

# CS 446/ECE 449: Machine Learning

Shenlong Wang

University of Illinois at Urbana-Champaign, 2024

## Logistics:

- Office hour today needs to be shifted a bit: 1:45pm - 2:30pm

## L4: Logistic Regression

## **Goals of this lecture**

- Understand logistic regression
- Understand how it fixes classification issues with linear regression
- Contrast linear and logistic regression
- Get to know an application of logistic regression

## **Reading Material**

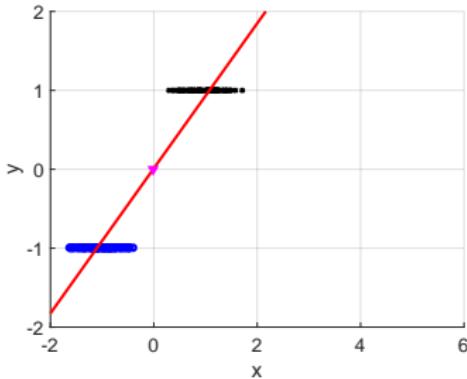
- K. Murphy; Machine Learning: A Probabilistic Perspective;  
Chapter 8

## **The Problem:** Linear regression for classification

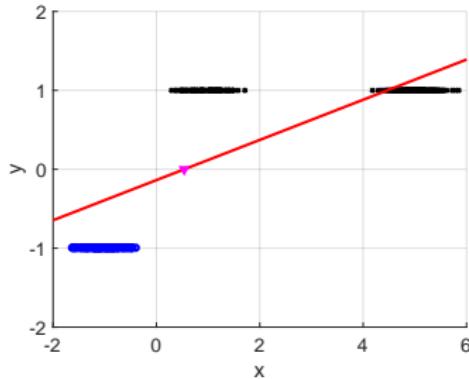
$$y^{(i)} \in \{-1, 1\}$$

1D-Model:

$$y^{(i)} = \text{sign}(w_1 x^{(i)} + w_0)$$



perfect classification



decision boundary shifted

Why is this?

## Why is this?

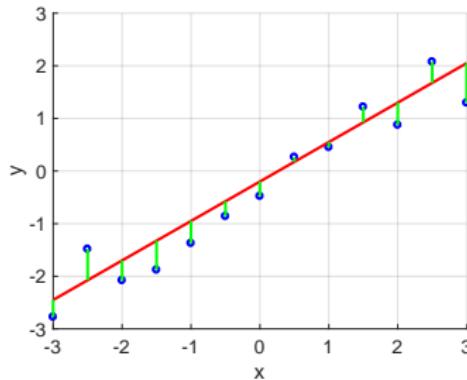
Assuming 1D-model

$$y = w_1 \cdot x + w_2$$

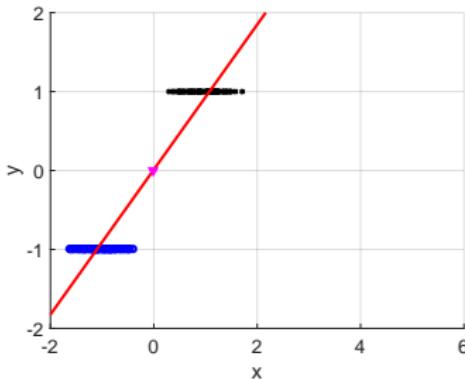
Linear regression finds parameters  $w_1$ ,  $w_2$  such that the squared error is small

$$\arg \min_{w_1, w_2} \frac{1}{2} \sum_{i=1}^N \left( y^{(i)} - w_1 \cdot x^{(i)} - w_2 \right)^2$$

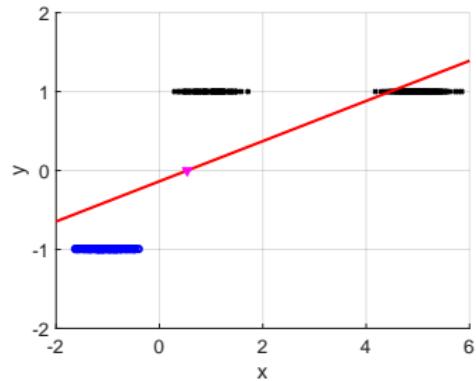
What exactly is the error?



