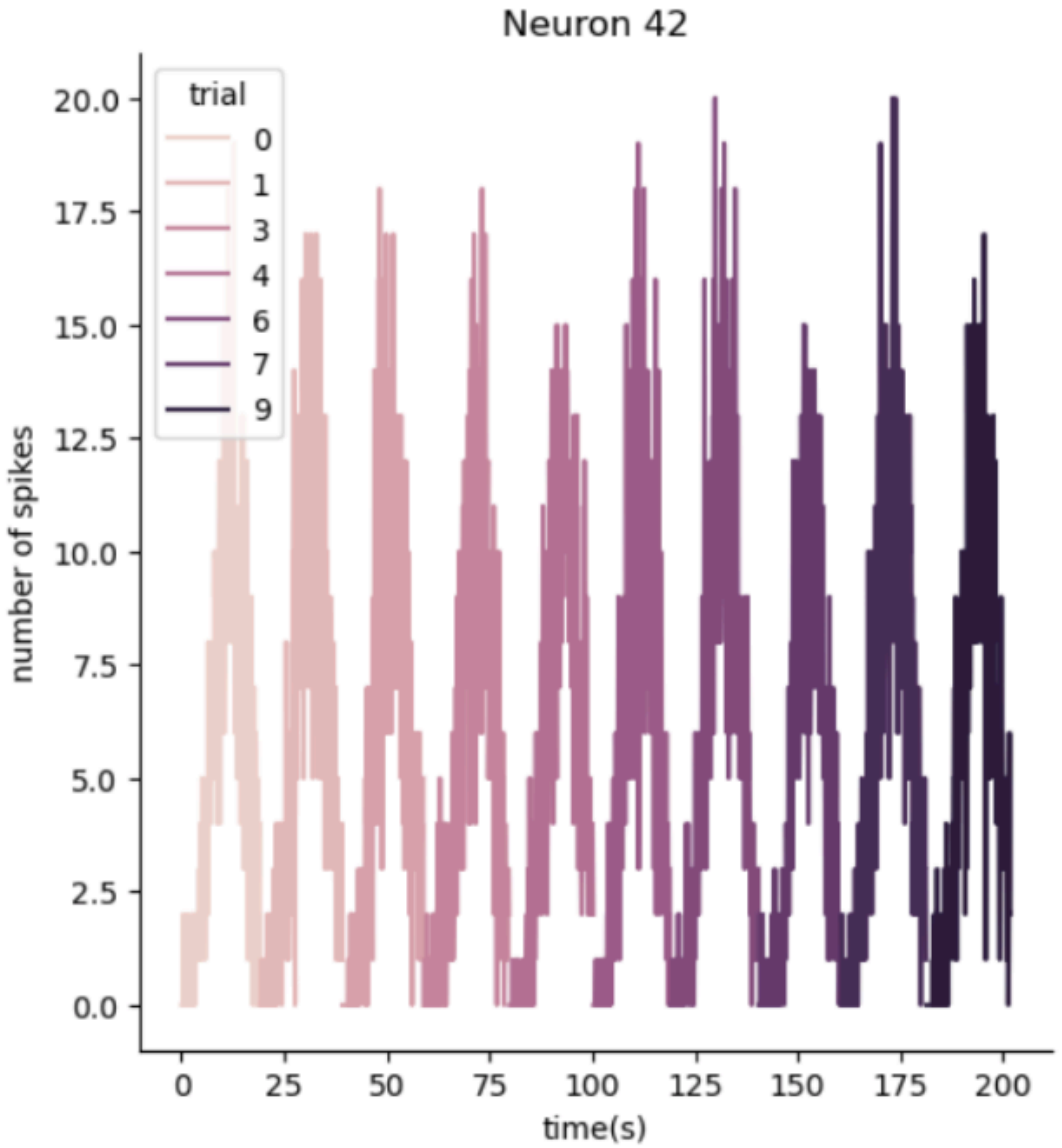
Data transformations in linear models

BIOENG-210 | Biological data science I: statistical learning

EPFL

