

Machine Learning

Lecture 2 : Introduction to Classification, Logistic Regression

Mohamad GHASSANY

EFREI PARIS

Classification

- ▶ Email: Spam / Not Spam?
- ▶ Online Transactions: Fraudulent (Yes/No)?
- ▶ Tumor: Malignant / Benign?
- ▶ Loan Demand (Credit Risk): Safe / Risky

- ▶ Email: Spam / Not Spam?
- ▶ Online Transactions: Fraudulent (Yes/No)?
- ▶ Tumor: Malignant / Benign?
- ▶ Loan Demand (Credit Risk): Safe / Risky

Classification: categorical output

- ▶ $y \in \{0, 1\}$
- ▶ 0: "Negative class"
- ▶ 1: "Positive Class"

- ▶ Email: Spam / Not Spam?
- ▶ Online Transactions: Fraudulent (Yes/No)?
- ▶ Tumor: Malignant / Benign?
- ▶ Loan Demand (Credit Risk): Safe / Risky

Classification: categorical output

- ▶ $y \in \{0, 1\}$
- ▶ 0: "Negative class"
- ▶ 1: "Positive Class"

.. and also multiclass classification

$$\text{Accuracy} = \frac{\text{Number of data points classified correctly}}{\text{all data points}}$$

Confusion Matrix

.. while in Regression (continuous output): Mean Squared Error (MSE).

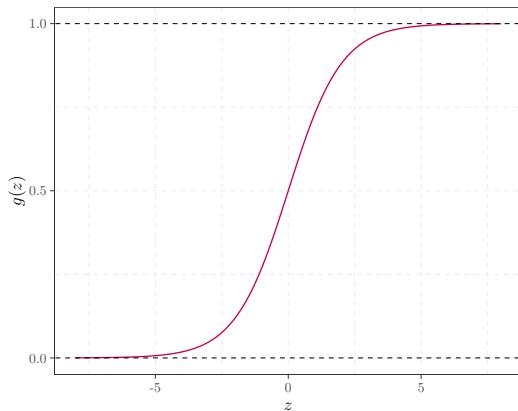
Logistic Regression

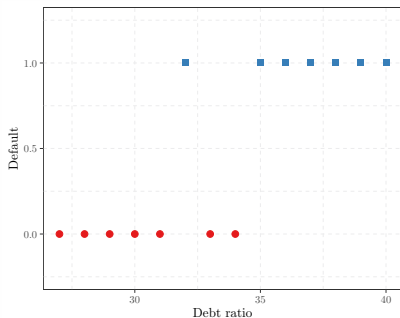
The logistic function (sigmoid)

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

The logistic function (sigmoid)

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





► $y \in \{0, 1\}$:

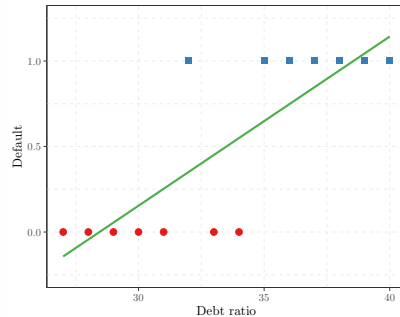
- "0": Negative class (here **no default**)
- "1": Positive class (here **default**)

► $f_{\omega}(x) = \omega'x$ can be > 1 ou < 0 !

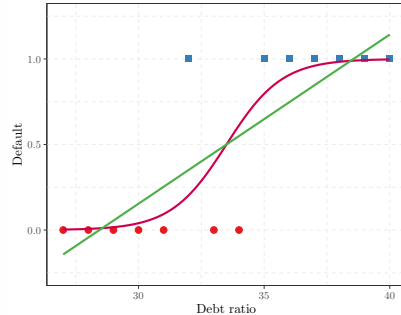
► Ideally $0 \leq f_{\omega}(x) \leq 1$ s.t.:

- If $f_{\omega}(x) \geq 0.5$, predict " $y = 1$ "
- If $f_{\omega}(x) < 0.5$, predict " $y = 0$ "

