Outline

Naïve Bayes

If $p > n$ (for example more genes p than patients n), LDA (and certainly QDA and RDA) runs into problems.
Recall that the covariance matrix Σ is estimated from n observations. If $p > n$, then

$$\hat{\Sigma} = \frac{1}{n} \sum_{k=1}^{K} \sum_{j:Y_j=k} (X_j - \hat{\mu}_k)(X_j - \hat{\mu}_k)^T$$

is singular. As the inverse of $\hat{\Sigma}$ is used in LDA, it will fail.
An extreme regularization is to estimate Σ as above but set all non-diagonal elements to 0, i.e. ignoring dependence between predictor variables completely. This is sometimes referred to as *Naive Bayes*. All correlations between variables are effectively ignored in this way.
Alternatively, one can estimate Σ by using the estimate $\hat{\Sigma}$ as above and adding $\lambda 1_p$ for some $\lambda > 0$, where 1_p is the p-dimensional identity matrix (makes only sense if data have been standardized initially).

Applications to Classification of Documents

Given documents such as emails, webpages, scientific articles, books etc., we might be interested in learning a classifier based on training data to automatically classify a new document. Possible classes could be spam/non-spam, academic/commercial webpages, maths/physics/biology etc.
Many popular techniques rely on simple probabilistic models for documents. Given a prespecified dictionary, we extract high-dimensional features such as absence/presence of a word (multivariate Bernoulli), number of occurences of a word (multinomial) etc.
Parameters within in class can be estimated through Maximum Likelihood. However Maximum Likelihood overfits so we will need to derive more robust alternative.