In our case:



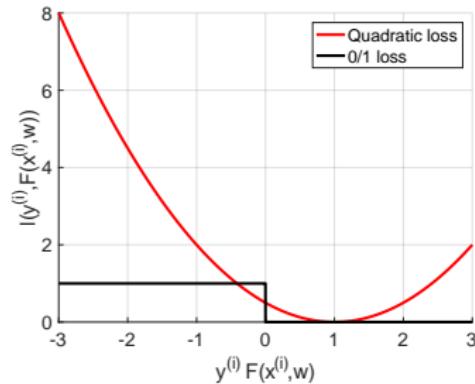
perfect classification



decision boundary shifted

## Linear regression: Quadratic loss (recall $y^{(i)} \in \{-1, 1\}$ )

$$\ell(y_i, \phi(x^{(i)})^\top w) = \frac{1}{2}(y^{(i)} - \phi(x^{(i)})^\top w)^2$$
$$= \frac{1}{2}(1 - y^{(i)} \underbrace{\phi(x^{(i)})^\top w}_{F(x^{(i)}, w)})^2$$
$$= \underbrace{F(x^{(i)}, w)}_{F(x^{(i)}, w, y^{(i)})}$$

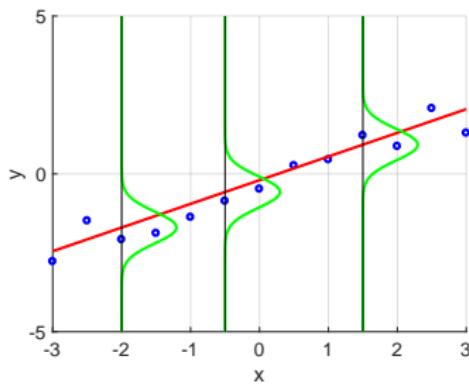


We penalize samples that are ‘very easy to classify.’

How to fix this?

A probabilistic interpretation of linear regression ( $y^{(i)} \in \mathbb{R}$ ):  
Model: Gaussian distribution

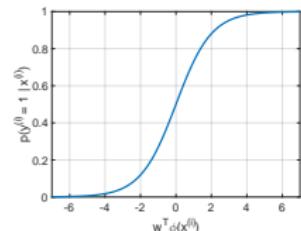
$$p(y^{(i)}|x^{(i)}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^\top \phi(x^{(i)}))^2\right)$$



## Logistic Regression:

Another probabilistic formulation for classification ( $y^{(i)} \in \{-1, 1\}$ ):

Model:



$$p(y^{(i)} = 1 | x^{(i)}) = \frac{1}{1 + \exp(-\mathbf{w}^T \phi(x^{(i)}))}$$
$$p(y^{(i)} = -1 | x^{(i)}) = 1 - p(y^{(i)} = 1 | x^{(i)}) = \frac{1}{1 + \exp(\mathbf{w}^T \phi(x^{(i)}))}$$

Taken together:

$$p(y^{(i)} | x^{(i)}) = \frac{1}{1 + \exp(-y^{(i)} \mathbf{w}^T \phi(x^{(i)}))}$$

What to do with this model?

$$p(y^{(i)}|x^{(i)}) = \frac{1}{1 + \exp(-y^{(i)} \mathbf{w}^T \phi(x^{(i)}))}$$

Recall that we are given a dataset  $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}$ . How about we choose  $\mathbf{w}$  which maximizes the likelihood/probability of this dataset?

