Lecture Note 10: Logistic Regression and Maximum Likelihood Estimation

1. Generalized Linear Models

The Generalized Linear Model (GLM) is a flexible extension of the traditional multiple linear regression model, designed to accommodate response variables (dependent variables) that do not follow a normal distribution or where the relationship between the response and predictors is not linear. Unlike multiple linear regression, which assumes that the response variable is continuous and normally distributed, GLMs allow for response variables that can be binary, count data, proportions, or positive continuous variables, among others. This is achieved by introducing a link function that connects the linear predictor (a linear combination of the independent variables) to the mean of the response variable's distribution.

$$y_i = f(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k) + \varepsilon_i$$

GLMs consist of three components:

1. **The Random Component** ε_i: Specifies the probability distribution of the response variable (e.g., normal, binomial, Poisson), extending the model's applicability to various types of data.
2. **The Systematic Component** $\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$: Similar to multiple linear regression, it is a linear combination of the independent variables.
3. **The Link Function** f: A crucial element that connects the systematic component to the mean of the random component's distribution. The choice of link function depends on the nature of the response variable and ensures that the model's predictions stay within a range that is plausible for the response variable's distribution (e.g., a logistic function for a binary outcome).

By incorporating these components, GLMs can model a wide range of data types and relationships, making them highly versatile in statistical modeling. They are used extensively across various fields, including medicine, finance, and environmental science, for tasks such as binary classification (logistic regression), count data analysis (Poisson regression), and beyond, providing a powerful framework for understanding complex relationships in data.

2. Logistic Regression

Logistic regression is a statistical model, a special case of generalized linear model, used in machine learning for binary classification tasks, where the goal is to predict outcomes that take on one of two possible states, often labeled as "0" or "1", "yes" or "no", "true" or "false", etc. It's a go-to method for problems with a binary dependent variable, and it can be extended to handle multi-class classification problems as well.

In the realm of Generalized Linear Models (GLM), when we employ the sigmoid or logistic function as the link function, the GLM is transformed into a logistic regression model. This sigmoid function is pivotal as it takes the linear predictor—a real number derived from input

variables weighted by coefficients—and converts it into a value between 0 and 1. This conversion process, resulting in an S-shaped curve known as the logistic curve, allows the output to be interpreted as a probability, indicating the likelihood of the dependent variable being in one of two binary categories, such as '1' or '0'. This interpretation is particularly useful in binary classification problems, where understanding the probability of outcomes and the impact of predictor variables is essential.

The estimated linear predictor or the system component is:

$$\hat{y}_{LR} = b_0 + b_1 X_1 + \cdots + b_k X_k.$$

The sigmoid or logistic function is:

$$\sigma(x) = \frac{e^x}{1+e^x} = \frac{exp(x)}{1+exp(x)}.$$

When the estimated linear predictor, derived from combining input variables and coefficients, is fed through the sigmoid function, we obtain the logistic regression function. This function produces the predicted probability of the dependent variable taking on the value of 1:

$$\hat{p} = \frac{exp(b_0+b_1X_1+\cdots+b_kX_k)}{1+exp(b_0+b_1X_1+\cdots+b_kX_k)}.$$

Through simple algebraic manipulation, we find that taking the exponential of the linear predictor yields the odds, which is the ratio of the estimated probability of an event occurring to the probability of it not occurring:

$$exp(b_0 + b_1 X_1 + \cdots + b_k X_k) = \frac{\hat{p}}{1-\hat{p}}.$$

Additionally, by applying the logarithm to the odds, we arrive at the estimated linear predictor, commonly known as the log-odds or logit, which forms the systemic component of the model.

$$\hat{y}_{LR} = b_0 + b_1 X_1 + \cdots + b_k X_k = \ln \frac{\hat{p}}{1-\hat{p}}.$$

Logistic regression represents a specialized case within the Generalized Linear Model (GLM) framework, where a linear regression model is crafted to estimate the log-odds of the dependent variable being '1'. Essentially, this means that the natural logarithm of the odds ratio (the probability of the event occurring versus it not occurring) is expressed as a linear combination of the predictor variables $X_1, \cdots, X_k$. In logistic regression, we are not directly estimating the probability of an event but rather the logarithm of the odds that the event occurs, which is then modeled to be directly proportional to the independent variables. This relationship allows for the easy interpretation of the effect of each predictor on the odds of the event in question, expanding the applicability of linear regression principles to binary outcome data.

2.1 Threshold for Making Predictions

In logistic regression, the output is the probability of an observation belonging to the positive class (usually denoted as '1'). To classify an observation as '0' or '1', we set a threshold

probability. By default, this threshold is often set at 0.5: if the model's predicted probability for an observation is greater than or equal to 0.5, the observation is predicted to be in the positive class ('1'); if it's less than 0.5, it's predicted to be in the negative class ('0').

