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"

Classification examples

- 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).

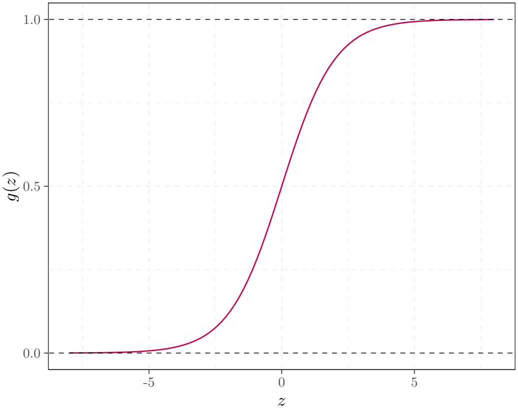
Logistic Regression

The logistic function (sigmoid)

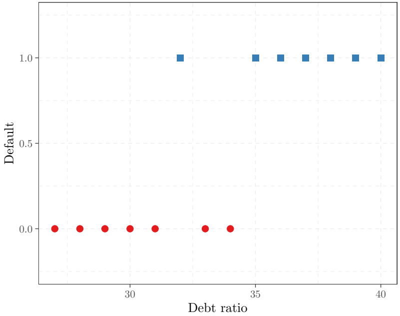
$$g(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$$

The logistic function (sigmoid)

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

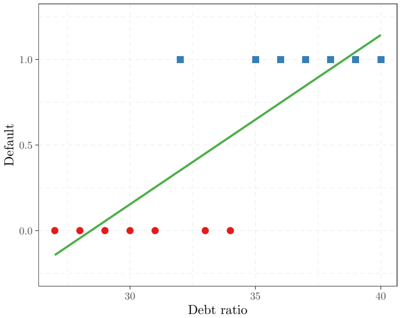


Logistic Regression: why not linear regression

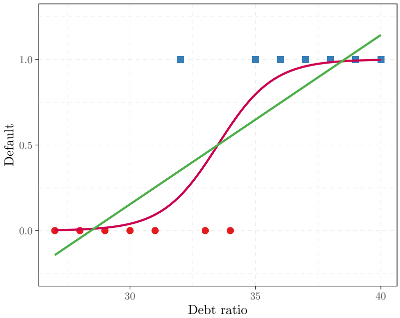


- $y \in \{0, 1\}$:
  - "0": Negative class (here no default)
  - "1": Positive class (here default)
- $f_\omega(x) = \omega' x$ can be > 1 ou < 0 !
- Ideally $0 \leqslant f_\omega(x) \leqslant 1$ s.t.:
  - If $f_\omega(x) \geqslant 0.5$, predict "$y = 1$"
  - If $f_\omega(x) < 0.5$, predict "$y = 0$"

- Let $f_\omega(x) = \omega' x$

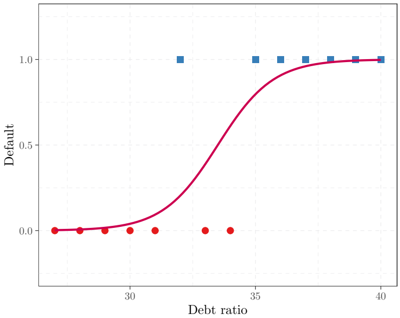


- Let $f_\omega(x) = \cancel{\omega' x} = g(\omega' x) = \dfrac{1}{1 + e^{-\omega' x}}$



Logistic Regression: intuition (3)

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



Logistic Regression: odds & log-odds (logit)

logistic score

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

odds (côtes)

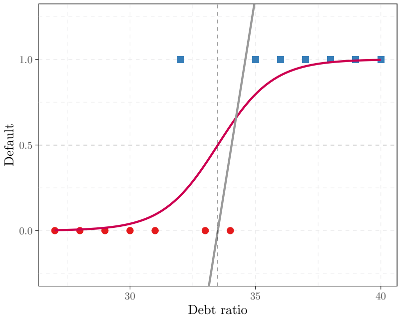
$$\frac{p(x)}{1 - p(x)} = e^{\omega' x}$$

log-odds or logit (logarithme des côtes)

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

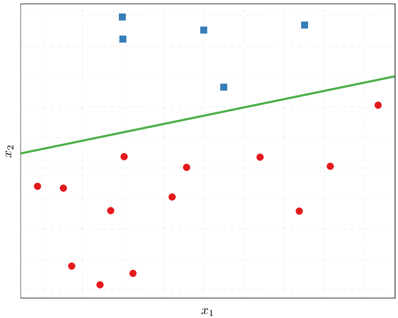
Logistic Regression: decision boundary

Logistic Regression: decision boundary



- We predict "$y = 1$" if $p(x) \geqslant 0.5$ which means $\omega' x \geqslant 0$
- $\omega_0 + \omega_1 x \geqslant 0 \Rightarrow x \geqslant -\dfrac{\omega_0}{\omega_1}$