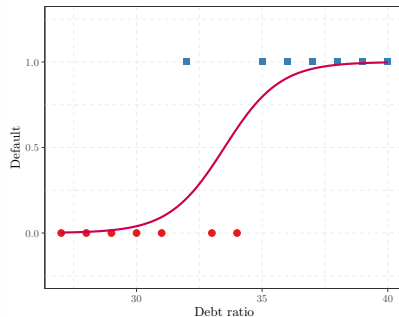
► Let $f_{\omega}(x) = \omega'x$



► Let $f_{\omega}(x) = \cancel{\omega'x} = g(\omega'x) = \frac{1}{1 + e^{-\omega'x}}$



- ▶ $0 \leq g(\omega'x) \leq 1$
- ▶ $f_{\omega}(x) = g(\omega'x)$ = estimated probability that $y = 1$ on input x
- ▶ Probability that $y = 1$, given x , parameterized by ω
- ▶ $g(\omega'x) = p(y = 1 | x) = p(x)$
- ▶ $y \in \{0, 1\}$ so $p(y = 1 | x) + p(y = 0 | x) = 1$



logistic score

$$p(x) = p(y = 1 | x) = \frac{e^{\omega'x}}{1 + e^{\omega'x}} = \frac{1}{1 + e^{-\omega'x}}$$

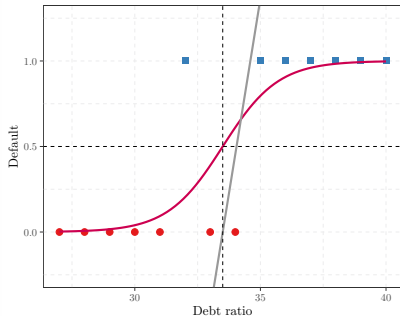
odds (côtes)

$$\frac{p(x)}{1 - p(x)} = e^{\omega'x}$$

log-odds or logit (logarithme des côtes)

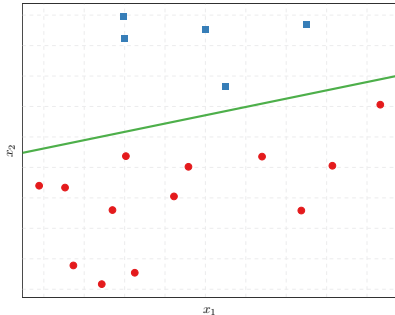
$$\log\left(\frac{p(x)}{1 - p(x)}\right) = \omega'x$$

Logistic Regression: decision boundary



► We predict " $y = 1$ " if $p(x) \geq 0.5$ which means $\omega'x \geq 0$

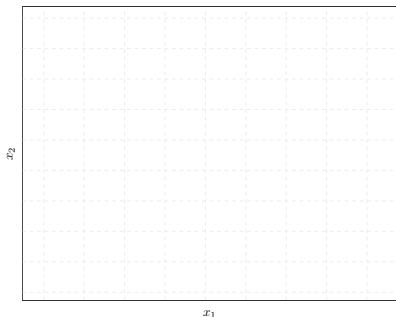
► $\omega_0 + \omega_1 x \geq 0 \Rightarrow x \geq -\frac{\omega_0}{\omega_1}$



- ▶ $p(x) = p(y = 1 | x) = f_{\omega}(x) = g(\omega'x)$
- ▶ Predict “ $y = 1$ ” if $p(x) \geq 0.5$ which means $\omega'x \geq 0$

- ▶ $\omega_0 + \omega_1 x_1 + \omega_2 x_2 \geq 0$ So

$$x_2 \geq -\frac{\omega_1}{\omega_2} x_1 - \frac{\omega_0}{\omega_2}$$



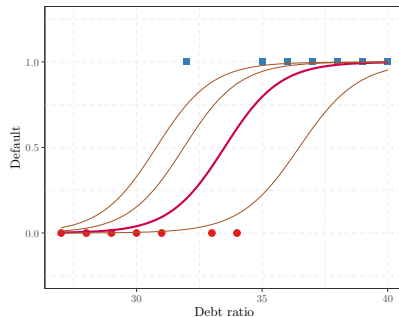
- ▶ Let

$$f_{\omega}(x) = g(\omega_0 + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_1^2 + \omega_4 x_2^2)$$
- ▶ For example, predict “ $y = 1$ ” if

$$-1 + x_1^2 + x_2^2 \geq 0$$
- ▶ Or, $f_{\omega}(x) = g(\omega_0 + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_1^2 + \omega_4 x_1^2 x_2 + \omega_5 x_1^2 x_2^2 + \dots)$

Logistic Regression: model estimation

- Parameters to estimate: $\omega = \{\omega_0, \omega_1\}$ if univariate
- $\omega = \{\omega_0, \omega_1, \dots, \omega_p\}$ if multivariate with p features
- How to choose parameters ω ?



