Logit and Multinomial Logit

Shreeyesh Menon

1 Introduction

Classification problems arise when the response variable takes categorical values and we have to select *labels* based on some observed *features*.

Two common cases are:

1. **Binary classification:** $Y \in \{0, 1\}$
2. **Multi-class classification:** $Y \in \{1, 2, \ldots, K\}$ with $K > 2$

Linear regression is not appropriate for such settings because its predictions are unbounded and do not naturally represent probabilities. Logistic-type models address this issue by mapping linear predictors into probability space using the logit link.

2 Binary Logit Model (Logistic Regression)

2.1 Model Specification

For binary classification, let $Y \in \{0, 1\}$. We model

$$\Pr(Y = 1 \mid X = x) = p(x).$$

Logistic regression assumes that the *log-odds* is linear in x:

$$\log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta^\top x.$$

Solving for $p(x)$ gives the logistic function:

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta^\top x)}}.$$

2.2 Interpretation

- The coefficients β_j represent the change in the log-odds associated with a one-unit increase in feature x_j, holding other variables fixed.
- The logistic function ensures predicted probabilities are between 0 and 1.
- The decision rule is usually $\hat{Y} = 1$ if $\hat{p}(x) > 0.5$.

2.3 Likelihood and Estimation

Given observations (x_i, y_i) for $i = 1, \ldots, n$, the likelihood is:

$$L(\beta) = \prod_{i=1}^{n} p(x_i)^{y_i} [1 - p(x_i)]^{1-y_i},$$

and the log-likelihood is:

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i \log p(x_i) + (1 - y_i) \log(1 - p(x_i)) \right].$$

There is no closed-form solution for maximizing $\ell(\beta)$, so numerical methods such as Newton–Raphson or gradient descent are used.

2.4 Decision Boundary

The decision boundary is obtained by setting $p(x) = 0.5$:

$$\beta_0 + \beta^\top x = 0.$$

Thus logistic regression yields a *linear* decision boundary in the input space.

3 Multinomial Logit Model

Binary logistic regression generalizes naturally to multiple categories using the *softmax* function.

3.1 Model Formulation

Let $Y \in \{1, \ldots, K\}$. For each class k, define a linear predictor:

$$\eta_k(x) = \beta_{k0} + \beta_k^\top x.$$

One class (typically K) is chosen as the reference class with $\eta_K(x) = 0$.

The multinomial logit model specifies:

$$\Pr(Y = k \mid X = x) = \frac{\exp(\eta_k(x))}{\sum_{j=1}^{K} \exp(\eta_j(x))}, \qquad k = 1, \ldots, K.$$

3.2 Interpretation

- Coefficients represent effects on the *log-odds* relative to the reference category:
$$\log \frac{\Pr(Y = k \mid X = x)}{\Pr(Y = K \mid X = x)} = \eta_k(x).$$
- Each class has its own set of parameters.
- The predicted class is the one with highest predicted probability.

3.3 Likelihood

The log-likelihood is:

$$\ell(\beta) = \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{I}(y_i = k) \log \Pr(Y = k \mid X = x_i).$$

Again, the parameters are estimated via numerical optimization.