SNS COLLEGE OF TECHNOLOGY
(AN AUTONOMOUS INSTITUTION)
COIMBATORE – 35
DEPARTMENT OF COMPUTER SIENCE AND ENGINEERING

UNIT II SUPERVISED LEARNING

Introduction - Linear Models for Regression – Linear Regression Models and Least Squares
Subset Selection
Shrinkage Methods – Derived Input Directions
Linear Models for Classification- Discriminant Analysis
Logistic Regression
Separating Hyperplanes

Logistic Regression

Logistic Regression is a widely used statistical method for binary classification problems. It models the relationship between one or more independent variables and a binary dependent variable (e.g., yes/no, success/failure) by estimating probabilities using a logistic function. Here's a comprehensive overview of logistic regression, including its key concepts, methodology, assumptions, evaluation metrics, and applications.

Key Concepts

1. **Binary Outcome**: Logistic regression is primarily used for binary outcomes, where the dependent variable can take one of two values (e.g., 0 or 1).
2. **Logistic Function**: The core of logistic regression is the logistic function, also known as the sigmoid function, which maps any real-valued number into the (0, 1) interval. It's defined as:

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

where z is a linear combination of the input features.

3. **Linear Combination**: The model can be expressed as:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

where β_0 is the intercept, β_i are the coefficients for each feature x_i.

4. **Probability Estimation**: The predicted probability of the positive class (e.g., y=1) is given by:

$$P(y=1|x) = \sigma(z) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n)}}$$

5. **Decision Boundary**: To classify an observation, a threshold (commonly 0.5) is applied to the predicted probabilities:
    - If $P(y=1|x) \geq 0.5$, classify as 1 (positive class).
    - If $P(y=1|x) < 0.5$, classify as 0 (negative class).

Methodology

1. **Data Preparation**: Organize the data into a feature matrix X and a target vector y.
2. **Model Fitting**:
    - Use maximum likelihood estimation (MLE) to estimate the coefficients (β).
    - The likelihood function for logistic regression is:

$$L(\beta) = \prod_{i=1}^{n} P(y_i|x_i)^{y_i} (1 - P(y_i|x_i))^{1-y_i}$$

    - The coefficients are estimated by maximizing this likelihood function.
3. **Optimization**: Techniques like gradient descent or optimization algorithms (e.g., Newton-Raphson) are often used to find the best-fitting parameters.
4. **Model Interpretation**:
    - The coefficients β_i represent the change in the log-odds of the outcome for a one-unit change in the predictor variable x_i.
    - The odds ratio can be computed as $e^{\beta i}$, which indicates how the odds of the event change with a one-unit increase in the predictor.

Assumptions

1. **Binary Dependent Variable**: The dependent variable should be binary.
2. **Independence of Observations**: The observations must be independent of each other.
3. **Linearity of the Logit**: The relationship between the log-odds of the dependent variable and the independent variables should be linear.
4. **No Multicollinearity**: Independent variables should not be too highly correlated with each other.

Evaluation Metrics

1. **Accuracy**: The proportion of correctly classified instances.

2. **Confusion Matrix**: A table that summarizes the performance of the classification model, showing true positives, true negatives, false positives, and false negatives.
3. **Precision**: The ratio of true positives to the sum of true positives and false positives.
4. **Recall (Sensitivity)**: The ratio of true positives to the sum of true positives and false negatives.
5. **F1 Score**: The harmonic mean of precision and recall, useful for imbalanced datasets.
6. **ROC Curve and AUC**: The Receiver Operating Characteristic curve shows the trade-off between true positive rate and false positive rate at various thresholds, with the Area Under the Curve (AUC) quantifying the overall model performance.

Applications

1. **Medical Diagnosis**: Predicting whether a patient has a disease based on test results and patient characteristics.
2. **Credit Scoring**: Assessing the likelihood of a borrower defaulting on a loan.
3. **Marketing**: Classifying customers as likely or unlikely to respond to a marketing campaign.
4. **Spam Detection**: Identifying whether an email is spam or not based on its content.

