1. LDA and QDA

When there are 2 classes, LDA has a similar setup as the (equal-variance) t-test for one predictor, or the Hotelling's T^2-test for high-dimensional X. The difference is that the roles of X and Y are switched in LDA, as compared to their roles in t-test/Hotelling's test. Similarly, QDA is analogous to unequal-variance t-test or Hotelling's test.

Both LDA and QDA require multivariate normality (a very strong assumption) for X in every class. When equal covariance matrix is assumed, we have LDA; otherwise, QDA.

In a nutshell, in LDA/QDA, given any $\boldsymbol{x}$, when all classes have equal prior probability, we seek to identify the nearest class, that is, the class that has the smallest squared Mahalanobis distance, $d_M^2(\boldsymbol{x}, \mu_k)$, between $\boldsymbol{x}$ and the class center μ_k. If the prior probabilities are unequal, an offset $-2\log(\pi_k)$ is added to $d_M^2(\boldsymbol{x}, \mu_k)$, giving more chance to the class that has a higher prior probability. In QDA, the covariance matrices may not be equal, another offset $\log(|\Sigma_k|)$ is added, giving more chance to the class that has more "concentration". The following sections have the details.

1.1. *LDA*

Under the assumptions of equal variance and normality for the distribution of X within each class, $X \sim N(\mu_k, \Sigma)$. The density of an observation $\boldsymbol{x}$ from class k is

$$p(\boldsymbol{x}|k) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\mu_k)^T\Sigma^{-1}(\boldsymbol{x}-\mu_k)\right).$$

Let K be the number of classes, and π_k be the prior probability for class k. The marginal density for $\boldsymbol{x}$ is $p(\boldsymbol{x}) = \sum_{k=1}^K \pi_k p(\boldsymbol{x}|k)$, and the posterior probability for $\boldsymbol{x}$ in class k is

$$p(k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}, k)}{p(\boldsymbol{x})} = \frac{\pi_k p(\boldsymbol{x}|k)}{p(\boldsymbol{x})} \propto \pi_k p(\boldsymbol{x}|k).$$

Given any $\boldsymbol{x}$, we want to identify the class k that has the *largest* $p(k|\boldsymbol{x})$, or equivalently, the largest $\pi_k p(\boldsymbol{x}|k)$, or the largest $\log(\pi_k) + \log(p(\boldsymbol{x}|k))$, or the largest

$$\log(\pi_k) - \frac{1}{2}(\boldsymbol{x}-\mu_k)^T\Sigma^{-1}(\boldsymbol{x}-\mu_k) = -\frac{1}{2}\left[d_M^2(\boldsymbol{x}, \mu_k) - 2\log(\pi_k)\right],$$

where $d_M(\boldsymbol{x}, \mu_k) = \sqrt{(\boldsymbol{x}-\mu_k)^T\Sigma^{-1}(\boldsymbol{x}-\mu_k)}$ is the Mahalanobis distance (sphered distance) between $\boldsymbol{x}$ and μ_k, the center of class k. That is, the class k that has the *largest* $p(k|\boldsymbol{x})$ is the class that has the *smallest*

$$d_M^2(\boldsymbol{x}, \mu_k) - 2\log(\pi_k).$$

When π_k's are equal for all classes, the class k that has the largest $p(k|\boldsymbol{x})$ is the one that has the smallest Mahalanobis distance between its center and $\boldsymbol{x}$. This result is quite intuitive.

When the prior probabilities are unequal, an offset $-2\log(\pi_k)$ is added to $d_M^2(\boldsymbol{x}, \mu_k)$. The larger prior probability π_k, the more downward offset $-2\log(\pi_k)$, giving class k a lower start and thus a higher chance to be classified to.

1.2. *QDA*

When we do not assume equal covariance matrices across classes, the above derivation still works except that Σ now becomes Σ_k, and the class k that has the *largest* $p(k|\boldsymbol{x})$ has the largest

$$\log(\pi_k) - \frac{1}{2}(\boldsymbol{x} - \mu_k)^T \Sigma_k^{-1}(\boldsymbol{x} - \mu_k) - \frac{1}{2}\log(|\Sigma_k|),$$

or equivalently, the *smallest*

$$d_M^2(\boldsymbol{x}, \mu_k) - 2\log(\pi_k) + \log(|\Sigma_k|).$$

Note that there is an additional offset $\log(|\Sigma_k|)$. The more "concentrated" class k is, the smaller $|\Sigma_k|$ and $\log(|\Sigma_k|)$, giving class k a higher chance to be classified to.

1.3. *Namesake, prior distribution, and number of parameters*

For QDA, given any $\boldsymbol{x}$, we want to identify k that maximizes

$$\begin{aligned}\delta_k(\boldsymbol{x}) &= \log(\pi_k) - \frac{1}{2}(\boldsymbol{x} - \mu_k)^T \Sigma_k^{-1}(\boldsymbol{x} - \mu_k) - \frac{1}{2}\log(|\Sigma_k|) \\ &= \log(\pi_k) - \frac{1}{2}\boldsymbol{x}^T \Sigma_k^{-1}\boldsymbol{x} + \mu_k^T \Sigma_k^{-1}\boldsymbol{x} - \frac{1}{2}\mu_k^T \Sigma_k^{-1}\mu_k - \frac{1}{2}\log(|\Sigma_k|).\end{aligned}$$

These functions, $\delta_k(\boldsymbol{x})$, are called discriminant functions. They are quadratic in $\boldsymbol{x}$, and thus the name "quadratic discriminant analysis".

For LDA, Σ_k becomes Σ. Given any $\boldsymbol{x}$, we want to identify k that maximizes

$$\begin{aligned}\delta_k(\boldsymbol{x}) &= \log(\pi_k) - \frac{1}{2}\boldsymbol{x}^T \Sigma^{-1}\boldsymbol{x} + \mu_k^T \Sigma^{-1}\boldsymbol{x} - \frac{1}{2}\mu_k^T \Sigma^{-1}\mu_k - \frac{1}{2}\log(|\Sigma|) \\ &= \log(\pi_k) + \mu_k^T \Sigma^{-1}\boldsymbol{x} - \frac{1}{2}\mu_k^T \Sigma^{-1}\mu_k + c,\end{aligned}$$

where c does not depend on k. These discriminant functions are linear functions of $\boldsymbol{x}$, and thus the name "linear discriminant analysis". The boundaries from the LDA are linear.

The prior distribution should be the true outcome distribution. If the data were obtained as a random sample, the prior can be estimated by the sample empirical distribution; otherwise, it may be quite different from the sample empirical distribution. When the prior is not specified by the analyst, a random sample is often assumed by software.

Let K be the number of classes and p be the dimension of the predictors. The number of parameters in the LDA is $(K-1)(p+1)$; for QDA, $(K-1)(\frac{p(p+1)}{2} + p + 1)$. For example, if $K = 2$ and $p = 10$, then LDA has 11 parameters and QDA has 66 parameters; if $K = 4$ and $p = 10$, then LDA has 33 parameters and QDA has 198 parameters. (A little technical details: Thinking of all the parameters needed to express μ_k and Σ_k, we might derive that the LDA requires $Kp + p(p+1)/2$ parameters and the QDA requires $Kp + Kp(p+1)/2$. Some of these parameters are redundant because for classification, it suffices to know the $K-1$ differences $\delta_1(\boldsymbol{x}) - \delta_K(\boldsymbol{x})$ instead of the K discriminant functions. In both methods, the number of redundant parameters is $\frac{p(p+1)}{2} + p - (K-1)$.)