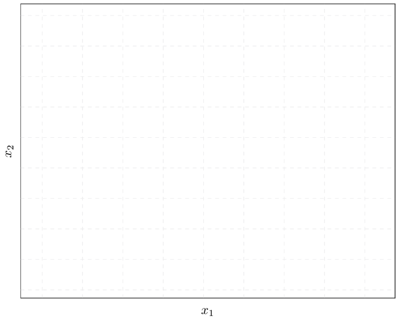
Logistic Regression: decision boundary (2 features)



- $p(x) = p(y = 1 \mid x) = f_\omega(x) = g(\omega' x)$
- Predict "$y = 1$" if $p(x) \geqslant 0.5$ which means $\omega' x \geqslant 0$
- $\omega_0 + \omega_1 x_1 + \omega_2 x_2 \geqslant 0$ So

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

Non linear decision boundaries



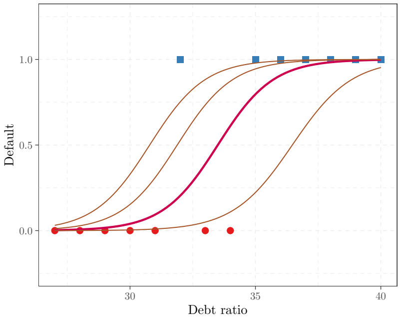
- Let $f_\omega(x) = g(\omega_0 + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_1^2 + \omega_4 x_2^2)$

- For example, predict "$y = 1$" if $-1 + x_1^2 + x_2^2 \geqslant 0$

- 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
- $\omega = \{\omega_0, \omega_1, \ldots, \omega_p\}$ if multivariate with p features
- How to choose parameters ω?



[1] check: https://shinyserv.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: $J(\omega_0, \omega_1) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \left(f_\omega \left(x^{(i)} \right) - y^{(i)} \right)^2$
- 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:

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

Logistic Regression: how to estimate the parameters

- 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:

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

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

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

- The likelihood:

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

- The likelihood:

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

- Maximizing the likelihood is same as maximizing its log:

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

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

Logistic Regression: model estimation

- The likelihood:

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

- Maximizing the likelihood is same as maximizing its log:

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

- 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. $\omega = \text{argmin}_{\omega} J(\omega)$
- $J(\omega) = -\dfrac{1}{n}\sum_{i=1}^{n} y^{(i)} \log f_{\omega}\left(x^{(i)}\right) + \left(1 - y^{(i)}\right)\log\left(1 - f_{\omega}\left(x^{(i)}\right)\right)$
- Contrary to the linear regression, this cost function **does not** have an analytical solution. We need an optimization technique.

- **Goal**: Find ω s.t. $\omega = \text{argmin}_\omega J(\omega)$
- $J(\omega) = -\frac{1}{n}\sum_{i=1}^{n} y^{(i)} \log f_\omega\left(x^{(i)}\right) + \left(1 - y^{(i)}\right)\log\left(1 - f_\omega\left(x^{(i)}\right)\right)$
- 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`{

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

`simultaneously for` $i = 0, \ldots, p$ }

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

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

- $\dfrac{\partial J(\omega)}{\partial \omega_i} = (y - f_\omega(x))x_i$

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

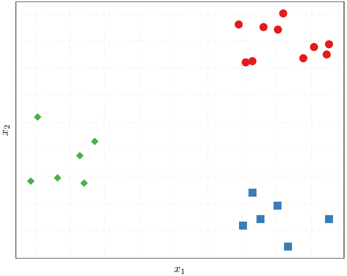
GD for logistic regression

- `initialize` ω `randomly`
- `repeat until convergence`$\{$

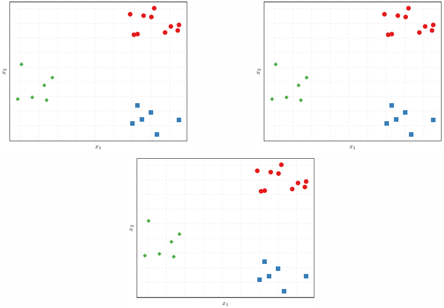
$$\omega_i^{new} = \omega_i^{old} - \alpha \frac{1}{n} \sum_{i=1}^{n} \left(f_\omega\left(x^{(i)}\right) - y^{(i)}\right).x_i^{(i)}$$

`simultaneously for` $i = 0, \ldots, p$ $\}$

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



- $f_\omega^{(i)}(x) = P(y = 1|x, \omega) \quad$ 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 $p(y = 1 \mid x)$