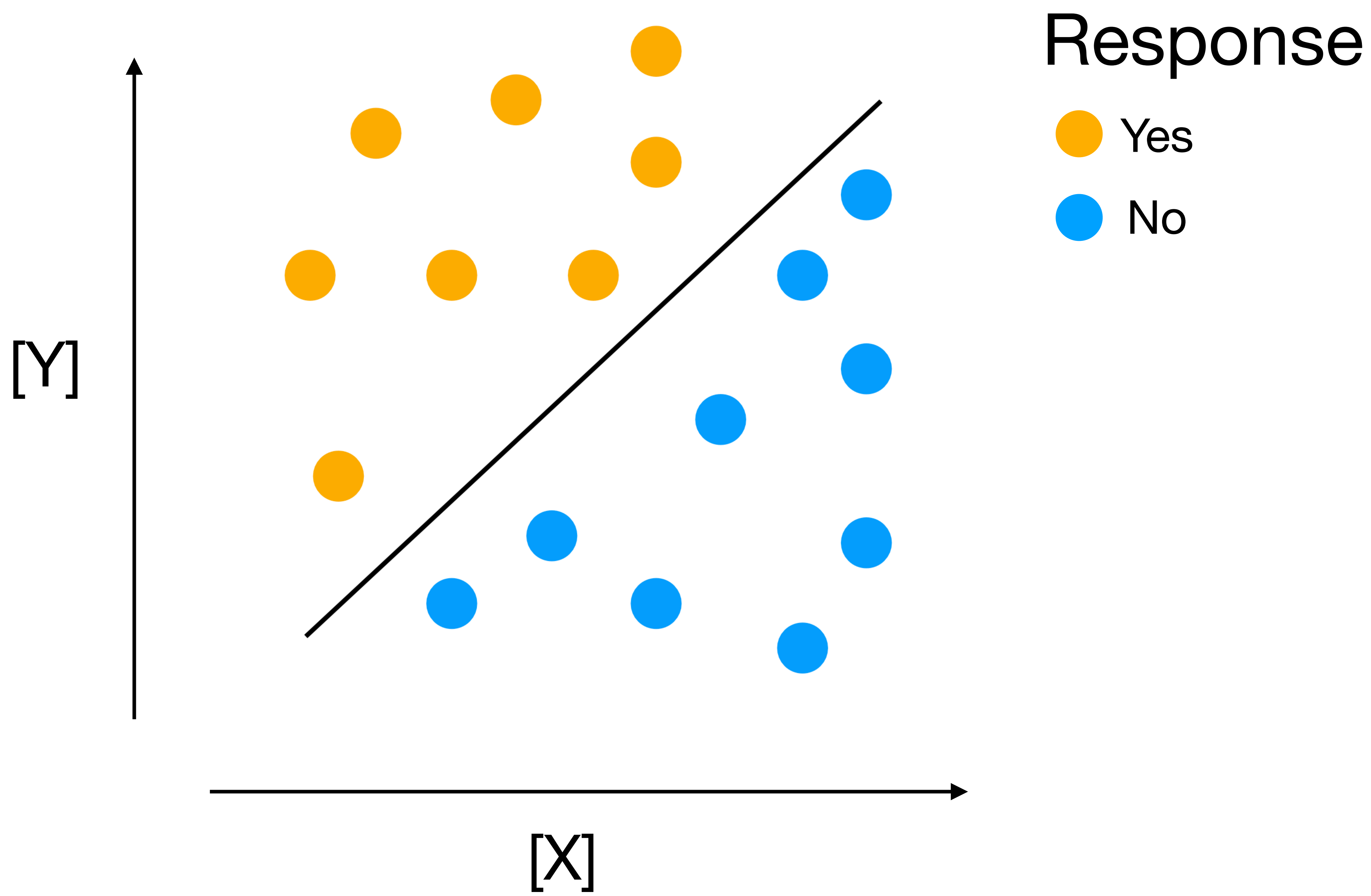
Hands-on 7: Generalized linear models

Poisson GLM



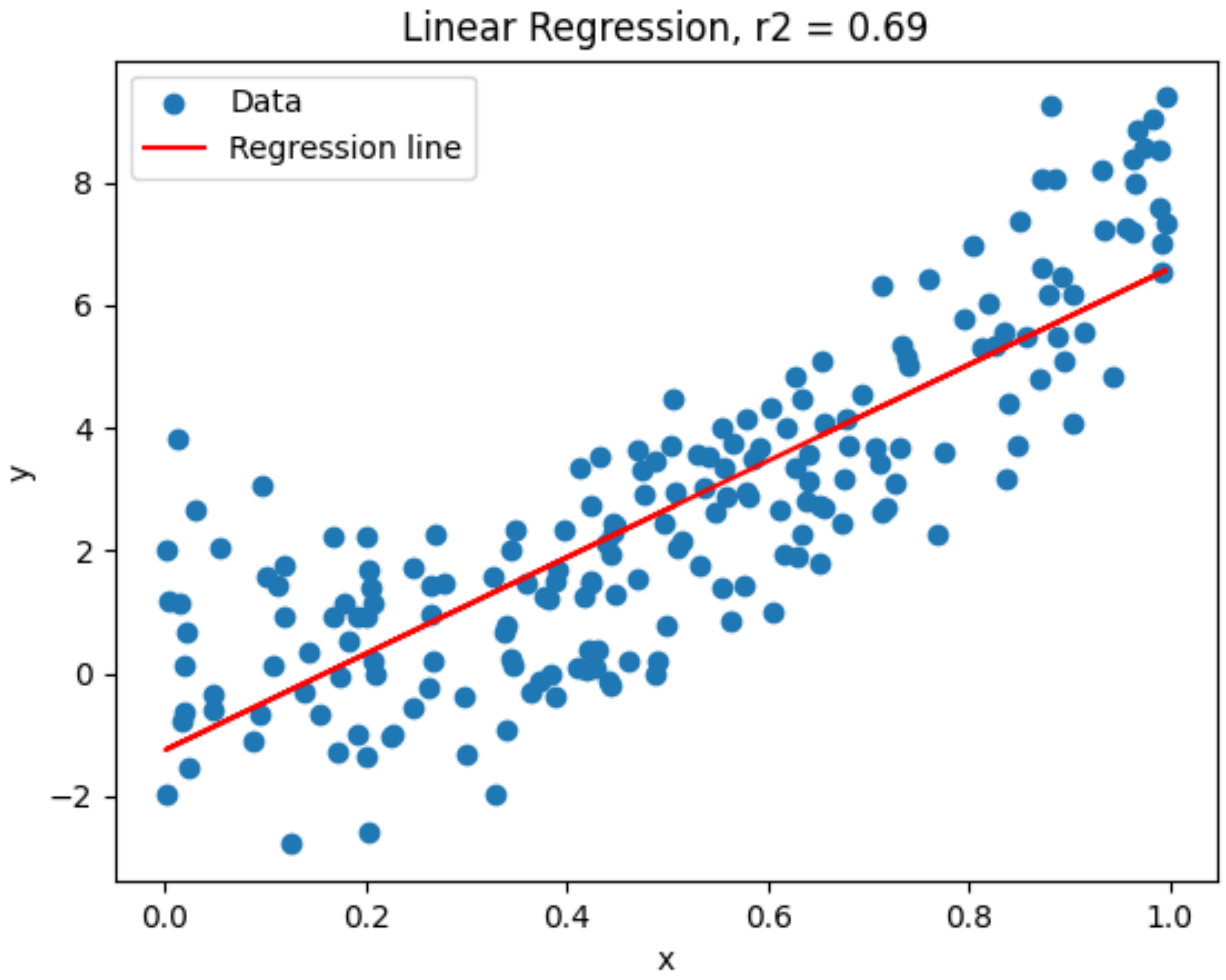
$$\log y = \beta_2 x^2 + \beta_1 x + \beta_0$$