¹check: <https://shiny serv.es/shiny/log-maximum-likelihood/>, by Eduardo García Portugués

Cost function of simple linear regression

- ▶ Model: $f_{\omega}(x) = \omega_0 + \omega_1 x = \omega'x$
- ▶ Parameters: ω_0 and ω_1
- ▶ Cost function: $J(\omega_0, \omega_1) = \frac{1}{n} \sum_{i=1}^n \frac{1}{2} (f_{\omega}(x^{(i)}) - y^{(i)})^2$
- ▶ Goal: $\min_{\omega_0, \omega_1} J(\omega_0, \omega_1)$

Non-convex in case of logistic regression !

- ▶ How to choose parameters ω ?
- ▶ $y \in \{0, 1\}$, Let's assume:

$$p(y = 1 \mid x, \omega) = f_{\omega}(x)$$

$$p(y = 0 \mid x, \omega) = 1 - f_{\omega}(x)$$

- ▶ How to choose parameters ω ?
- ▶ $y \in \{0, 1\}$, Let's assume:

$$p(y = 1 \mid x, \omega) = f_{\omega}(x)$$

$$p(y = 0 \mid x, \omega) = 1 - f_{\omega}(x)$$

- ▶ We represent $y \mid x, \omega \sim \mathcal{B}(f_{\omega}(x))$
- ▶ We can write:

$$p(y \mid x, \omega) = (f_{\omega}(x))^y (1 - f_{\omega}(x))^{1-y} \quad y \in \{0, 1\}$$

- ▶ How to choose parameters ω ?
- ▶ $y \in \{0, 1\}$, Let's assume:

$$p(y = 1 \mid x, \omega) = f_{\omega}(x)$$

$$p(y = 0 \mid x, \omega) = 1 - f_{\omega}(x)$$

- ▶ We represent $y \mid x, \omega \sim \mathcal{B}(f_{\omega}(x))$
- ▶ We can write:

$$p(y \mid x, \omega) = (f_{\omega}(x))^y (1 - f_{\omega}(x))^{1-y} \quad y \in \{0, 1\}$$

- ▶ Given the n observations and assuming independance, we estimate ω by maximizing the **likelihood**:

$$\mathcal{L}(\omega) = \prod_{i=1}^n p(y^{(i)} \mid x^{(i)}, \omega)$$

► The likelihood:

$$\begin{aligned}\mathcal{L}(\omega) &= \prod_{i=1}^n p(y^{(i)} | x^{(i)}, \omega) \\ &= \prod_{i=1}^n (f_{\omega}(x^{(i)}))^{y^{(i)}} (1 - f_{\omega}(x^{(i)}))^{1-y^{(i)}}\end{aligned}$$

- The likelihood:

$$\begin{aligned}\mathcal{L}(\omega) &= \prod_{i=1}^n p(y^{(i)} | x^{(i)}, \omega) \\ &= \prod_{i=1}^n (f_{\omega}(x^{(i)}))^{y^{(i)}} (1 - f_{\omega}(x^{(i)}))^{1-y^{(i)}}\end{aligned}$$

- Maximizing the likelihood is same as maximizing its log:

$$\begin{aligned}\ell(\omega) &= \log(\mathcal{L}(\omega)) \\ &= \sum_{i=1}^n y^{(i)} \log f_{\omega}(x^{(i)}) + (1 - y^{(i)}) \log(1 - f_{\omega}(x^{(i)}))\end{aligned}$$

- Maximizing $\ell(\omega)$ is same as minimizing: $-\frac{1}{n}\ell(\omega)$

- The **likelihood**:

$$\begin{aligned}\mathcal{L}(\omega) &= \prod_{i=1}^n p(y^{(i)} | x^{(i)}, \omega) \\ &= \prod_{i=1}^n (f_{\omega}(x^{(i)}))^{y^{(i)}} (1 - f_{\omega}(x^{(i)}))^{1-y^{(i)}}\end{aligned}$$

- Maximizing the likelihood is same as maximizing its log:

$$\begin{aligned}\ell(\omega) &= \log(\mathcal{L}(\omega)) \\ &= \sum_{i=1}^n y^{(i)} \log f_{\omega}(x^{(i)}) + (1 - y^{(i)}) \log(1 - f_{\omega}(x^{(i)}))\end{aligned}$$

- Maximizing $\ell(\omega)$ is same as minimizing: $-\frac{1}{n}\ell(\omega)$
- Let $J(\omega) = -\frac{1}{n}\ell(\omega)$, a **convex cost function** for the logistic regression model (known as *binary cross entropy*).