Adjusting the threshold probability has significant implications for the model's performance, particularly concerning its precision (the proportion of true positive results in all positive predictions) and recall (the proportion of true positive results in all actual positives). Setting a higher threshold (closer to 1) makes the model more conservative in predicting positive outcomes, which can increase precision but decrease recall. This means the model requires stronger evidence to classify an observation as positive, potentially missing some true positives but reducing false positives. Conversely, a lower threshold increases recall but may reduce precision, as the model classifies more observations as positive, catching more true positives but also increasing false positives.

The choice of threshold depends on the specific needs and priorities of the task at hand. For example, in medical testing, a high recall might be prioritized to ensure all positive cases are identified, even at the cost of higher false positives. In contrast, in spam detection, high precision might be more desirable to minimize the risk of classifying important emails as spam. Ultimately, adjusting the threshold allows for fine-tuning of the model's sensitivity and specificity to align with the objectives of the classification task, highlighting the importance of considering the balance between different types of errors and the overall context in which the model is applied.

3. Maximum Likelihood Estimation

In logistic regression, finding the optimal coefficients for the linear predictor presents a challenge, as the least squares method used in linear regression is not suitable for a binary outcome. This is because, unlike linear regression, which predicts a continuous value, logistic regression predicts a probability that the target variable is either '1' or '0'.

To reconcile this, we adapt our approach: rather than fitting a line to our data in the least squares sense, we fit a probability curve to the occurrence of the '1' outcome. This is done using a method called maximum likelihood estimation, which seeks to maximize the likelihood that the parameters of the logistic model result in the observed outcomes.

We can consider the predicted probability from the logistic regression model as a score. If our target variable is indeed '1', we score the model's predicted probability, $\hat{p}$, of being '1'. If the actual target is '0', then we score the model's predicted probability of being '0', which is $1 - \hat{p}$. To represent the model's prediction performance on an individual data point, we can use a bit of algebraic expression that combines these two scenarios:

$$\hat{p}^{y}(1-\hat{p})^{1-y}.$$

When $y = 1$, the expression evaluates to $\hat{p}$, and when $y = 0$, it simplifies to $1 - \hat{p}$. These outcomes exactly represent the model's accuracy in predicting the correct outcome, effectively functioning as the score the predictive model earns for its forecasts.

To accurately model the probability of making correct predictions across all observations, we assume that each instance in the dataset is independent. Under this assumption, the overall probability, or likelihood, of the model correctly predicting every data point is the product of individual probabilities for each observation. This likelihood acts as the objective function in our quest to determine the optimal coefficients for the model's linear component, leading us to employ a method known as maximum likelihood estimation (MLE) for optimization.

$$likelihood = \prod_{i=1}^{n} \hat{p}_i^{\,y_i}(1-\hat{p}_i)^{1-y_i}.$$

Directly working with the product of probabilities, however, poses significant computational challenges. A mathematical solution involves taking the logarithm of the likelihood, resulting in the log-likelihood, which is substantially easier to manage. Given that the logarithm function is monotonically increasing, maximizing the likelihood and its logarithmic counterpart will lead to the same optimal coefficients.

$$log - likelihood = \sum_{i=1}^{n}(y_i \ln \hat{p}_i + (1-y_i)\ln(1-\hat{p}_i))$$

While theoretically, the derivatives of the log-likelihood with respect to the model parameters can be set to zero to find an analytical solution, in practice, finding these optimal values is typically achieved using gradient-based optimization techniques. This practical approach is highly recommended for those interested in exploring the intricacies of logistic regression further.

3.1 MLE for Multiple Linear Regression

Maximum Likelihood Estimation (MLE) is a powerful statistical technique that can be applied to a wide range of models, including linear regression, to estimate the model parameters that best explain the observed data. Conceptually, applying MLE to linear regression models involves a specific assumption about the nature of the error terms in the model.

In the context of linear regression, the model predicts the dependent variable Y as a linear combination of one or more independent variables X, plus an error term ϵ. The model can be expressed as:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon.$$

To apply MLE to this setup, we assume that the error terms ϵ follow an independent, identically distributed (i.i.d.) normal distribution with a mean of zero and a constant variance σ^2. This assumption is key because it implies that the dependent variable Y, given the independent variables X, also follows a normal distribution centered around the linear prediction

$$\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon.$$

Under these conditions, MLE seeks to find the values of the coefficients β_0, β_1, ..., β_p and the variance σ^2 that maximize the likelihood of observing the actual data. Conceptually, this means finding the set of parameters that make the observed data most probable under the specified model.

The process involves:

1. **Formulating the Likelihood Function**: This function represents the probability of observing the data given specific values of the model parameters. Since the errors are assumed to be normally distributed, the likelihood function is derived from the probability density function of the normal distribution.
2. **Maximizing the Likelihood**: The values of β_0, β_1, ..., β_p and the variance σ^2 that maximize this likelihood function are considered the best estimates for the model parameters. This often involves taking the logarithm of the likelihood function to obtain the log-likelihood, which is easier to work with mathematically. Maximizing the log-likelihood through differentiation yields the MLE estimates of the parameters.
3. **Interpreting the Results**: The MLE estimates of the coefficients β_0, β_1, ..., β_p tell us about the relationship between the independent variables and the dependent variable, while the estimate of σ^2 provides information about the variability of the observations around the regression line.