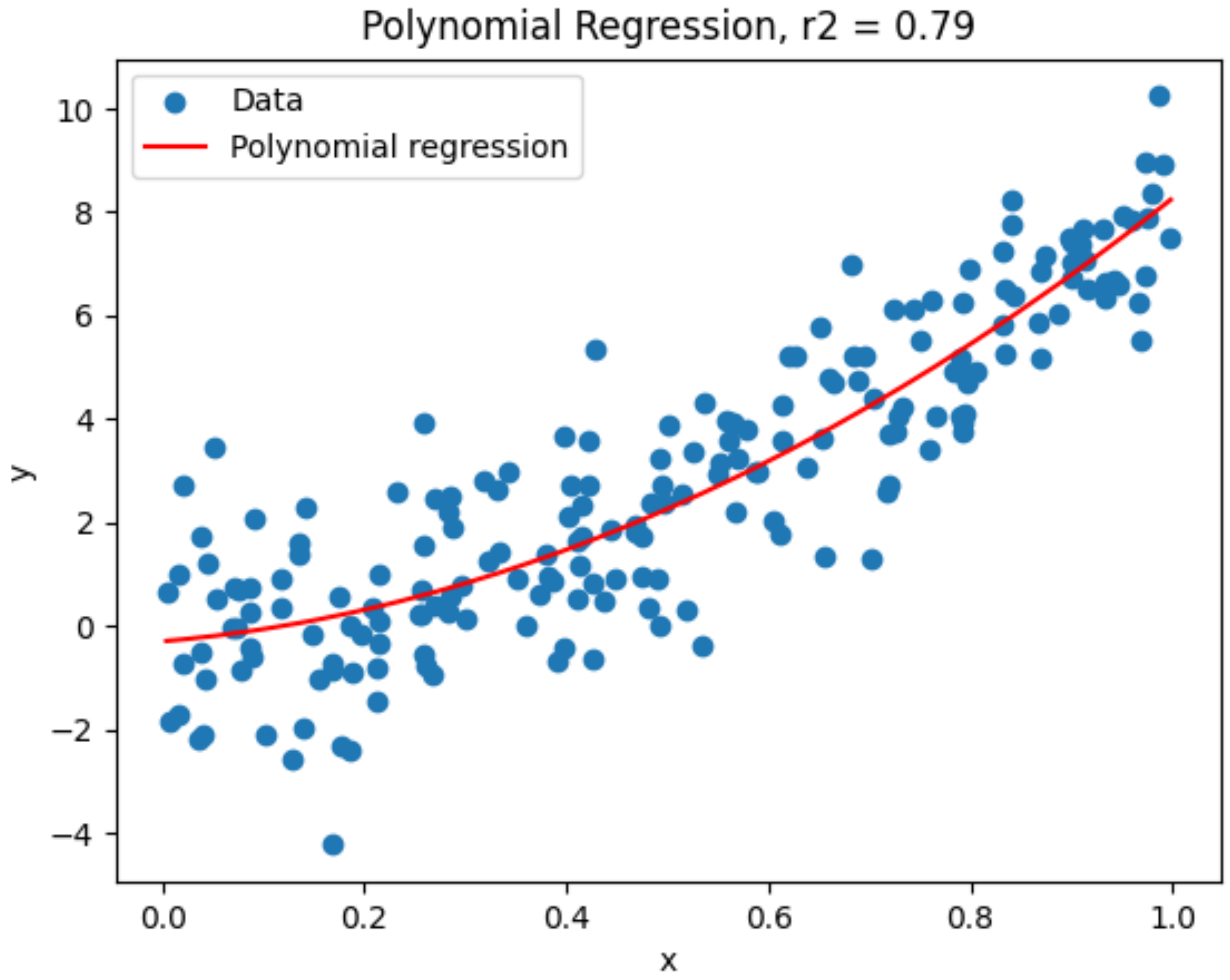
Logistic regression



$$p = \sigma(\beta_x [X] + \beta_y [Y] + \beta_0)$$

Linear models: beyond linearity





Applying transformations to features

Do not confuse with link functions! (Applied in the GLM to the target)

