logistic regression

Referred to as Binomial regression in the two class problem.

Goal: Model the probability of being in each class given its predictors by estimating the following functions:

\[
P(Y = 1|X = x) = \frac{e^{\beta^T x}}{1 + e^{\beta^T x}}
\]

\[
P(Y = 0|X = x) = 1 - \frac{e^{\beta^T x}}{1 + e^{\beta^T x}} = \frac{1}{1 + e^{\beta^T x}}
\]
Given \( n \) data points \( \{ x_i \}_{i=1}^n \) drawn independently from \( p(x; \theta) \), where the form of \( p(x) \) is known but \( \theta \) is unknown, then \( \hat{\theta}_{\text{MLE}} \) is the Maximum Likelihood Estimate which maximizes the Likelihood of the data.

\[
\hat{\theta}_{\text{MLE}} = \operatorname{argmax}_\theta l(\theta)
\]

In this case, we wish to find \( \hat{\beta} \) which maximizes \( \ell(\beta) \) where

\[
\ell(\beta) = \log(L(\beta)) = \sum_{i=1}^{n} \log(f(x_i; \beta))
\]

\[
f(x_i; \beta) = \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}} y_i \left(\frac{1}{1+e^{\beta^T x_i}}\right)^{1-y_i}
\]
logistic regression

In order to find $\hat{\beta}$ which maximizes $\ell(\beta)$, we set $\frac{\partial \ell}{\partial \beta} = 0$ and solve $\beta$.

$$
\ell(\beta) = \sum_{i=1}^{n} \log f(x_i; \beta)
$$

$$
= \sum_{i=1}^{n} y_i \log \left( \frac{e^{\beta^T x_i}}{1 + e^{\beta^T x_i}} \right) + (1 - y_i) \log \left( \frac{1}{1 + e^{\beta^T x_i}} \right)
$$

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

$$
= \sum_{i=1}^{n} y_i \beta^T x_i - \log (1 + e^{\beta^T x_i})
$$
logistic regression

\[
\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} y_i x_i - \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}} x_i
\]

We see that \( \hat{\beta} \) cannot be found analytically so we can use a numerical method; the Newton Raphson algorithm is widely used:

1) initialize \( x_0 \)

2) \( x_{k+1} = x_k - f''(x_k)^{-1}f'(x_k) \)

3) repeat until convergence (ie. \( |x_{k+1} - x_k| < \epsilon \))
logistic regression

For convenience, let $p_i = \frac{e^{\beta^T x_i}}{1 + e^{\beta^T x_i}}$ and $1 - p_i = \frac{1}{1 + e^{\beta^T x_i}}$.

Compute the first derivative (Score vector)

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} (y_i - p_i) x_i$$

Compute the second derivative (Hessian matrix)

$$\frac{\partial^2 \ell}{\partial \beta \partial \beta^T} = -\sum_{i=1}^{n} x_i p_i (1 - p_i) x_i^T$$

Now we can apply the Newton Raphson algorithm to maximize $\ell(\beta)$

$$\beta^{t+1} \leftarrow \beta^t - \left( \frac{\partial^2 \ell}{\partial \beta^t \partial \beta^T} \right)^{-1} \frac{\partial \ell}{\partial \beta^t}$$
Recalling some matrix algebra, We can convert all summations to matrix operations.

$$\frac{\partial \ell}{\partial \beta} = X^T (y - p)$$

$$\frac{\partial^2 \ell}{\partial \beta \partial \beta^T} = -X^TWX; \ W_{ii} = p_i(1 - p_i), \ W_{ij} = 0$$

The Newton Raphson algorithm can now be expressed as

$$\beta^{t+1} \leftarrow \beta^t + (X^TWX)^{-1}X^T(y - p)$$
logistic regression

Alternatively, the algorithm can be expressed as:

\[
\beta^{t+1} \leftarrow \beta^t + (X^T WX)^{-1} X^T (y - p) \\
\leftarrow (X^T WX)^{-1} \left[ X^T WX \beta^t + X^T (y - p) \right] \\
\leftarrow (X^T WX)^{-1} X^T WZ
\]

where \( Z = X \beta^t + W^{-1}(y - p) \)

This algorithm is also known as Iteratively Re-weighted Least Squares (IRLS)

\[
\beta^{new} \leftarrow \text{argmin}_\beta (Z - X \beta)^T W (Z - X \beta)
\]
**Note:** For a $d$-dimensional $\mathbf{x}$ this model has $d$ adjustable parameters. By contrast to LDA we have: $2d$ parameters for the means and $d(d + 1)/2$ parameters for the covariance matrix. Together with the class priors, LDA gives a total of $d(d + 5)/2 + 1$ parameters which grows quadratically in $d$, in contrast to the linear growth of parameters ($d$ parameters) of logistic regression. For large $d$, there is a clear advantage for working with the logistic regression model directly.