Mathematics for Machine Learning

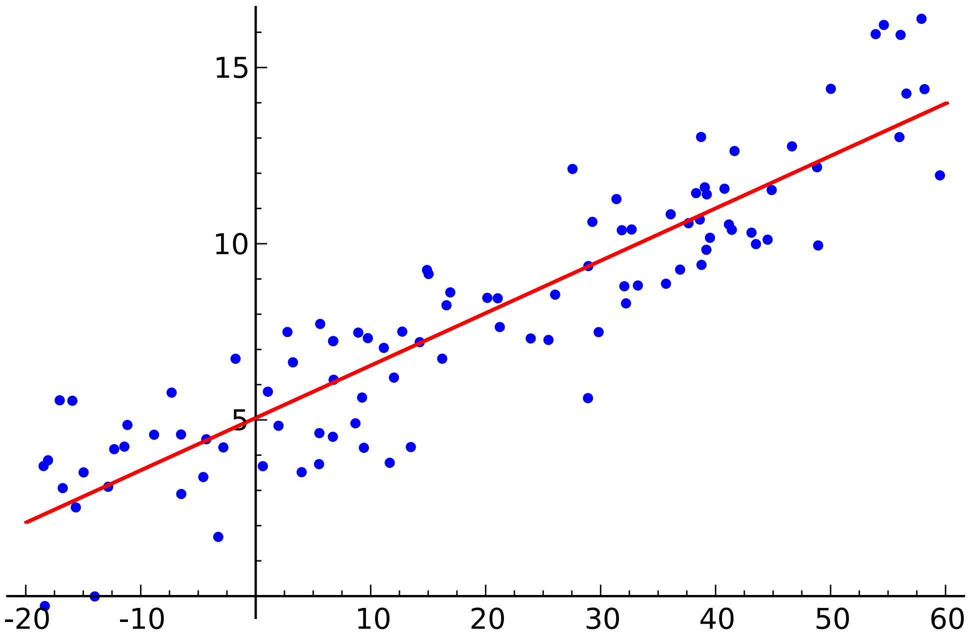
MAD-B3-2526-S2-MAT0611

Fitting Logistic Regression via MLE

Linear Regression Recap

The Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$



where:

- Y: response variable (continuous)
- $X_1, \ldots, X_p$: predictor variables
- $\beta_0, \beta_1, \ldots, \beta_p$: coefficients (parameters)
- $\epsilon \sim N(0, \sigma^2)$: error term

Goal: Estimate coefficients to predict Y from $\mathbf{X}$

Linear Regression: Matrix Notation

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}; \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}; \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

Linear Regression: Ordinary Least Squares (OLS)

Objective

Minimize the sum of squared residuals:

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2$$

In matrix form:

$$\text{RSS} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

Solution (closed-form):

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

We'll look into matrix algebra in future lectures.

Logistic Regression Recap

Binary Classification

$$P(Y=1|\mathbf{X}) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\cdots+\beta_p X_p)}} = \sigma(\beta_0+\beta_1 X_1+\cdots+\beta_p X_p)$$

Logit form:

$$\log\left(\frac{P(Y=1|\mathbf{X})}{1-P(Y=1|\mathbf{X})}\right) = \beta_0+\beta_1 X_1+\cdots+\beta_p X_p$$

Interpretation:

- Linear relationship between predictors and log-odds
- β_j: change in log-odds for one unit increase in X_j

Logistic Regression: Estimation

Maximum Likelihood Estimation

Likelihood for observations $(y_i, \mathbf{x}_i)$:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i}$$

$$\text{where} \quad p_i = P(Y_i = 1|\mathbf{x}_i) = \frac{1}{1 + e^{-\mathbf{x}_i^T \boldsymbol{\beta}}}$$

Log-likelihood:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

Logistic Regression: MLE Solution

No closed-form solution!

Use **iterative optimization**:

- Newton-Raphson method
- Iteratively Reweighted Least Squares (IRLS)
- Gradient descent variants

Algorithm (Newton-Raphson):

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} - \mathbf{H}^{-1}\nabla\ell$$

where:

- $\nabla\ell$: gradient (score vector)
- $\mathbf{H}$: Hessian matrix (second derivatives of log-likelihood)

Logistic Regression: Gradient (Score Vector)

Log-likelihood:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

Gradient w.r.t. β_j:

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell}{\partial p_i} \cdot \frac{\partial p_i}{\partial \beta_j}$$

Key derivatives:

$$\frac{\partial \ell}{\partial p_i} = \frac{y_i}{p_i} - \frac{1 - y_i}{1 - p_i} = \frac{y_i - p_i}{p_i(1 - p_i)}$$

$$\frac{\partial p_i}{\partial \beta_j} = p_i(1 - p_i)x_{ij}$$

Logistic Regression: Gradient (Simplified)

Combining the derivatives:

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - p_i}{p_i(1-p_i)} \cdot p_i(1-p_i)x_{ij} = \sum_{i=1}^{n}(y_i - p_i)x_{ij}$$

Gradient vector (score):

$$\nabla \ell = \mathbf{X}^T(\mathbf{y} - \mathbf{p})$$

where:

- $\mathbf{X}$: design matrix $(n \times (p+1))$
- $\mathbf{y}$: outcome vector $(n \times 1)$
- $\mathbf{p}$: predicted probabilities $(n \times 1)$

Logistic Regression: Hessian Matrix

Second derivative w.r.t. β_j and β_k:

$$\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} = \sum_{i=1}^{n} \frac{\partial}{\partial \beta_k}[(y_i - p_i)x_{ij}]$$

Key observation: y_i doesn't depend on β_k

$$\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} = -\sum_{i=1}^{n} x_{ij} \cdot \frac{\partial p_i}{\partial \beta_k} = -\sum_{i=1}^{n} x_{ij} \cdot p_i(1 - p_i)x_{ik}$$

Simplification:

$$\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} = -\sum_{i=1}^{n} p_i(1 - p_i)x_{ij}x_{ik}$$

Logistic Regression: Hessian (Matrix Form)

Hessian matrix:

$$\mathbf{H} = -\mathbf{X}^T \mathbf{W} \mathbf{X}$$

where $\mathbf{W}$ is a diagonal matrix with weights:

$$\mathbf{W} = \mathrm{diag}(w_1, w_2, \ldots, w_n), \quad w_i = p_i(1 - p_i)$$

Properties:

- **Negative definite**: $\mathbf{H} \preceq 0$ (log-likelihood is concave)
- **Unique maximum**: Guarantees convergence to global optimum
- **Fisher Information**: $\mathcal{I}(\boldsymbol{\beta}) = -\mathbb{E}[\mathbf{H}] = \mathbf{X}^T \mathbf{W} \mathbf{X}$

Newton-Raphson Update

Update formula:

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} - \mathbf{H}^{-1}\nabla\ell$$

Substituting gradient and Hessian:

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y} - \mathbf{p})$$

Algorithm:

1. Initialize $\boldsymbol{\beta}^{(0)}$ (often $\mathbf{0}$)
2. Compute $\mathbf{p}$ using current $\boldsymbol{\beta}$
3. Compute $\mathbf{W} = \mathrm{diag}(p_i(1 - p_i))$
4. Update $\boldsymbol{\beta}$
5. Check convergence: $||\nabla\ell|| < \epsilon$