CS 378 Lecture 4

Classification 3: Multiclass

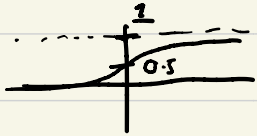
"Now! ... *That* should clear up a few things around here!"

Announcements

- A1 due in one week
- Social impact response: in-class Thurs / open for a week
- Seating chart

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:

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

Update from SGD:

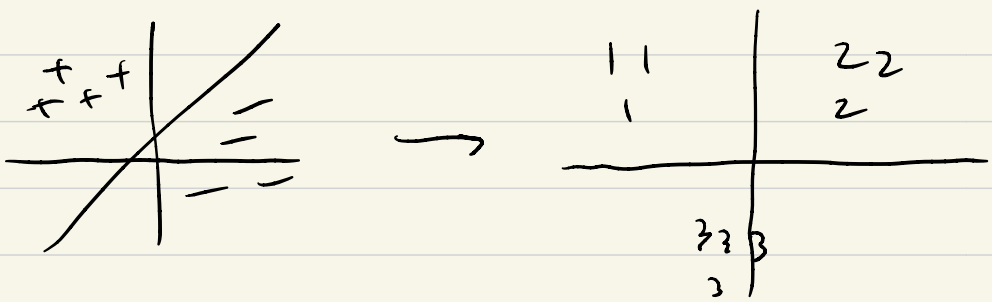
$$\bar{w} \leftarrow \bar{w} + \alpha f(\bar{x}^{(i)})\left(1 - P(y = +1 | \bar{x}^{(i)})\right) \quad \text{if } y^{(i)} = +1$$
$$\bar{w} \leftarrow \bar{w} - \alpha f(\bar{x}^{(i)})\left(1 - P(y = -1 | \bar{x}^{(i)})\right) \quad \text{if } y^{(i)} = -1$$

Connections to perceptron (see video / lecture note)

Optimization: for now, SGD w/tuned step size. Later: better methods
Convergence is always imperfect in this course!

Today Multiclass classification
MC Perc
MC LR
MC examples

Multiclass Classification



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

one-vs-all: each class has a binary classifier "in or out"

1: + 2: + 3: − is it 1 or 2?

Compare probs

We won't do one-vs-all, too hacky.

Key idea Two ways to think about MC

one weight vector per class ("different weights")

diff. feature vector per class for the same example (diff. features)

Different weights

$\bar{w}_1, \bar{w}_2, \bar{w}_3$ weight vector per class

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

loop over $y \in Y$, compute, take the highest index

returns either 1, 2, or 3

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

$\bar{w}_1 = [2, 5.6, -3]$ $\bar{w}_2 = [1.2, -3.1, 5.7]$

score for class 1 = 7.6 class 2 = -1.9

$\Rightarrow$ return class 1

Different features One $\bar{w}$, each class gives different feats to an ex

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

Ex Topic classification

$\mathcal{Y} = \{\text{health}_1, \text{sports}_2, \text{science}_3\}$

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

$f(\bar{x}) = [$ drug patients baseball $]$

$= [\ 1 \quad 1 \quad 0\ ]$

Diff feats: copy-pastes $f(\bar{x})$ for each class

$f(\bar{x}, y) = [\_\_\_, \_\_\_, \_\_\_]$

health sports science

$f(\bar{x}, y = \text{health}) = [1, 1, 0, 0\,0\,0, 0\,0\,0]$

$f(\bar{x}, y = \text{sports}) = [0, 0, 0, 1, 1, 0, 0, 0, 0]$

$f(\bar{x}, y = \text{science}) = \ldots$

$\bar{w} = [2, 5.6, -3, 1.2, -3.1, 5.7, \ldots]$

$\bar{w}^T f(\bar{x}, y = \text{health}) = 7.6$

$\bar{w}^T f(\bar{x}, y = \text{sci}$

$\bar{w}^T f(\bar{x}, y = \text{sports}) = -1.9$

why? Diff weights: neural net

Diff feats: Eisenstein book, structured prediction (sequence of POS tags)

Multiclass Perceptron (diff. weights)

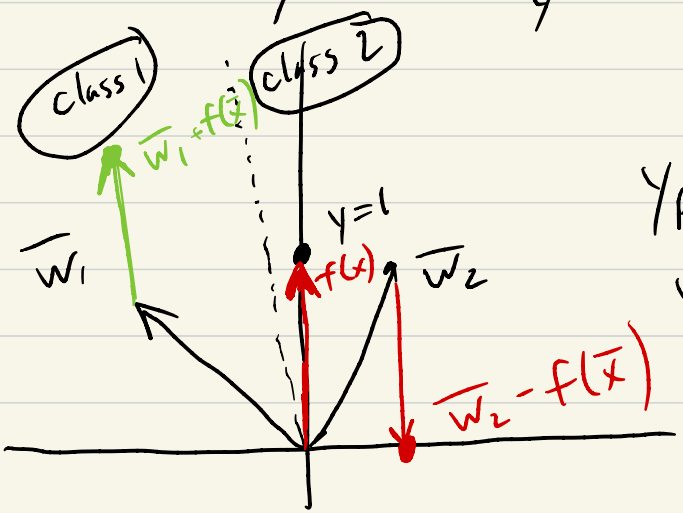
for t in epochs:
  for i in data:
    $(\bar{x}^{(i)}, y^{(i)})$

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

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

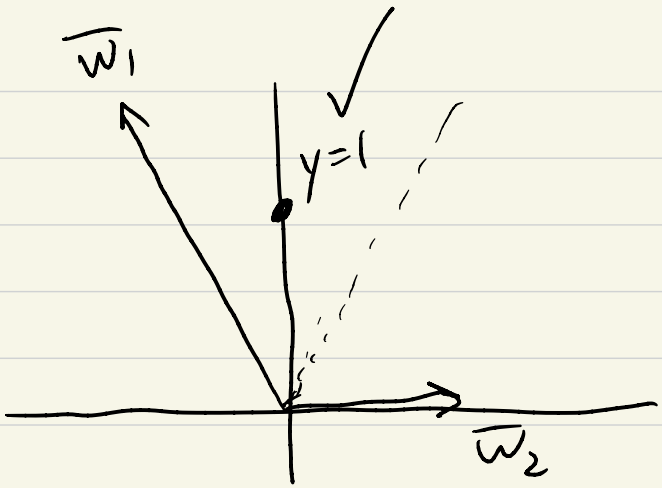
$$\bar{w}_{y_{pred}} \leftarrow \bar{w}_{y_{pred}} - \alpha f(\bar{x})$$

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



$y_{pred} = 2$

$y^{(i)} = 1$



Multiclass Logistic Regression

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

prob. dist. over the classes

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

Is it a probability dist.?

Update from SGD:

Suppose we have $(\bar{x}^{(i)}, y^{(i)})$

Always make an update

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

For all other y':

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

Ex Probs $[0.9, 0.1, 0]$

$y^{(i)} = 2$

$\bar{w}_2$ "add 0.9"

$\bar{w}_1$ "subtract 0.9"

$\bar{w}_3$ "subtract 0"

Instapoll:

Single ex $(\bar{x}, y) = [1\ 1\ 0]$, class 1

$\bar{w}_1 = \bar{w}_2 = \bar{w}_3 = [0\ 0\ 0]$ to start

① Probs = $[1/3, 1/3, 1/3]$

② After one update:

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

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

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

③ What happens?