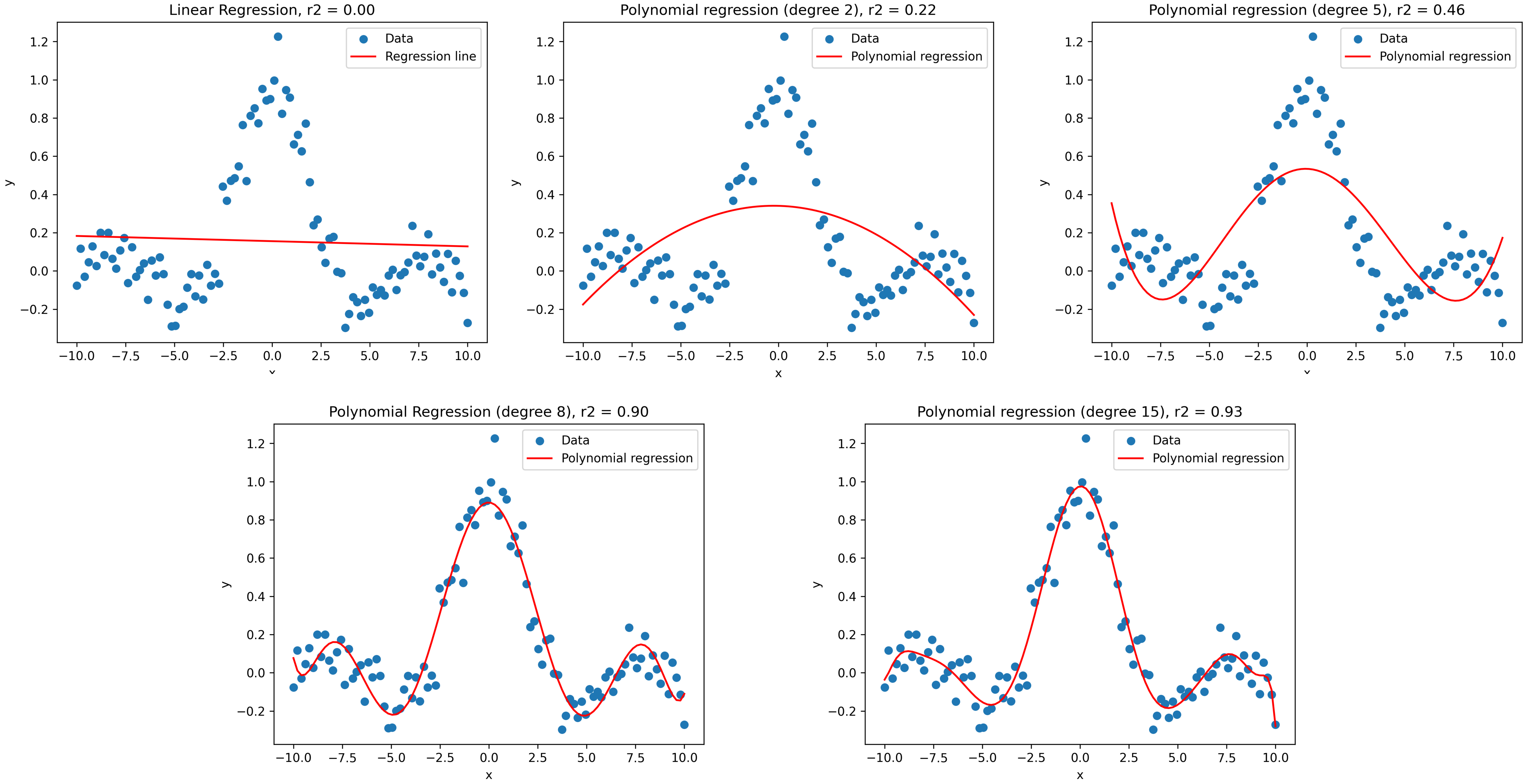
Design matrix

$$\text{n\_samples} \quad \updownarrow \begin{pmatrix} x & \text{Intercept} \\ 1.3 & 1 \\ 2.4 & 1 \\ 0.3 & 1 \\ 0.9 & 1 \\ 1.6 & 1 \\ 3.9 & 1 \\ 0.1 & 1 \\ \ldots & \ldots \end{pmatrix}$$

Design matrix with transformation

