# BIOS 825: Linear Classification Methods (part 2)

Alexander McLain

September 25, 2023

# LDA and Logistic regression

- Recall that if $X \sim MVN(\mu_k, \Sigma_k)$ for $G = k$ and $\Sigma_k = \Sigma$ for all $k = 1, 2, \ldots, K$

$$\begin{aligned}\log \frac{\Pr(G = k|X = x)}{\Pr(G = \ell|X = x)} = \log \frac{\pi_k f_k(x)}{\pi_\ell f_\ell(x)} &= \log \frac{\pi_k}{\pi_\ell} - \frac{1}{2}(\mu_k + \mu_\ell)'\Sigma^{-1}(\mu_k - \mu_\ell) \\ &\quad +x'\Sigma^{-1}(\mu_k - \mu_\ell) \\ &= b_0 + \boldsymbol{b}_1' x\end{aligned}$$

  where $b_0 = \log \frac{\pi_k}{\pi_\ell} - \frac{1}{2}(\mu_k + \mu_\ell)'\Sigma^{-1}(\mu_k - \mu_\ell)$ and $\boldsymbol{b}_1 = \Sigma^{-1}(\mu_k - \mu_\ell)$.
- What does this look like?

# Logistic Regression

- If we are classifying $K$ classes using logistic regression

$$
\begin{aligned}
\log \frac{\Pr(G = 1|X = x)}{\Pr(G = K|X = x)} &= \beta_{01} + \boldsymbol{\beta}_1' \boldsymbol{x} \\
\log \frac{\Pr(G = 2|X = x)}{\Pr(G = K|X = x)} &= \beta_{02} + \boldsymbol{\beta}_2' \boldsymbol{x} \\
&\vdots \\
\log \frac{\Pr(G = K-1|X = x)}{\Pr(G = K|X = x)} &= \beta_{0K-1} + \boldsymbol{\beta}_{K-1}' \boldsymbol{x}
\end{aligned}
$$

# Logistic Regression

- We can then estimate the conditional probability of being in each class with

$$\begin{aligned}
\Pr(G = k|X = x) &= \frac{\exp(\beta_{0k} + \boldsymbol{\beta}'_k \boldsymbol{x})}{1 + \sum_{\ell=1}^{K} \exp(\beta_{0\ell} + \boldsymbol{\beta}'_\ell \boldsymbol{x})} \quad \text{for} \quad k = 1, \ldots, K-1 \\
\Pr(G = K|X = x) &= \frac{1}{1 + \sum_{\ell=1}^{K} \exp(\beta_{0\ell} + \boldsymbol{\beta}'_\ell \boldsymbol{x})}
\end{aligned}$$

- Let $\Pr(G = k|X = x; \theta) = p_k(\boldsymbol{x}; \theta)$ where $\theta = \{\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{K-1}\}$

# MLE Logistic Regression

- Here $\theta$ can be estimated my maximizing the log-likelihood

$$\ell(\theta) = \sum_{i=1}^{N} \log p_{g_i}(\boldsymbol{x}_i; \theta)$$

- Which for $K = 2$ can be written in terms of $\boldsymbol{\beta} = \{\beta_{01}, \boldsymbol{\beta}_1\}$ as

$$\ell(\theta) = \sum_{i=1}^{N} \left\{ y_i \boldsymbol{\beta}' \boldsymbol{x}_i - \log(1 + e^{\boldsymbol{\beta}' \boldsymbol{x}_i}) \right\}$$

where $\boldsymbol{x}$ now contains an intercept term.

# MLE Logistic Regression

- To maximize we take the derivative wrt $\boldsymbol{\beta}$ which yields

$$\frac{\partial \ell(\theta)}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} \boldsymbol{x}_i \left\{ y_i - p_{g_i}(\boldsymbol{x}_i; \boldsymbol{\beta}) \right\} = 0,$$

where $\frac{\partial \ell(\theta)}{\partial \boldsymbol{\beta}}$ is a $p+1$ dimensional vector. Since the first term in $\boldsymbol{x}_i$ is 1 the first score equation yields $\sum_{i=1}^{N} y_i = \sum_{i=1}^{N} p(\boldsymbol{x}_i; \boldsymbol{\beta})$

