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

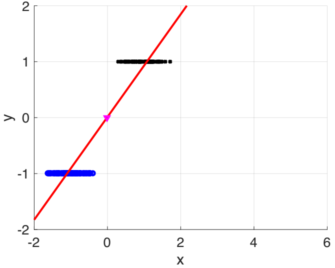
- K. Murphy; Machine Learning: A Probabilistic Perspective; Chapter 8

The Problem: Linear regression for classification

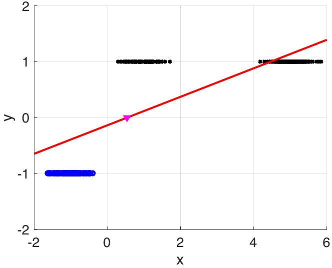
$$y^{(i)} \in \{-1, 1\}$$

1D-Model:

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



perfect classification



decision boundary shifted

Why is this?

Why is this?

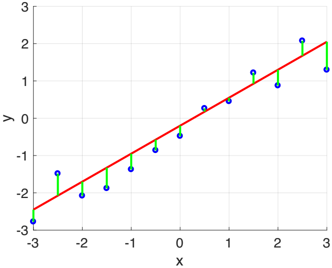
Assuming 1D-model

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

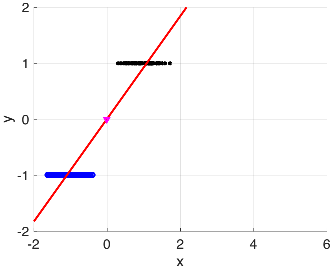
Linear regression finds parameters w_1, w_2 such that the squared error is small

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

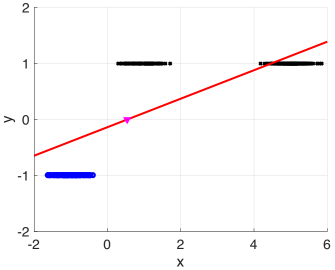
What exactly is the error?



In our case:



perfect classification

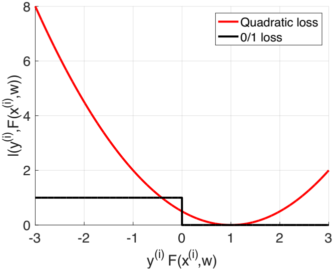


decision boundary shifted

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

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

$$\overset{\left(y^{(i)}\right)^2=1}{=} \frac{1}{2}(1 - \underbrace{y^{(i)} \underbrace{\phi(x^{(i)})^\top \boldsymbol{w}}_{F(x^{(i)},\boldsymbol{w})}}_{F(x^{(i)},\boldsymbol{w},y^{(i)})})^2$$



We penalize samples that are 'very easy to classify.'

How to fix this?

$$P(y \mid X, w) = \mathrm{Ber}(y \mid \mu(x)) \; ; \quad \text{where}$$

$$\mu(x) = E[y \mid x] = p(y=1 \mid x)$$

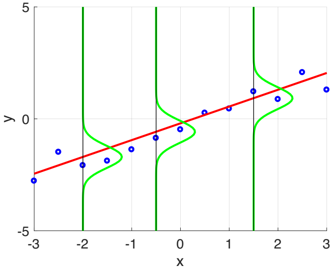
Next: define: $\mu(x) = \mathrm{Sigm}(w^T x)$, where

$$\mathrm{Sigm}(\eta) \triangleq \frac{1}{1+\exp(-\eta)} = \frac{e^{\eta}}{e^{\eta}+1}$$

$$\Longrightarrow P(y \mid X, w) = \mathrm{Ber}(y \mid \mathrm{Sigm}(w^T X))$$

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

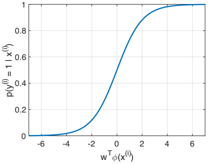
$$p(y^{(i)}|x^{(i)}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y^{(i)} - \boldsymbol{w}^\top \phi(x^{(i)}))^2\right)$$



Logistic Regression:

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

Model:



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

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

Taken together:

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

What to do with this model?

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

Recall that we are given a dataset $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}$. How about we choose $\boldsymbol{w}$ 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 $\boldsymbol{w}$ to maximize probability:

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

Model:

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

Task:

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

Combined:

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

Comparison

Linear regression

Program:

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

Logistic regression

Program:

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

Empirical risk minimization:

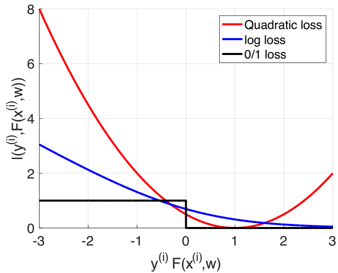
$$\min_{\boldsymbol{w}} \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \ell(y^{(i)}, F(x^{(i)}, w))$$

Linear regression:

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

Logistic regression:

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



How to optimize

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

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

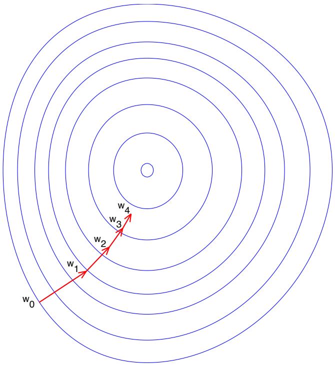
$$\sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \frac{-y^{(i)}\exp(-y^{(i)}\boldsymbol{w}^T\phi(x^{(i)}))}{1+\exp(-y^{(i)}\boldsymbol{w}^T\phi(x^{(i)}))}\phi(x^{(i)}) = 0$$

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

How to optimize

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

Gradient descent: (walking down a mountain)



To solve

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

we can use its gradient:

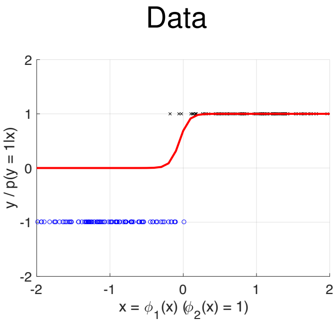
$$\nabla_{\boldsymbol{w}} f(\boldsymbol{w}) = \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \frac{-y^{(i)}\exp(-y^{(i)}\boldsymbol{w}^T\phi(x^{(i)}))}{1+\exp(-y^{(i)}\boldsymbol{w}^T\phi(x^{(i)}))}\phi(x^{(i)})$$

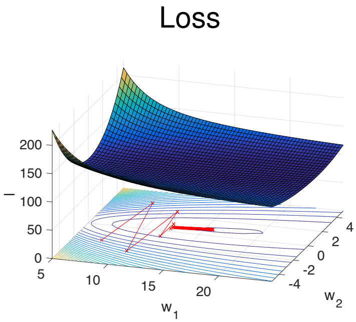
Simple algorithm: Initialize $t=0$, $\boldsymbol{w}_t$, and stepsize α

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

More complex algorithms may be 'better.'

Example:


Data
2
1
0
-1
-2
y / p(y = 1|x)
-2
-1
0
1
2
x = φ1(x) (φ2(x) = 1)


Loss
200
150
100
50
0
l
5
10
15
20
w1
4
2
0
-2
-4
w2

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