}^T f(\bar{x}, y=\text{health}) = 7.6$

Why? DW: neural nets
$\underline{\underline{\bar{w}_y}}^T \underbrace{f(\bar{x})}$ our NN

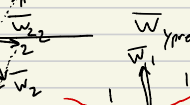
DF: Eisenstein book
Structured prediction: POS tagging

Multiclass Perceptron $\boxed{\text{DW}}$

for t in epochs
  for i in data:
    $y_{pred} = \underset{y \in \mathcal{Y}}{\text{argmax}}\ \bar{w}_y^T f(\bar{x})$
    if $y_{pred} \neq y^{(i)}$:
      $\bar{w}_{y_{pred}} \leftarrow \bar{w}_{y_{pred}} - \alpha f(\bar{x})$
      $\bar{w}_{y^{(i)}} \leftarrow \bar{w}_{y^{(i)}} + \alpha f(\bar{x})$



current ex
$y_{pred}=2$
$\bar{w}_{y_{pred}} - \alpha f(\bar{x})$



Example 3 classes
[1,1,0] class 1
[1,0,1] class 2
default pred = 3

$\bar{w}_1 = \bar{0}$ $\bar{w}_2 = \bar{0}$, $\bar{w}_3 = \bar{0}$

after ex 1: $y_{pred}=3$
$\bar{w}_1 = 1\ 1\ 0$
$\bar{w}_2 = 0\ 0\ 0$
$\bar{w}_3 = -1\ -1\ 0$

ex 2: $y_{pred}=1$

Answer
$$\boxed{\begin{array}{l} \bar{w}_1 = 0\ 1\ -1 \\ \bar{w}_2 = 1\ 0\ 1 \\ \bar{w}_3 = -1\ -1\ 0 \end{array}}$$

Multiclass Perc w/ different features

for ..
  for...
    $y_{pred} = \underset{y \in \mathcal{Y}}{\text{argmax}}\ \bar{w}^T f(\bar{x}, y)$ hypothesized y, NOT $y^{(i)}$
    if $y_{pred} \neq y^{(i)}$:
      $\bar{w} \leftarrow \bar{w} + \alpha f(\bar{x}, y^{(i)}) - \alpha f(\bar{x}, y_{pred})$

Multiclass Logistic Regression

$$P(y=\hat{y}|\bar{x}) = \frac{e^{\bar{w}_{\hat{y}}^T f(\bar{x})}}{\sum_{y' \in \mathcal{Y}} e^{\bar{w}_{y'}^T f(\bar{x})}}$$ $\boxed{\text{DW}}$