### Assumption:

Samples/Data points are i.i.d.

$$p(\mathcal{D}) = \prod_{(x^{(i)}, y^{(i)}) \in \mathcal{D}} p(y^{(i)}|x^{(i)})$$

Choose  $\mathbf{w}$  to maximize probability:

$$\max_{\mathbf{w}} \prod_{(x^{(i)}, y^{(i)}) \in \mathcal{D}} p(y^{(i)}|x^{(i)})$$

Model:

$$p(y^{(i)}|x^{(i)}) = \frac{1}{1 + \exp(-y^{(i)} \mathbf{w}^T \phi(x^{(i)}))}$$

Task:

$$\arg \max_{\mathbf{w}} \prod_{(x^{(i)}, y^{(i)}) \in \mathcal{D}} p(y^{(i)}|x^{(i)}) = \arg \min_{\mathbf{w}} \sum_{(x^{(i)}, y^{(i)}) \in \mathcal{D}} -\log p(y^{(i)}|x^{(i)})$$

Combined:

$$\min_{\mathbf{w}} \sum_{(x^{(i)}, y^{(i)}) \in \mathcal{D}} \log \left( 1 + \exp(-y^{(i)} \mathbf{w}^T \phi(x^{(i)})) \right)$$

## Comparison

Linear regression

Logistic regression

Program:

$$\min_{\mathbf{w}} \sum_{(x^{(i)}, y^{(i)}) \in \mathcal{D}} \underbrace{\frac{1}{2} (1 - y^{(i)} \underbrace{\mathbf{w}^T \phi(x^{(i)})}_F)^2}_{F(x^{(i)}, \mathbf{w}, y^{(i)})}$$

Program:

$$\min_{\mathbf{w}} \sum_{(x^{(i)}, y^{(i)}) \in \mathcal{D}} \log \left( 1 + \exp \left( - y^{(i)} \underbrace{\mathbf{w}^T \phi(x^{(i)})}_F \right) \right) \underbrace{\vphantom{\log} \exp \left( - y^{(i)} \mathbf{w}^T \phi(x^{(i)}) \right)}_{F(x^{(i)}, \mathbf{w}, y^{(i)})}$$

**Empirical risk minimization:**

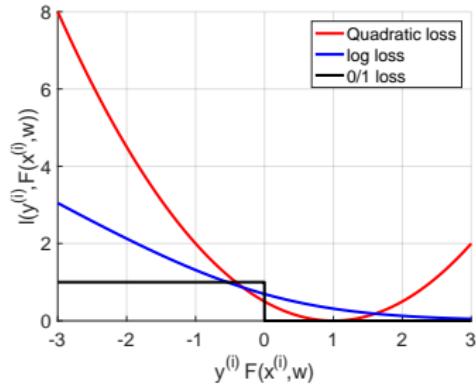
$$\min_{\mathbf{w}} \sum_{(x^{(i)}, y^{(i)}) \in \mathcal{D}} \ell(y^{(i)}, F(x^{(i)}, \mathbf{w}))$$

Linear regression:

$$\min_{\mathbf{w}} \sum_{(x^{(i)}, y^{(i)}) \in \mathcal{D}} \frac{1}{2} (1 - y^{(i)} \underbrace{\mathbf{w}^T \phi(x^{(i)})}_{F(x^{(i)}, \mathbf{w})})^2$$

Logistic regression:

$$\min_{\mathbf{w}} \sum_{(x^{(i)}, y^{(i)}) \in \mathcal{D}} \log \left( 1 + \exp(-y^{(i)} \underbrace{\mathbf{w}^T \phi(x^{(i)})}_{F(x^{(i)}, \mathbf{w})}) \right)$$



## How to optimize

$$\min_{\mathbf{w}} \sum_{(x^{(i)}, y^{(i)}) \in \mathcal{D}} \log \left( 1 + \exp(-y^{(i)} \underbrace{\mathbf{w}^T \phi(x^{(i)})}_{F(x^{(i)}, \mathbf{w})}) \right)$$

Can we set the gradient to zero and solve for  $\mathbf{w}$ ?