# MLE Logistic Regression

- The Newton-Raphson algorithm can be used to solve the score equations where, a single Newton update of $\boldsymbol{\beta}^{old}$ is

$$\boldsymbol{\beta}^{new} = \boldsymbol{\beta}^{old} - \left( \frac{\partial^2 \ell(\theta)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \right)^{-1} \frac{\partial \ell(\theta)}{\partial \boldsymbol{\beta}}$$

where

$$\frac{\partial^2 \ell(\theta)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = \sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i' p_{g_i}(\boldsymbol{x}_i; \boldsymbol{\beta}) \left\{ 1 - p_{g_i}(\boldsymbol{x}_i; \boldsymbol{\beta}) \right\}$$

# MLE Logistic Regression

- It's convenient to write this in matrix notation as

$$\frac{\partial \ell(\theta)}{\partial \boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{y} - \boldsymbol{p}), \quad \text{and} \quad \frac{\partial^2 \ell(\theta)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = -\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}$$

  where $\boldsymbol{W}$ is an $N \times N$ diagonal matrix where $W_{ii} = p_{g_i}(\boldsymbol{x}_i; \boldsymbol{\beta}) \{1 - p_{g_i}(\boldsymbol{x}_i; \boldsymbol{\beta})\}$.
- The Newton step can be represented as

$$\begin{aligned}
\boldsymbol{\beta}^{new} &= \boldsymbol{\beta}^{old} - (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{p}) \\
&= (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\{\boldsymbol{X}\boldsymbol{\beta}^{old} + \boldsymbol{W}^{-1}(\boldsymbol{y} - \boldsymbol{p})\} \\
&= (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{z}
\end{aligned}$$

# Logistic Regression with IRLS

- The last line re-expressed the with a weighted least squares step

$$\boldsymbol{z} = \boldsymbol{X}\beta^{old} + \boldsymbol{W}^{-1}(\boldsymbol{y} - \boldsymbol{p}) \quad \text{and } z_i = \boldsymbol{x}_i\beta^{old} + \frac{y_i - \hat{p}_i}{\hat{p}_i(1 - \hat{p}_i)}$$

  referred to as the *adjusted response*.
- Since $\boldsymbol{W}$ changes with $\beta$ this method is solved iteratively, and referred to as *iteratively reweighted least squares* or IRLS.
- For each step

$$\beta^{new} \leftarrow \text{argmin}_{\beta}(\boldsymbol{z} - \boldsymbol{X}'\beta)'\boldsymbol{W}(\boldsymbol{z} - \boldsymbol{X}'\beta).$$

  IRLS can be extended to the $K \geq 3$ multiclass case.

# Logistic versus LDA

- MLE is valid under general exponential family assumptions. Logistic is more robust.
- If no Gaussian or equal variances the logistic could be better.
- Logistic is less sensitive to outliers.
- LDA is more efficient than logistic.
- To get "good discrimination" logistic requires larger sample size than LDA.

# $L_1$ penalized logistic regression

- The $L_1$ Lasso penalty discussed previously can be used for variable selection and shrinkage.
- For the 2 class case we look to find

$$\max_{\boldsymbol{\beta}} \left[ \sum_{i=1}^{N} \left\{ y_i \boldsymbol{\beta}' \boldsymbol{x}_i - \log(1 + e^{\boldsymbol{\beta}' \boldsymbol{x}_i}) \right\} - \lambda \sum_{j=1}^{p} |\beta_j| \right]$$

- This is a difficult concave optimization, but the IRLS solution can be applied.
- This leads to a repeated application of a weighted lasso algorithm.

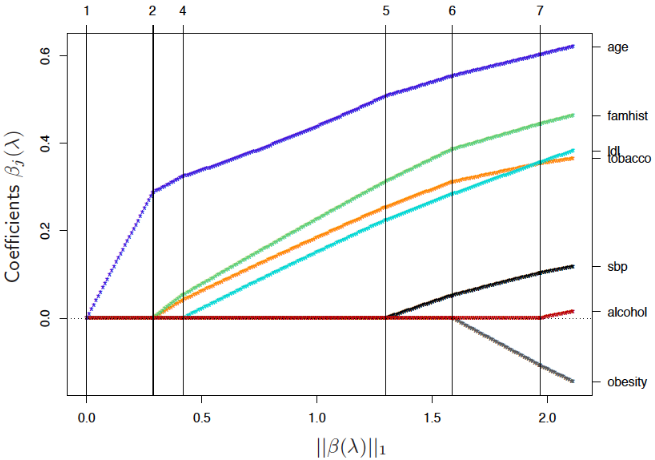


Figure: From ESL (online version). Note that the paths are not linear.