In summary, applying MLE to linear regression under the assumption of normally distributed errors leverages the probabilistic nature of the model to estimate parameters that make the observed data as likely as possible, offering a conceptually intuitive and statistically rigorous method for parameter estimation in linear regression models.

3.2 Discussion on MLE

Maximum Likelihood Estimation (MLE) is a statistical method used for estimating the parameters of a model, characterized by its approach of maximizing the probability of the observed data (likelihood) under the given model. This technique is grounded in the idea that the chosen parameters should make the observed data as likely as possible, as if the model itself generated the data. From a generative perspective, MLE is not just about fitting a model to the data; it's about identifying the parameters that would most plausibly bring the observed data into existence from a specific population distribution.

The concept of MLE as a generative process is particularly powerful and has broad applications across many fields of science and engineering. When we apply MLE, we are essentially asking: given a set of observations, what are the parameter values of our model that would most likely produce these observations if we were to simulate or generate the data from the model? This perspective shifts the focus from mere prediction to understanding the underlying process that generates the data, enhancing the interpretability and applicability of the model.

For example, in the context of a linear regression model, applying MLE involves assuming a particular form of distribution for the residuals (e.g., a normal distribution), then finding the parameter values that maximize the likelihood of observing the given data points. This is equivalent to asking: if we were to generate data points using our linear model with a certain set

of parameters and the assumed distribution of residuals, which parameters would make the actual observed data most probable?

This generative viewpoint is immensely powerful for several reasons:

- **Flexibility**: It allows for the modeling of complex phenomena by assuming different distributions for the data, making MLE adaptable to a wide range of scenarios.
- **Predictive Power**: By capturing the underlying data generation process, models estimated using MLE can offer strong predictive capabilities, even in situations not represented in the training data.
- **Interpretability**: Understanding the process that generates the data provides insights into the mechanisms at play, facilitating more informed decision-making and hypothesis testing.

Moreover, the generative idea underpinning MLE is fundamental to many modern machine learning approaches, including Bayesian methods and generative models like Generative Adversarial Networks (GANs). These applications extend the concept of maximizing likelihood to more complex and nuanced models, further illustrating the versatility and enduring relevance of the MLE approach in both traditional statistics and contemporary machine learning.

4. Discussion on Logistic Regression

Logistic regression stands as a cornerstone statistical method for binary classification problems, offering a framework to estimate the probability of a binary outcome based on one or more predictor variables. Its utility is particularly noted in its provision of odds ratios, which shed light on the influence of each feature on the outcome, thereby offering a nuanced understanding of the data's dynamics.

One of the model's strengths lies in its flexibility to accommodate both linear and nonlinear relationships. This is achieved through the application of transformations on the predictor variables or the incorporation of interaction terms within the model, allowing for a broader capture of the underlying patterns in the data. Furthermore, logistic regression demonstrates resilience against minor noise in the output variable, making it a robust choice for real-world data that often comes with some degree of imperfection.

However, the model operates under several assumptions that, if violated, may impact its performance and the validity of its conclusions. A fundamental assumption is the linear relationship between the log-odds (logit) of the outcome and each predictor. This linearity assumption is pivotal for the model's interpretability and calculation, yet it restricts the model's ability to capture more complex, nonlinear relationships directly without transformations or interactions.

Moreover, logistic regression may struggle with datasets where the relationship between the predictors and the outcome does not conform well to the log-odds scale. Complex relationships that deviate significantly from linearity can lead to model misfit unless properly addressed through the aforementioned methods.

Another critical aspect to consider is the model's vulnerability to overfitting when dealing with a large number of features relative to the number of observations. Overfitting occurs when the model becomes too complex, capturing noise rather than the underlying pattern, which diminishes its predictive power on new data. Conversely, underfitting can arise in situations where the relationships between variables are inherently nonlinear, and the logistic regression model, without adequate adjustments, fails to capture the essence of these relationships.

Despite these limitations, logistic regression's simplicity, interpretability, and efficiency maintain its status as a go-to method for binary classification tasks. Its ability to provide meaningful insights into the importance of predictors and their relationships with the outcome, coupled with the availability of techniques to address its limitations, such as feature selection and engineering, ensures its continued relevance in the analytical domain.

In summary, while logistic regression is a powerful tool for classification problems, a deep understanding of its assumptions and limitations is essential for its effective application. Recognizing when and how to adjust the model to accommodate the peculiarities of your data can greatly enhance the accuracy and interpretability of your results, making logistic regression a versatile and invaluable method in the statistical analysis toolkit.

5. Example (forthcoming)