Therefore the discriminant function is effectively given by $\delta_k(x) = \hat{f}_k(x)$. This classification effective partitions the domain using linear decision boundaries.

Remark 4.1. It is easy to check that the entries of $\hat{f}_k(x)$ sum to 1, since

$$\begin{aligned}\hat{f}_k(x)^T 1_{K\times 1} &= (1, x^T)(X^T X)^{-1} X^T Y 1_{K\times 1} = (1, x^T)(X^T X)^{-1} X^T 1_{N\times 1} \\ &= (1, x^T)(X^T X)^{-1} X^T X_{\cdot 1} = [(1, x^T)(X^T X)^{-1} X^T X]_{\cdot 1} = 1.\end{aligned}$$

However $\hat{f}_k(x)$ is not a vector of probabilities, since it is a linear function of x and thus can easily contain values outside of the interval $[0, 1]$. Nevertheless, as seen from the classification function, larger entries $\hat{f}_k(x)$ to correspond to more favourable classes.

Remark 4.2. A disadvantage of classification via linear regression is the *masking* effect. In cases where $K >> p$, the domain of the linear function $\hat{f}_k(x)$ has much smaller dimension than its range. Since only the largest component $\hat{f}_k(x)$ is represented in the classification function, some classes may be completely neglected if for every x they are dominated by another class (even if they were say, the second largest out of K). This effect can be partially alleviated by including higher order terms (or other functions of x) and thus increasing the input dimension of the regression.

4.3 Discriminant analysis

Discriminant analysis is an example of a Bayesian classifier where each class is assigned a prior probability $P(G = k) = \pi_k$. Bayes' theorem tells us that

$$P(G = k \,|\, X = x) = \frac{P(X = x \,|\, G = k)\pi_k}{\sum_{l=1}^{K} P(X = x \,|\, G = l)\pi_l}.$$

We further assume that each class follows a multivariate Gaussian prior distribution $P(X = x \,|\, G = k) =\sim \mathcal{N}(\mu_k, \Sigma_k)$, so

$$P(X = x \,|\, G = k) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k)\right)$$

Recall that the optimal classifier is then

$$\hat{G}(x) = \arg\max_k P(G = k \,|\, X = x) = \arg\max_k P(X = x \,|\, G = k)\pi_k.$$

If we assume that all classes share a common covariance matrix $\Sigma_k = \Sigma$, then the decision boundary between two class can be simplified to

$$\ln \frac{P(G = k \,|\, X = x)}{P(G = l \,|\, X = x)} = \ln \frac{\pi_k}{\pi_l} - \frac{1}{2}(x - \mu_k)^T \Sigma^{-1}(x - \mu_k) + \frac{1}{2}(x - \mu_l)^T \Sigma^{-1}(x - \mu_l) \tag{4.1}$$

$$= \ln \frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k + \mu_l)^T \Sigma^{-1}(\mu_k - \mu_l) + x^T \Sigma^{-1}(\mu_k - \mu_l). \tag{4.2}$$

The cancellation of the quadratic term $x^T\Sigma^{-1}x$ means that the decision boundary is linear in x. Therefore this case is known as a *linear discriminant analysis* or *LDA*. As a result the discriminant function can be simplified to

$$\delta_k(x) = \ln \pi_k + x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k.$$

If the common covariance assumption is dropped, then the cancellation do no occur in (4.1) and the resulting decision boundaries will be quadratic in x. This case is known as a *quadratic discriminant analysis* or *QDA*. The discriminant function is

$$\delta_k(x) = \ln \pi_k - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k).$$

Suppose the training data are $(x_i, G_i), i = 1, \ldots, N$ and the number of observations in class K is given by N_k. In order to estimate the parameters π_k, μ_k and Σ_k, we may use the usual unbiased estimators for Gaussian distributions

$$\hat{\pi}_k = \frac{N_k}{N}, \quad \hat{\mu}_k = \sum_{G_i=k} \frac{x_i}{N_k},$$

$$\hat{\Sigma} = \frac{\sum_{k=1}^K \sum_{G_i=k}(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T}{N-K} \quad \text{for LDA},$$

$$\hat{\Sigma}_k = \frac{\sum_{G_i=k}(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T}{N_k - 1} \quad \text{for QDA}.$$

Both LDA and QDA are popular methods for classification that do not suffer from the masking effect which linear regression has. LDA is a simpler method with lower variance which performs well if the decision boundaries are believed to be linear. QDA has better bias if the data genuinely has different covariances, but it has higher variance due to the increased number of estimated parameters.

As an alternative to QDA, quadratic decision boundaries can be obtained by performing LDA on an enlarged data space which contains quadratic functions of x. QDA is still the preferred approach, although the extended LDA could be convenient if there is not enough data in some classes to estimate Σ_k accurately. The two approaches usually produce similar results.

Even if the initial training data is from a mixture model where the exact classification G_i are unknown, it is still possible to perform discriminant analysis (with decent accuracy) as long as we first estimate the parameters using methods such as the EM algorithm (see notes on EM algorithm).

4.4 Logistic regression

Logistic regression can be viewed as a more general version of LDA, in the sense that it only postulates that the decision boundaries of the Bayesian classifier are linear in