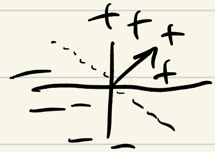
CS 378 Lecture 3
Classification 2: Logistic Regression, Optimization

Announcements

- A1
- Seating chart
- Video on course website



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

Bag-of-words featurization

$\bar{x}$ = movie was great

$$\Rightarrow f(\bar{x}) = \begin{array}{c} \begin{bmatrix} 0 & 1 & 0 & 0 & \dots & 1 & \dots & 1 \end{bmatrix} \\ \text{the} \quad \text{was} \quad \text{a} \quad \text{of} \quad \quad \text{movie} \quad \quad \text{great} \end{array}$$

Perceptron: initialize $\bar{w} = \bar{0}$

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

$$\begin{array}{lcc}
 & & \text{feats} \\
 & y & \begin{array}{ccc} g & b & n \end{array} \\
\text{good} & +1 & [1 \ 0 \ 0] = f(\bar{x}^{(1)}) \\
\text{not good} & -1 & [1 \ 0 \ 1] \\
\text{bad} & -1 & [0 \ 1 \ 0]
\end{array}$$

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

Ex 1: $\bar{w}^T f(\bar{x}^{(1)}) = 0 \overset{?}{>} 0 \Rightarrow \overset{y_{pred}}{-1}$

$\bar{w} \leftarrow \bar{w} + \alpha \cdot [1 \ 0 \ 0]$ $\quad \alpha = 1$

$\bar{w} = [1 \ 0 \ 0]$

Ex 2: $\bar{w}^T f(\bar{x}^{(2)}) = 1 \overset{?}{>} 0 \Rightarrow \overset{y_{pred}}{+1}$

$\bar{w} \leftarrow \bar{w} - \alpha \cdot [1 \ 0 \ 1]$

$\bar{w} = [0 \ 0 \ -1]$

Ex 3: correct

After ex 1 again: $[1 \ 0 \ -1]$

Today

- Logistic regression
- Optimization
- Sentiment systems (if time)

Logistic Regression

Discriminative probabilistic model

Models a distribution $P(y \mid \bar{x})$

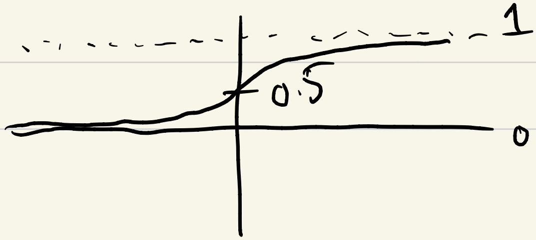
label instance

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 function



Decision boundary: predict +1 if

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

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

equivalent to

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

in order for it to be a dist. over y

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

Learning

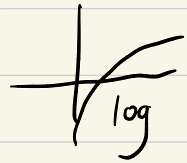
For a dataset $\{(\bar{x}^{(i)}, y^{(i)})\}_{i=1}^{D}$,

maximize the data likelihood

logistic reg. w/weights $\overline{w}$

$$\max_{\overline{w}} \prod_{i=1}^{D} P(y^{(i)} | \overline{x}^{(i)})$$

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



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

$$= \min_{\overline{w}} \sum_{i=1}^{D} \underbrace{- \log P(y^{(i)} | \overline{x}^{(i)})}_{\text{loss}(\overline{x}^{(i)}, y^{(i)}, \overline{w})}$$

For stochastic gradient descent:

need $\frac{\partial}{\partial \overline{w}} \text{loss}(\overline{x}, y, \overline{w})$

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

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

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

CALCULUS

Logistic regression update

$$\bar{w} \leftarrow \bar{w} - \alpha \frac{\partial}{\partial \bar{w}} \text{loss}$$

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

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

for $y^{(i)} = -1$

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

① Think about when $y^{(i)} = +1$

What happens when:

$P(y^{(i)} = +1 \mid \bar{x})$ is close to 1? not much update!

close to 0? perc update

close to 0.5?

Optimization

$$\sum_{i=1}^{D} \text{loss}(\bar{x}^{(i)}, y^{(i)}, \bar{w}) \triangleq \mathcal{L}\left((\bar{x}^{(i)}, y^{(i)})_{i=1}^{D}, \bar{w}\right)$$

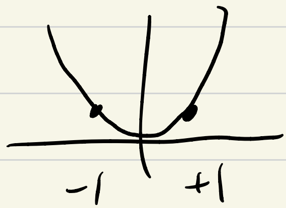
function of $\bar{w}$

$\mathcal{L}(\bar{w})$

SGD: $\bar{w} \leftarrow \bar{w} - \alpha \frac{\partial}{\partial \bar{w}} \mathcal{L}(\bar{w})$

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

$x = -1$



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

Suppose $\alpha = 1$

SGD: start at $x = -1$

$$\Rightarrow x \leftarrow x - 1 \cdot \frac{\partial}{\partial x}\mathcal{L}(x)$$

$$\leftarrow -1 - 1 \cdot (-2)$$

$$x = 1$$

oscillates with $\alpha = 1$

If $\alpha = 1/2$, converges instantly

How to choose step size?

- Different constants
- Large $\to$ small, e.g. $1/t$ where t = epoch number

$\frac{1}{\sqrt{t}}$, ... other options

e^{-t}: decreases too fast

How to do something smarter?

Newton: $\overline{w} \leftarrow \overline{w} - \underbrace{\left(\frac{\partial^2}{\partial \overline{w}^2} \mathcal{L}\right)^{-1}}_{\text{inverse Hessian}} \frac{\partial}{\partial \overline{w}} \mathcal{L}$

$n \times n$ matrix

n = # of features parameters

Approximate 2nd-order methods: Adagrad, Adadelta, Adam, AdamW ...

Regularization: not used