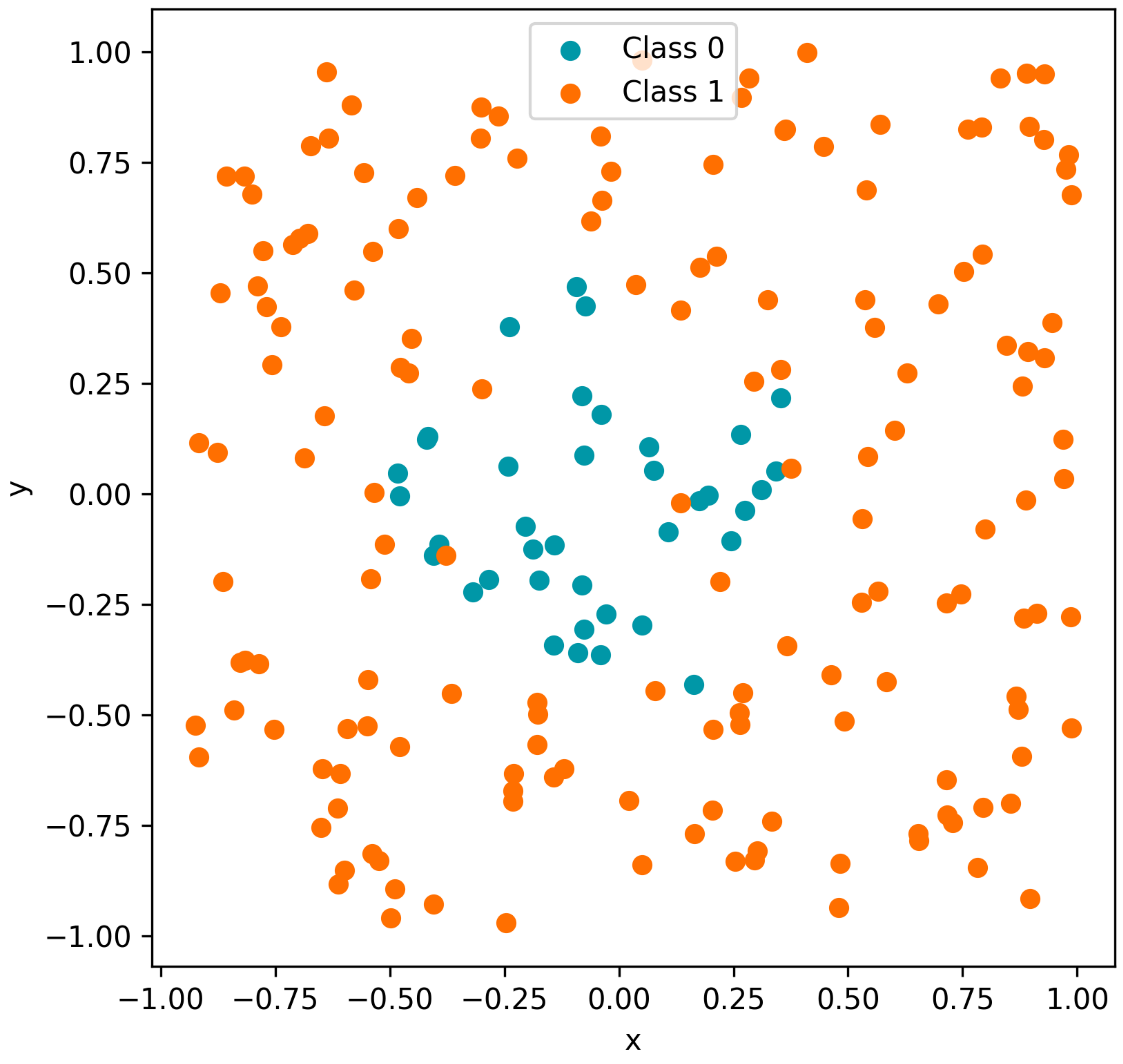
$$\text{n\_samples} \quad \updownarrow \begin{pmatrix} x^2 & x & \text{Intercept} \\ 1.69 & 1.3 & 1 \\ 5.76 & 2.4 & 1 \\ 0.09 & 0.3 & 1 \\ 0.81 & 0.9 & 1 \\ 2.56 & 1.6 & 1 \\ 15.21 & 3.9 & 1 \\ 0.01 & 0.1 & 1 \\ \ldots & \ldots & \end{pmatrix}$$