- ▶ **Goal:** Find ω s.t. $\omega = \operatorname{argmin}_{\omega} J(\omega)$
- ▶ $J(\omega) = -\frac{1}{n} \sum_{i=1}^n y^{(i)} \log f_{\omega}(x^{(i)}) + (1 - y^{(i)}) \log (1 - f_{\omega}(x^{(i)}))$
- ▶ Contrary to the linear regression, this cost function **does not** have an **analytical** solution. We need an optimization technique.

- ▶ **Goal:** Find ω s.t. $\omega = \operatorname{argmin}_{\omega} J(\omega)$
- ▶ $J(\omega) = -\frac{1}{n} \sum_{i=1}^n y^{(i)} \log f_{\omega}(x^{(i)}) + (1 - y^{(i)}) \log (1 - f_{\omega}(x^{(i)}))$
- ▶ Contrary to the linear regression, this cost function **does not** have an **analytical** solution. We need an optimization technique.

GD for logistic regression

- ▶ initialize ω 'randomly'
- ▶ repeat until convergence{

$$\omega_i^{\text{new}} = \omega_i^{\text{old}} - \alpha \frac{\partial J(\omega)}{\partial \omega_i}$$

simultaneously for $i = 0, \dots, p$ }

- ▶ Recall that $g(z) = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}}$
- ▶ Notice that $g'(z) = g(z)(1 - g(z))$

- ▶ Recall that $g(z) = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}}$
- ▶ Notice that $g'(z) = g(z)(1 - g(z))$
- ▶ $\frac{\partial J(\omega)}{\partial \omega_i} = (y - f_\omega(x))x_i$

- ▶ Recall that $g(z) = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}}$
- ▶ Notice that $g'(z) = g(z)(1 - g(z))$
- ▶ $\frac{\partial J(\omega)}{\partial \omega_i} = (y - f_\omega(x))x_i$

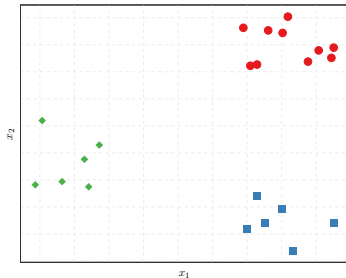
GD for logistic regression

- ▶ initialize ω randomly
- ▶ repeat until convergence{

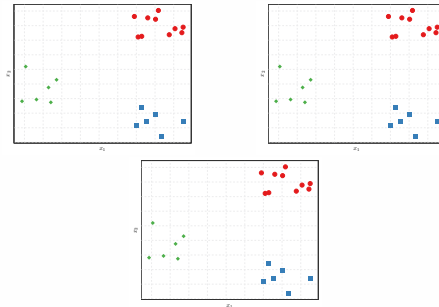
$$\omega_i^{\text{new}} = \omega_i^{\text{old}} - \alpha \frac{1}{n} \sum_{i=1}^n (f_\omega(x^{(i)}) - y^{(i)}) \cdot x_i^{(i)}$$

simultaneously for $i = 0, \dots, p \}$

- Weather: Sunny, Cloudy, Rain, Snow
- Medical diagrams: Not ill, Cold, Flu
- News articles: Sport, Education, Technology, Politics



- $f_{\omega}^{(i)}(x) = P(y = i|x, \omega)$ for $i = 1, 2, 3$
- Train a logistic regression classifier for each class i to predict the probability that $y = i$
- On a new input x , to make a prediction, pick the class i that maximizes $f_{\omega}^{(i)}(x)$



- ▶ Very famous method and maybe the most used
- ▶ Adapted for a binary y
- ▶ Relation with linear regression
- ▶ Linear decision boundary, but can be non linear using other hypothesis
- ▶ Direct calculation of $p(y = 1 | x)$