Logistic Regression

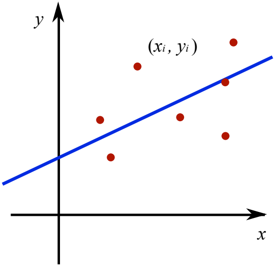
Foundations of Data Analysis

April 5, 2022

Logistic Regression: Estimating the parameters of a logistic model.

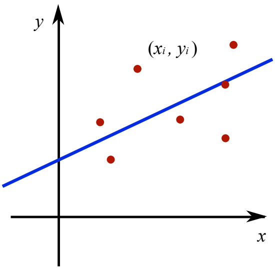
Classification as Regression

Regression problem:
Given *x* (independent variable), predict *y* (dependent variable).

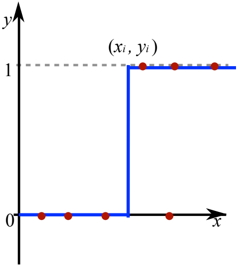


Classification as Regression

Regression problem:
Given x (independent variable),
predict y (dependent variable).

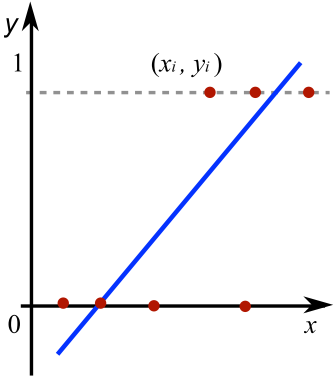


Classification problem:
Given x (features),
predict y (labels).



Classification as Regression

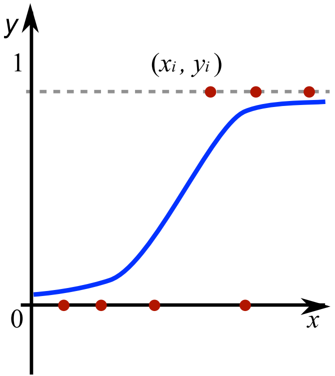
What if given x, we wanted to predict $p(y = 1 \mid x)$?



Linear fit to $p(y = 1 \mid x)$ goes outside $[0,1]$!

Classification as Regression

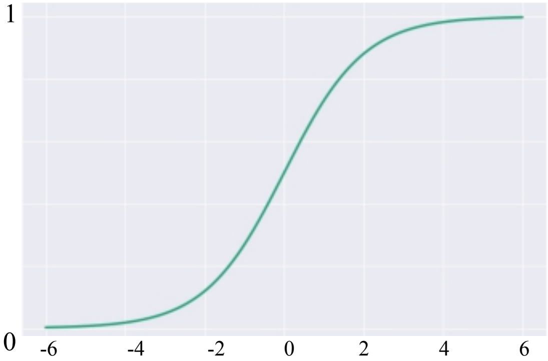
We want to use a nonlinear function with outputs in [0, 1].



This is *logistic regression*.

Logistic Function

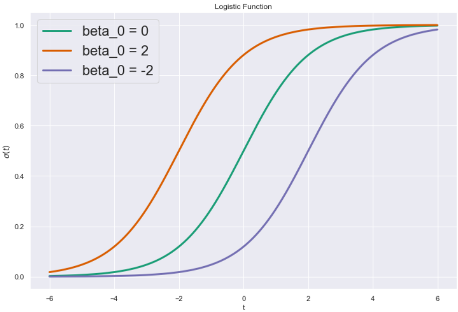
$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

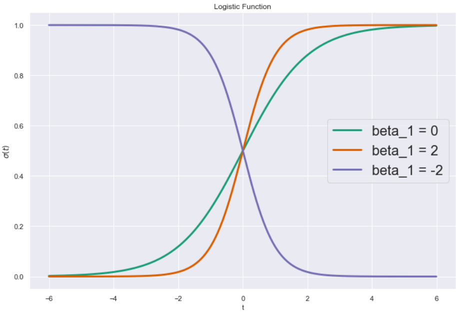


Linear Predictor Inside Logistic Function

a.k.a. Sigmoid function

$$p(y|x) = \sigma(\beta_0 + \beta_1 x) = \frac{1}{1 + e^{-\beta_0 - \beta_1 x}}$$



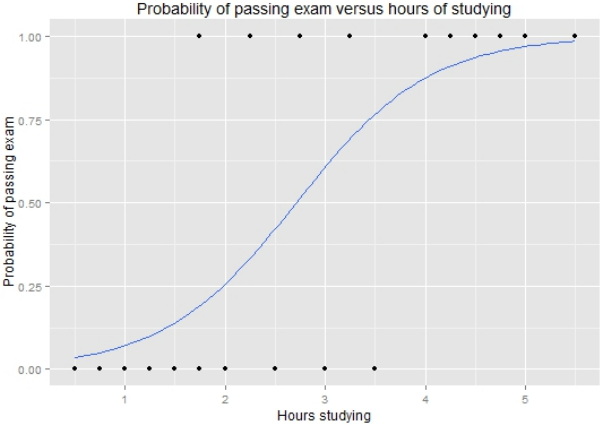


β0 : "intercept"

β1 : "slope"

Example (from Wikipedia)

Pass/fail of exam (y) vs. Hours spent studying (x)