Example 1: Regressing a complex function

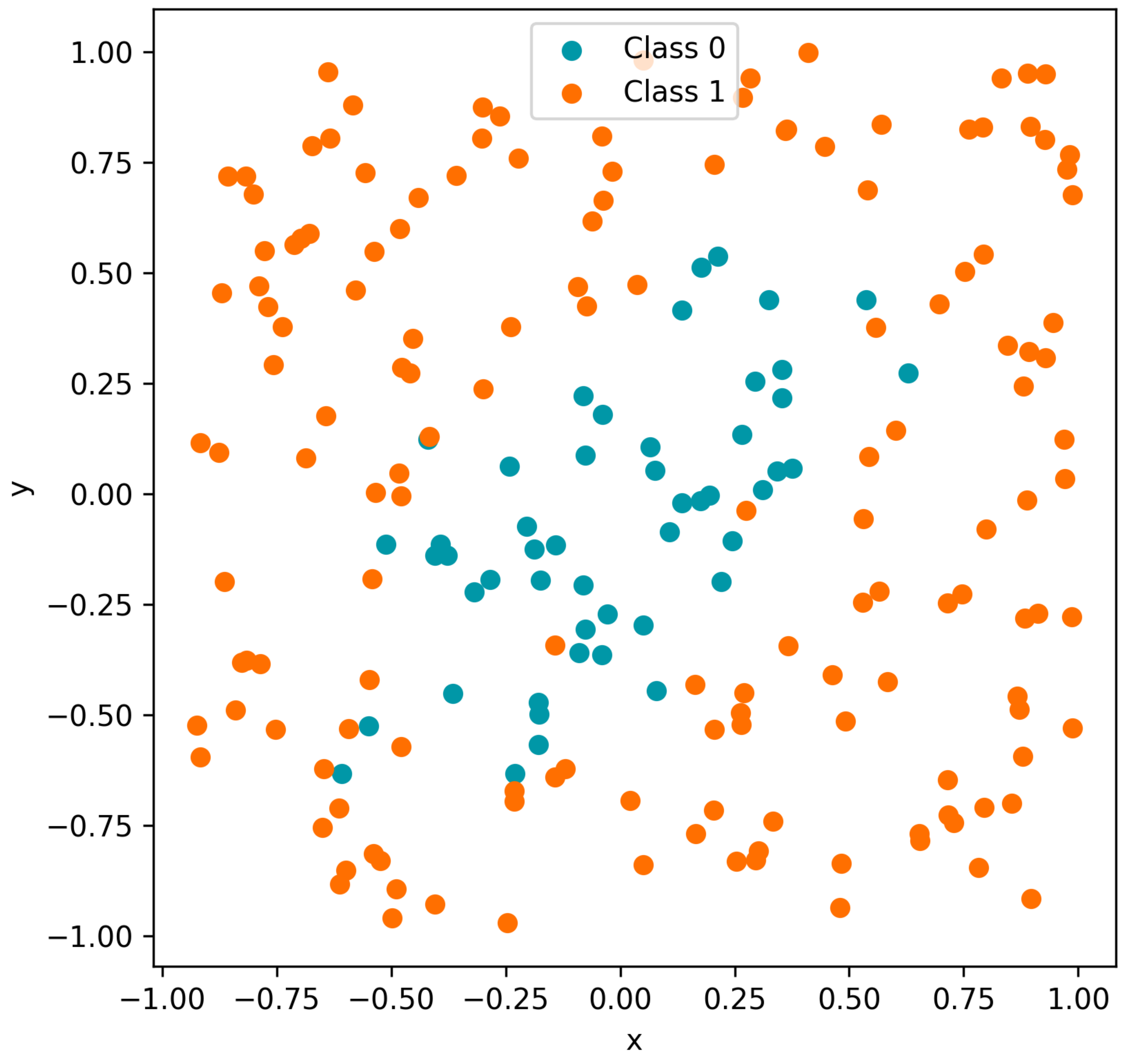


Example 2: Non-linear decision boundaries

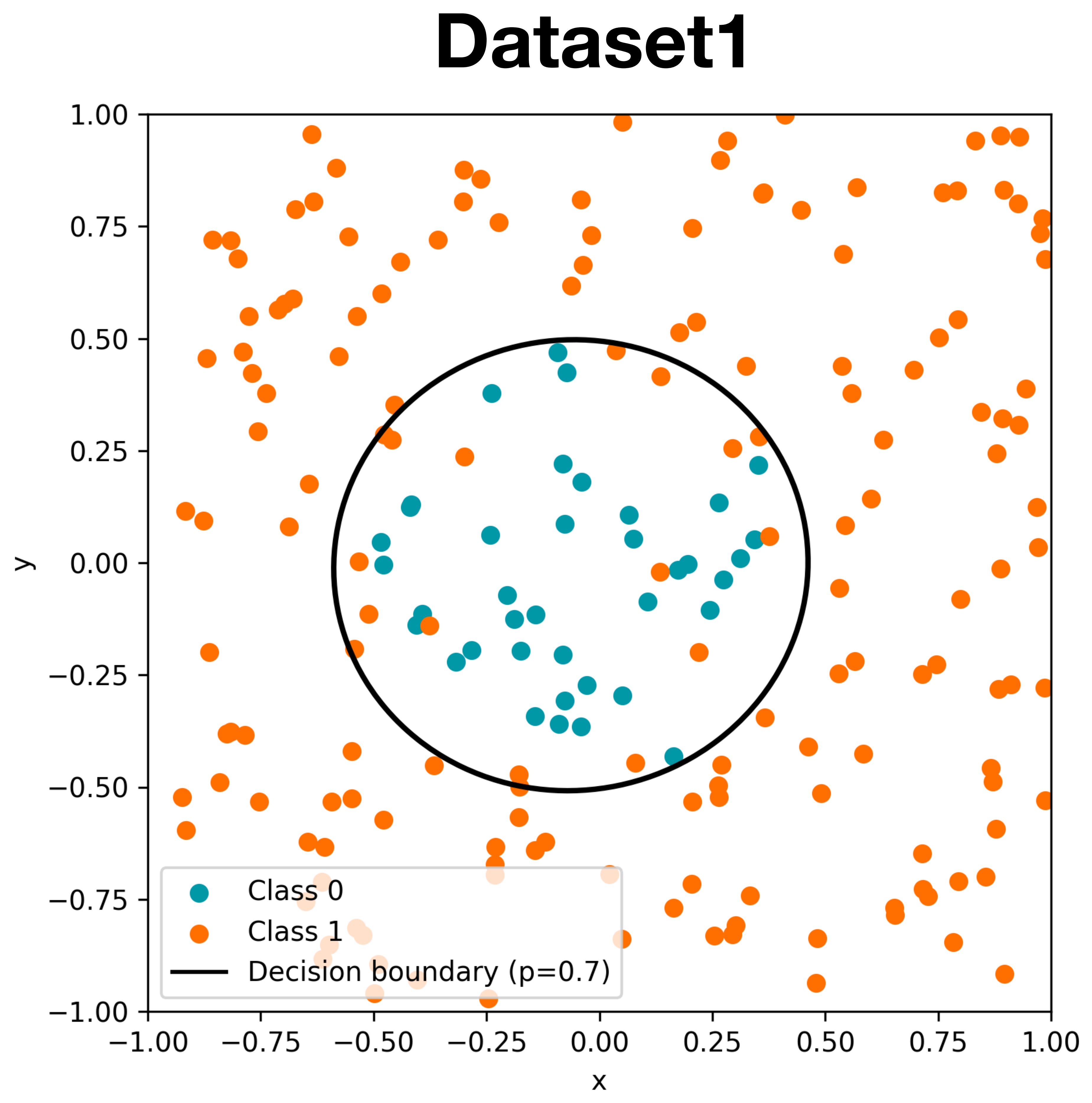
Dataset1



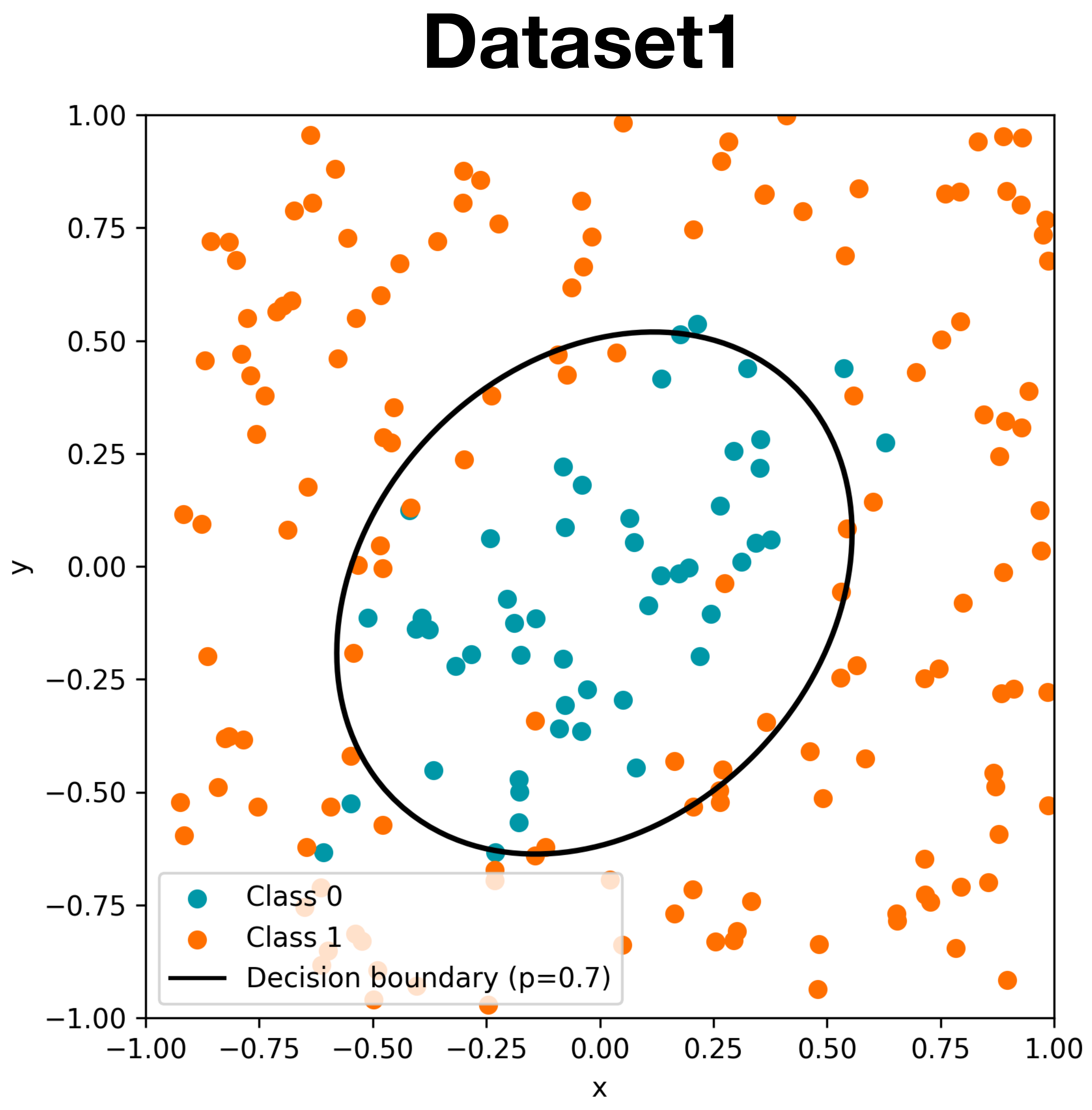
Dataset2



Example 2: Non-linear decision boundaries



$$p = \sigma(\beta_x x^2 + \beta_y y^2)$$



$$p = \sigma(\beta_x x^2 + \beta_y y^2 + \beta_{xy} xy)$$

Some take home messages

- By adding feature transformations, **linear models** can be made as **complex** as you want.
- Adding **too many features** can easily lead to **overfitted models**.
- If we have a model for our data, **combining linear models** with the **correct data transformations** leads to **interpretable** and **effective** models.
- **Data transformations** apply to **features**. On the other hand, **link functions** are transformations applied to the **target**.