Advantages and Limitations

Advantages:

- **Simplicity**: Easy to understand and implement.
- **Interpretability**: Coefficients provide insight into the relationship between predictors and the outcome.
- **Probabilistic Outputs**: Provides probabilities that can be useful for decision-making.

Limitations:

- **Linearity Assumption**: May not perform well if the relationship between the features and the log-odds is not linear.
- **Sensitivity to Outliers**: Can be influenced by extreme values in the dataset.
- **Imbalanced Classes**: May struggle if one class significantly outnumbers the other.

Conclusion

Logistic Regression is a foundational technique in statistical modeling and machine learning, particularly useful for binary classification tasks. Its interpretability, combined with the ability to provide probabilities, makes it a popular choice across various fields.

Extra

Logistic Regression

The objective of logistic regression is to model the **posterior probabilities of K classes via linear functions in x, while at the same time ensuring that they sum to one and remain in [0,1].**

$$log\frac{Pr(G=1|X=x)}{Pr(G=K|X=x)} = \beta_{0_1} + \beta_1^T x$$

$$log\frac{Pr(G=2|X=x)}{Pr(G=K|X=x)} = \beta_{0_2} + \beta_2^T x$$

$$\cdots$$

$$log\frac{Pr(G=k-1|X=x)}{Pr(G=K|X=x)} = \beta_{0_{(k-1)}} + \beta_{k-1}^T x$$

Logistic regression, self-generated.

Here K-1 is the number of logit transformations(log-odds). The last class is used as the denominator in the log-odds ratio. We can obtain ***Pr(G=k|X=x)*** using exponents

$$Pr(G=K|X=x) = \frac{exp(\beta_{k0} + \beta_k^T x)}{1 + \sum_{l=1}^{K-1} exp(\beta_{0_l} + \beta_l^T x)}, K = 1, \ldots, K-1$$

$$Pr(G=K|X=x) = \frac{1}{1 + \sum_{l=1}^{K-1} exp(\beta_{0_l} + \beta_l^T x)}$$

Obtaining ***Pr(G=k|X=x),*** *self-generated.*

To simplify the notation we will use Θ as the entire parameter set Θ = ***{β01,β1^T,…,β0(k-1),β(k-1)T}*** and thus denote the probabilities ***Pr(G=k|X=x)=Pk(x:Θ).***

Fitting logistic regression models

Logistic models are usually fitted by maximum likelihood, using the conditional likelihood of G given X. Since P(G|X) specifies the conditional distribution, the multinomial distribution is appropriate. The log-likelihood for N observations is:

$$l(\theta) = \sum_{i=1}^{N} log P_{g_i}(x_i; \theta)$$

Log-likelihood for N observations, self-generated.

where ***Pk(xi:Θ)=Pr(G=k|X=xi,Θ).***

In the two-class case, the algorithm is simple, it is convenient to code the two classes ***gi = {1,2}***via a ***0,1*** response ***yi*** where:

- ***yi =1*** if ***gi =1***
- ***yi = 0*** if ***gi = 2***

Let ***p1(x:Θ)=p(x:Θ)*** and ***p2(x:Θ)=1-p(x:Θ).*** The log-likelihood can be written as

$$l(\beta) = \sum_{i=1}^{N} \{y_i log\, p(x_i : \beta) + (1-y_i) log\, (1-p(x_i : \beta))\} = \sum_{i=1}^{N} \{y_i \beta^T x_i - log\, (1+e^{\beta^T x_i})\}$$

Log-likelihood for N observations, self-generated.

Here ***β = {β_01,β_1},*** and we assume that the vector of inputs ***xi*** includes the constant term **1** to accommodate the intercept. To maximize the log-likelihood we set the derivates to **0**.

$$\frac{dl(\beta)}{d\beta} = -\sum_{i=1}^{N} x_i (y_i - p(x_i : \beta)) = 0$$

The first derivative, self-generated.