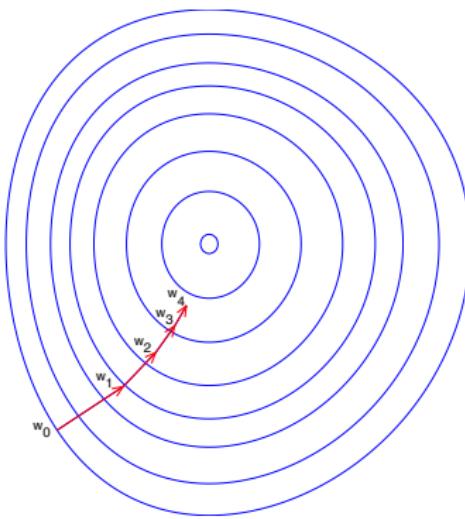
$$\sum_{(x^{(i)}, y^{(i)}) \in \mathcal{D}} \frac{-y^{(i)} \exp(-y^{(i)} \mathbf{w}^T \phi(x^{(i)}))}{1 + \exp(-y^{(i)} \mathbf{w}^T \phi(x^{(i)}))} \phi(x^{(i)}) = 0$$

No analytic solution for  $\mathbf{w}$  in general

## How to optimize

$$\min_{\mathbf{w}} \sum_{(x^{(i)}, y^{(i)}) \in \mathcal{D}} \log \left( 1 + \exp(-y^{(i)} \underbrace{\mathbf{w}^T \phi(x^{(i)}))}_{F(x^{(i)}, \mathbf{w})} \right)$$

Gradient descent: (walking down a mountain)



To solve

$$\min_{\mathbf{w}} f(\mathbf{w}) := \sum_{(x^{(i)}, y^{(i)}) \in \mathcal{D}} \log \left( 1 + \exp(-y^{(i)} \underbrace{\mathbf{w}^T \phi(x^{(i)})}_{F(x^{(i)}, \mathbf{w})}) \right)$$

we can use its gradient:

$$\nabla_{\mathbf{w}} f(\mathbf{w}) = \sum_{(x^{(i)}, y^{(i)}) \in \mathcal{D}} \frac{-y^{(i)} \exp(-y^{(i)} \mathbf{w}^T \phi(x^{(i)}))}{1 + \exp(-y^{(i)} \mathbf{w}^T \phi(x^{(i)}))} \phi(x^{(i)})$$

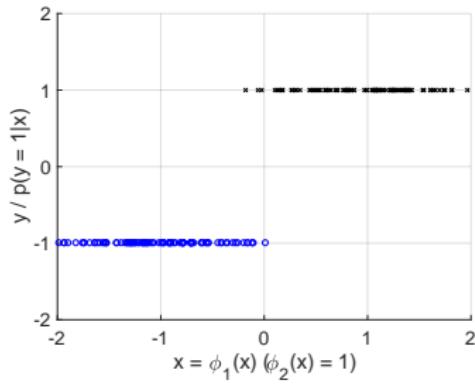
Simple algorithm: Initialize  $t = 0$ ,  $\mathbf{w}_t$ , and stepsize  $\alpha$

- Compute gradient  $\mathbf{g}_t = \nabla_{\mathbf{w}} f(\mathbf{w}_t)$
- Update parameters  $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \alpha \mathbf{g}_t$
- Update  $t \leftarrow t + 1$

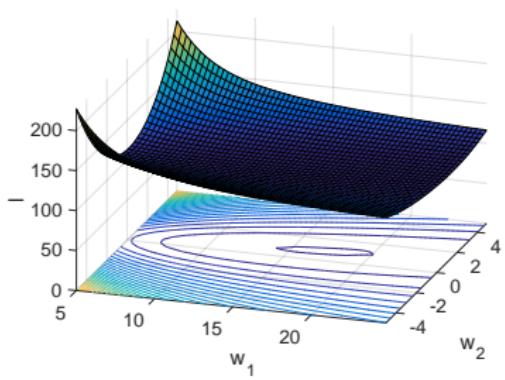
More complex algorithms may be ‘better.’

