

Logistic Regression

CMSC 422

SOHEIL FEIZI

sfeizi@cs.umd.edu

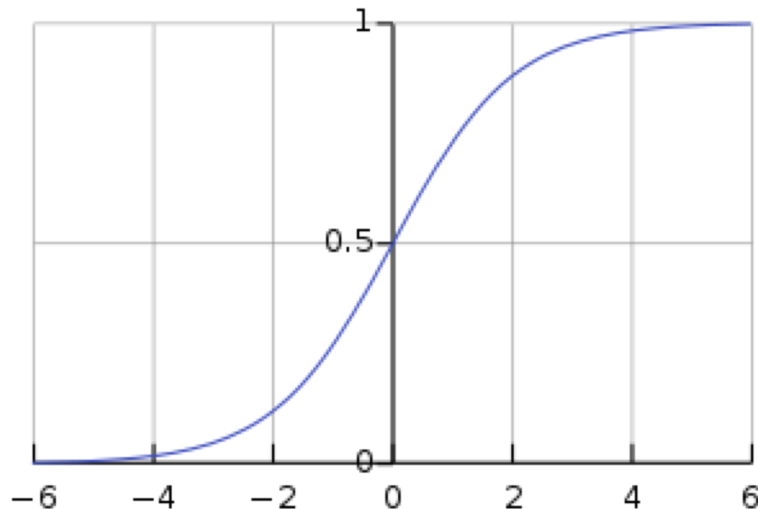
Slides partially adapted from MARINE CARPUAT

Logistic Regression

- Binary classification

$$P(Y^{(i)} = 1|X^{(i)}, \theta) = g(< \theta, X^{(i)} >)$$

$$P(Y^{(i)} = 0|X^{(i)}, \theta) = 1 - g(< \theta, X^{(i)} >)$$



Sigmoid function

$$g(z) = \frac{1}{1 + \exp(-z)}$$

Logistic Regression

- Maximum Likelihood

$$\max_{\theta} \prod_{i=1}^N P(Y^{(i)} | X^{(i)}, \theta)$$



$$\max_{\theta} \prod_{i=1}^N g(< \theta, X^{(i)} >)^{Y^{(i)}} (1 - g(< \theta, X^{(i)} >))^{1-Y^{(i)}}$$



$$\max_{\theta} \sum_{i=1}^N Y^{(i)} \log g(< \theta, X^{(i)} >) + (1 - Y^{(i)}) \log(1 - g(< \theta, X^{(i)} >))$$

Cross-entropy loss function

How to solve it?

- Gradient Descent

- A good property of sigmoid:

$$\nabla_z g(z) = g(z)(1 - g(z))$$

- SGD: $\theta_{k+1} = \theta_k + \eta(Y^i - g(\langle \theta, X^i \rangle))X^{(i)}$
- Why? Intuition behind the updates

Multiclass classification

- Real world problems often have multiple classes (text, speech, image, biological sequences...)
- How can we perform multiclass classification?
 - Straightforward with decision trees or KNN
 - Can we use the perceptron algorithm?

Reductions for Multiclass Classification

TASK: MULTICLASS CLASSIFICATION

Given:

1. An input space $\mathcal{X}$ and number of classes K
2. An unknown distribution $\mathcal{D}$ over $\mathcal{X} \times [K]$

Compute: A function f minimizing: $\mathbb{E}_{(x,y) \sim \mathcal{D}} [f(x) \neq y]$

TASK: BINARY CLASSIFICATION

Given:

1. An input space $\mathcal{X}$
2. An unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, +1\}$

Compute: A function f minimizing: $\mathbb{E}_{(x,y) \sim \mathcal{D}} [f(x) \neq y]$

How many classes can we handle in practice?

- In most tasks, number of classes $K < 100$
- For much larger K
 - we need to frame the problem differently
 - e.g, machine translation or automatic speech recognition

Reduction 1: OVA

- “One versus all” (aka “one versus rest”)
 - Train K-many binary classifiers
 - classifier k predicts whether an example belong to class k or not
- At test time,
 - If only one classifier predicts positive, predict that class
 - Break ties randomly

Algorithm 12 ONEVERSUSALLTRAIN($\mathbf{D}^{multiclass}$, BINARYTRAIN)

```
1: for  $i = 1$  to  $K$  do
2:    $\mathbf{D}^{bin} \leftarrow$  relabel  $\mathbf{D}^{multiclass}$  so class  $i$  is positive and  $\neg i$  is negative
3:    $f_i \leftarrow$  BINARYTRAIN( $\mathbf{D}^{bin}$ )
4: end for
5: return  $f_1, \dots, f_K$ 
```

Algorithm 13 ONEVERSUSALLTEST($f_1, \dots, f_K, \hat{\mathbf{x}}$)

```
1:  $score \leftarrow \langle 0, 0, \dots, 0 \rangle$  // initialize  $K$ -many scores to zero
2: for  $i = 1$  to  $K$  do
3:    $y \leftarrow f_i(\hat{\mathbf{x}})$ 
4:    $score_i \leftarrow score_i + y$ 
5: end for
6: return  $\operatorname{argmax}_k score_k$ 
```

Time complexity

- Suppose you have N training examples, in K classes. How long does it take to train an OVA classifier
 - if the base binary classifier takes $O(N)$ time to learn?
 - if the base binary classifier takes $O(N^2)$ time to learn?

Reduction 2: AVA

- All versus all (aka all pairs)
- How many binary classifiers does this require?

Algorithm 14 ALLVERSUSALLTRAIN($\mathbf{D}^{multiclass}$, BINARYTRAIN)

```
1:  $f_{ij} \leftarrow \emptyset, \forall 1 \leq i < j \leq K$ 
2: for  $i = 1$  to  $K-1$  do
3:    $\mathbf{D}^{pos} \leftarrow$  all  $\mathbf{x} \in \mathbf{D}^{multiclass}$  labeled  $i$ 
4:   for  $j = i+1$  to  $K$  do
5:      $\mathbf{D}^{neg} \leftarrow$  all  $\mathbf{x} \in \mathbf{D}^{multiclass}$  labeled  $j$ 
6:      $\mathbf{D}^{bin} \leftarrow \{(\mathbf{x}, +1) : \mathbf{x} \in \mathbf{D}^{pos}\} \cup \{(\mathbf{x}, -1) : \mathbf{x} \in \mathbf{D}^{neg}\}$ 
7:      $f_{ij} \leftarrow$  BINARYTRAIN( $\mathbf{D}^{bin}$ )
8:   end for
9: end for
10: return all  $f_{ij}$ s
```

Algorithm 15 ALLVERSUSALLTEST(all f_{ij} , $\hat{\mathbf{x}}$)

```
1:  $score \leftarrow \langle 0, 0, \dots, 0 \rangle$  // initialize  $K$ -many scores to zero
2: for  $i = 1$  to  $K-1$  do
3:   for  $j = i+1$  to  $K$  do
4:      $y \leftarrow f_{ij}(\hat{\mathbf{x}})$ 
5:      $score_i \leftarrow score_i + y$ 
6:      $score_j \leftarrow score_j - y$ 
7:   end for
8: end for
9: return  $\operatorname{argmax}_k score_k$ 
```

Time complexity

- Suppose you have N training examples, in K classes. How long does it take to train an AVA classifier
 - if the base binary classifier takes $O(N)$ time to learn?
 - if the base binary classifier takes $O(N^2)$ time to learn?