Multivariate Predictor

If x is multivariate: $x = (x^{(1)}, x^{(2)}, \ldots, x^{(d)})$,

$$p(y|x) = \sigma(\beta_0 + \beta_1 x^{(1)} + \beta_2 x^{(2)} + \ldots + \beta_d x^{(d)})$$

$$= \frac{1}{1 + e^{-\beta_0 - \beta_1 x^{(1)} - \beta_2 x^{(2)} - \ldots - \beta_d x^{(d)}}}$$

(Note: just multivariate linear regression inside σ)

Multivariate Predictor

Data matrix X with n data points (rows):

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} \ldots x_{1d} \\ 1 & x_{21} & x_{22} \ldots x_{2d} \\ \cdot & \cdot & \cdot\,\cdot \\ \cdot & \cdot & \cdot\,\cdot \\ 1 & x_{n1} & x_{n2} \ldots x_{nd} \end{pmatrix}$$

Logistic regression evaluated for the i-th data point (i-th row vector):

$$p(y \mid X_{i\bullet}) = \sigma(X_{i\bullet}\beta)$$

(Note: $X_{i\bullet}\beta$ is the dot product between i-th row and β)

How To Estimate Parameter β ?

Maximize likelihood:

1. Compute derivative (gradient) of likelihood w.r.t. β
2. Solve for β that makes this derivative zero

Likelihood Function

Use Bernoulli likelihood:

$$L(\beta; X, y) = \prod \sigma(X_{i\bullet}\beta)^{y_i}(1 - \sigma(X_{i\bullet}\beta)^{1-y_i})$$

Log-Likelihood Function

$$l(\beta; X, y) = \ln L(\beta; X, y)$$

$$= \sum_{i=1}^{n} (y_i - 1)X_{i\bullet}\beta - \ln(1 + e^{-X_{i\bullet}\beta})$$

Gradient of Log-Likelihood Function

$$\nabla \ell(\beta; X, y) = \begin{bmatrix} \frac{\partial \ell}{\partial \beta_0} \\ \frac{\partial \ell}{\partial \beta_1} \\ \cdot \\ \cdot \\ \frac{\partial \ell}{\partial \beta_d} \end{bmatrix}$$

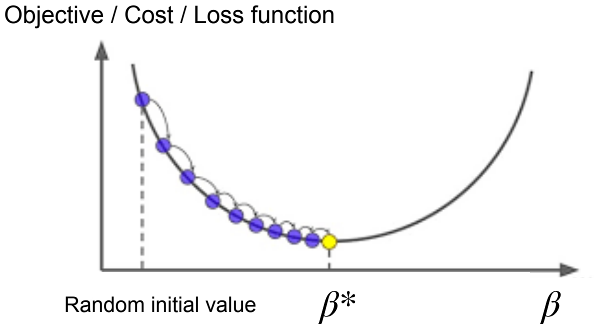
$$\frac{\partial \ell}{\partial \beta_k} = \sum_{i=1}^{n} [(y_i - 1) - \frac{e^{-X_{i\bullet}\beta}}{1 + e^{-X_{i\bullet}\beta}}]$$

Problem! Can't solve for that makes this zero!

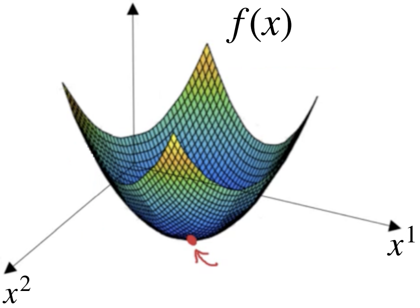
Gradient Ascent / Descent

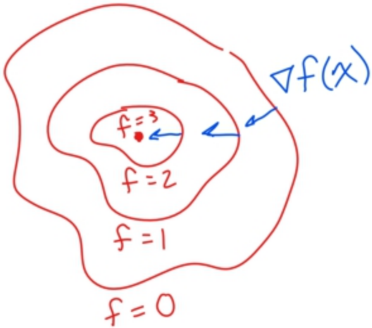
Take a small step (learning rate) in the gradient direction

Repeat until the gradient is zero

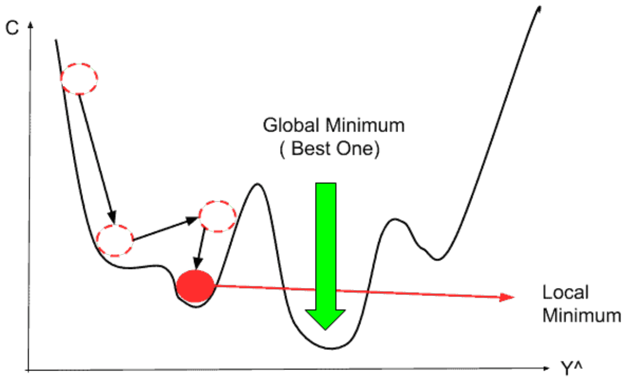


Gradient Ascent / Descent





Optimization of Functions With Local Min/Max



Algorithm for Logistic Regression

Set ϵ = small threshold

Set δ = step size (learning rate) along gradient

Initialize β

While $||\nabla \ell|| > \epsilon$

Update $\beta \leftarrow \beta + \delta \nabla \ell(\beta)$

Effects of Learning Rate