These are p+1 equations non-linear in ***β.*** To solve their equations, we can use the Newton-Raphson algorithm, which needs the second derivative or the hessian matrix.

$$\frac{d^2l(\beta)}{d\beta d\beta^T} = -\sum_{i=1}^{N} x_i x_i^T p(x_i : \beta)(1 - p(x_i : \beta)))$$

The second derivative, self-generated.

starting with ***βold***, a single Newton update is:

$$\beta^{new} = \beta^{old} - (\frac{d^2l(\beta)}{d\beta d\beta^T})^{-1}\frac{dl(\beta)}{d\beta}$$

βold *Newton update, self-generated.*

where the derivatives are evaluated at ***βold.*** In matrix notation, we can write it as:

$$\frac{dl(\beta)}{d\beta} = X^T(Y - P) \vdots \frac{d^2l(\beta)}{d\beta d\beta^T} = -X^T W X$$

Derivatives in matrix notation, self-generated

Then, the Newton step looks like this:

$$\beta^{new} = \beta^{old} + (X^T W X)^{-1} X^T (y - p) = (X^T W X)^{-1} X^T (X\beta^{old} + W^{-1}(y - p)) = (X^T W X)^{-1} X^T W_z$$

$$where\ z = X\beta^{old} + W^{-1}(y - p)$$

Newton steps in matrix notation, self-generated.

z is known as the adjusted response. This model is an iterative reweighted least-squares(IRLS) since at each iteration it solves the weighted least squares problem:

$$\beta^{new} \leftarrow argmin_x (z - X\beta)^T W (Z - X\beta)$$

IRLS maximization problem, self-generated.

Normally ***β=0*** is a good starting point for the iterative process.

Quadratic approximation and inference

The maximum-likelihood parameter estimates ***β^*** satisfying a self-consistency relationship, they are coefficients of a weighted least-squares fit, where the response is:

$$z_i = X_i^T \hat{\beta} + \frac{y_i - \hat{p}_i}{\hat{p}_i(1 - \hat{p}_i)}$$

The weighted least-squares response, self-generated.

the weights are ***wi = pi^(1-pi^)***, both depending on ***β^*** itself. Tho connection with least-squares offers a lot to us:

- The weighted residual sum of squares is the familiar Pearson Chi-Squared statistic.
- Asymptotic likelihood theory says that if the model is correct, ***β*** is consistent, so it converges to the true ***β.***
- The Central Limit Theorem shows that the distribution of ***β^*** converges to ***N(β,(X^TWX)^(-1)).***

As model building for logistic regression models can be costly due to the required iteration, there exist some popular shortcuts to avoid the iteration, some of them are Rao score test to Wald test, they are based on the maximum-likelihood fit of the current model.

L1 regularized logistic regression

The L1 penalty can be used in any linear regression model. For logistic regression, we can add it as:

$$max_{\beta_0,\beta}\{\sum_{i=1}^{N}[y_i(\beta_0 + \beta^T x_i) - log(1 + e^{\beta_0+\beta^T x_i})] - \lambda \sum_{j=1}^{p} |\beta_j|\}$$

L1-regularization for logistic regression, self-generated.

We can solve it with IRLS using the quadratic approximation. The score equation for non-zero coefficient variables have the form:

$$x_j^T(y - p) = \lambda \; sign(\beta_j)$$

Non-zero coefficient variables, self-generated.

which generalizes the LAR. Path algorithms as LAR for lasso are more difficult because the coefficients are pice-wise smooth rather than linear. Nevertheless, progress can be estimated with quadratic approximations.

Logistic regression vs linear discriminant analysis

Both models seem to be the same, they have exactly the same form, but the **logistic regression is more general in that It makes fewer assumptions.**

The logistic regression takes the density function *Pr(x)* and fits the parameters of *P(G|X)* by maximizing the conditional likelihood. Meanwhile, LDA first the parameters by maximizing the full log-likelihood, based on the prior density.

By relying on more assumptions, logistic regression has more information about the parameters and can obtain lower variances on estimation.

Observations far from the decision boundaries are important for the covariance matrix estimation, which makes LDA not robust to gross outliers.

If the data in a plane can be separated by a line, the logistic regression will not converge to the best solution, otherwise, LDA will do it.

Conclusion

Logistic regression and linear discriminant analysis are both good approaches to perform a simple classification. Logistic regression is a more robust model because of the lower amount of assumptions, but in practice, both models perform similarly.