Example:

Data

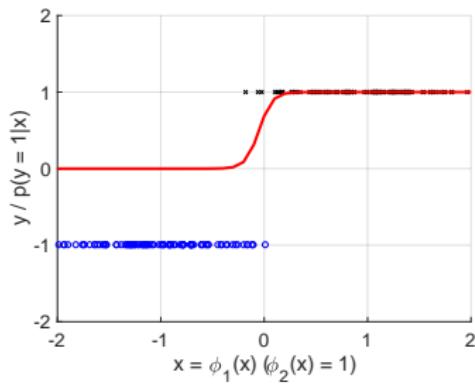


Loss

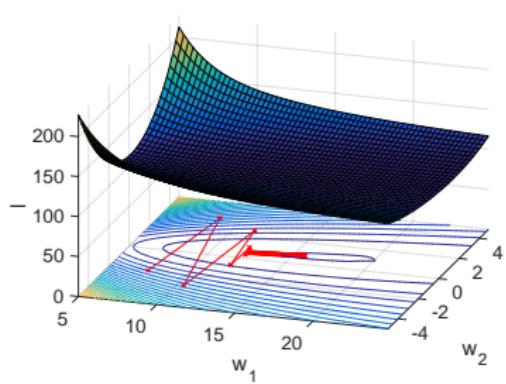


Example:

Data

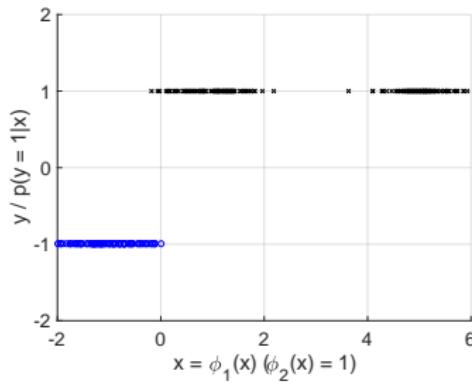


Loss

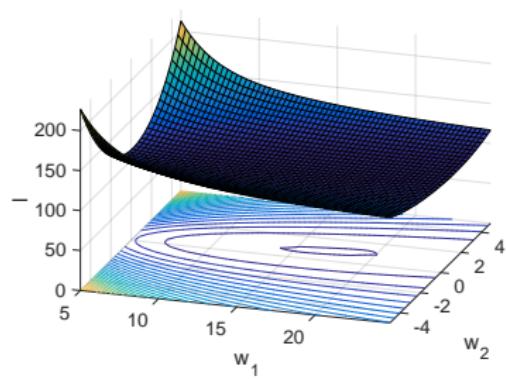


Example:

Data

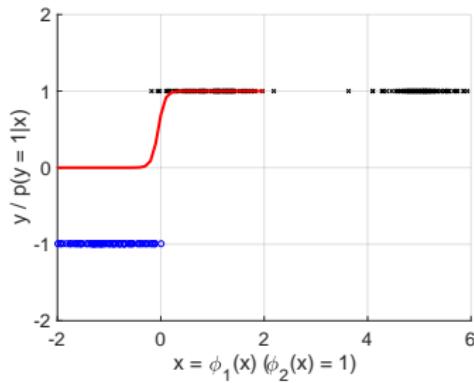


Loss

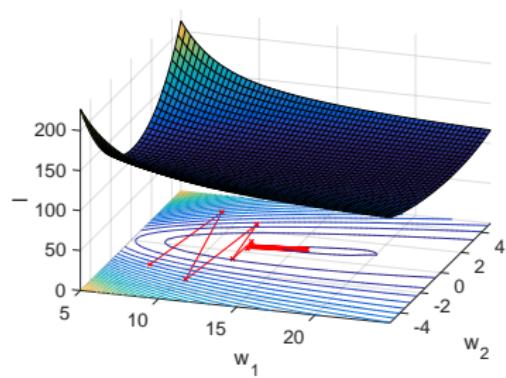


Example:

Data



Loss



## **Comparison:**

Linear regression:

- Closed form solution
- Gaussian probability model
- Not too well suited for classification

Logistic regression:

- Well suited for binary classification
- Logistic probability model
- No closed form solution

## **Quiz:**

- Which loss is used for logistic regression?
- What is the difference between logistic and linear regression?
- How to optimize linear and logistic regression?

## **Important topics of this lecture:**

- Linear regression
- Logistic regression

## **Up next:**

- Basics about optimization techniques