Application of LDA and QDA on MNIST dataset

Arkonil Dhar Saumyadip Bhowmik Shreya Pramanik
Shubha Sankar Banerjee Souvik Bhattacharya

Summer Project 2021
Indian Institute of Technology Kanpur

July 16, 2021

Data Description The training data consists of 50000 observations on the label (response y) which has $K = 10$ classes viz. $\{0, 1, 2, \ldots, 9\}$ and 784 features (the independent variables $X = (X_1, \ldots, X_{784})$).

In a typical learning algorithm such as Logistic Regression, we try to directly model the conditional probability of y given x i.e. $p(y|x)$. In case of discriminant analysis, we try to model the distribution of predictors X separately in each response classes (i.e. given y), and then use Bayes' theorem to flip these around into estimates for $p(y = k|X = x)$, $\forall k = 1(1)10$. In our case, since we will be dealing with LDA and QDA, the distributions are assumed to be Normal.

Bayes' Theorem for Classification

We wish to classify an observation into one of k classes which are inherently distinct and unordered in nature. Let $\pi_k = p(y = k)$, $\forall k = 1(1)10$ be the prior probability that a randomly chosen observation arises from the k^{th} class. Let $f_k(x) = p(X = x|y = k)$ denote the density function of X assuming that it arises from the class k. The Bayes' theorem states that

$$p(y = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{k=1}^{10} f_k(x)\pi_k} \tag{1}$$

This is the posterior probability that an observation X belongs to the k^{th} class. The prior probability is estimated by simply calculating the fraction of total observations that fall in the k^{th} class, i.e.

$$\hat{\pi}_k = \frac{\sum_{i=1}^{n} 1\{y_i = k\}}{n}, \ \forall k = 1(1)10 \tag{2}$$

Decision Rule

We calculate the posterior probability of an obsevation arising from all the possible classes one at a time. We then assign the label based on which class has the highest posterior probability of the observation arising from that class.

Linear Discriminant Analysis

In the Eq (1), note that x is a vector of $p = 784$ elements. We then assume that X is drawn from a p-dimentional *multivariate Gaussian distribution*. Note that a multivariate gaussian distribution assumes that each component is distributed as univariate normal distribution with some amount of correlation with the other components.
Thus the assumption we make here is $X \sim \mathcal{N}(\mu, \Sigma)$, where $\mathbb{E}(X) = \mu^{p\times 1}$ and $\mathbf{Cov}(X) = \Sigma^{p\times p}$. The density is then of the followinf form,

$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right) \tag{3}$$

For each of the k classes, we calculate the means of all the components of observation belonging to a particular class and obtain the estimate of class-specific mean vector $\hat{\mu}_k,\ \forall k = 1(1)10$ and assume that the **variance and covriance of the components are same throughout all the K classes** and then end up with the estimate of Σ as $\hat{\Sigma}$. Thus,

$$\hat{\mu}_k = \frac{\sum_{i=1}^{n} 1\{y_i = k\}x_i}{\sum_{i=1}^{n} 1\{y_i = k\}},\ \forall k = 1(1)10 \tag{4}$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_{y_i})(x_i - \hat{\mu}_{y_i})',\ \forall k = 1(1)10 \tag{5}$$

Plugging in estimates (4) and (5) into the density (3), and using it and the prior estimate in (2) in (1), we get the estimated posterior probability of an observation arising from a class k as:

$$\hat{p}(y = k|X = x) = \frac{\hat{\pi}_k \frac{1}{(2\pi)^{p/2}|\hat{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(x-\hat{\mu}_k)'\hat{\Sigma}^{-1}(x-\hat{\mu}_k)\right)}{\sum_{k=1}^{10} \hat{\pi}_k \frac{1}{(2\pi)^{p/2}|\hat{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(x-\hat{\mu}_k)'\hat{\Sigma}^{-1}(x-\hat{\mu}_k)\right)} \tag{6}$$

Finding the class having higher posterior density is equivalent to finding the class for which the following discriminant function is the highest.

$$\hat{\delta}_k(x) = x'\hat{\Sigma}^{-1}\hat{\mu}_k - \frac{1}{2}\hat{\mu}_k'\hat{\Sigma}^{-1}\hat{\mu}_k + \log\hat{\pi}_k \tag{7}$$

For two class comparison (say for classes k and l, $\in k \neq l$), we compute the region where the posterior probability of an observation arising from either of the two classes is equal.

$$\begin{aligned}
& \hat{\delta}_k(x) = \hat{\delta}_l(x) \\
\implies \quad & x'\hat{\Sigma}^{-1}\hat{\mu}_k - \frac{1}{2}\hat{\mu}_k'\hat{\Sigma}^{-1}\hat{\mu}_k + \log\hat{\pi}_k = x'\hat{\Sigma}^{-1}\hat{\mu}_l - \frac{1}{2}\hat{\mu}_l'\hat{\Sigma}^{-1}\hat{\mu}_l + \log\hat{\pi}_l \\
\implies \quad & x'\hat{\Sigma}^{-1}(\hat{\mu}_k - \hat{\mu}_l) = \frac{1}{2}\hat{\mu}_k'\hat{\Sigma}^{-1}\hat{\mu}_k - \frac{1}{2}\hat{\mu}_l'\hat{\Sigma}^{-1}\hat{\mu}_l + \log(\hat{\pi}_l/\hat{\pi}_k)
\end{aligned} \tag{8}$$

Note that this region is defined by a *linear* function of the p dimensional feature vectors, and hence results in a linear decision boundary.

Further as evident from Eq (7), the form of the discriminant function is linear in x and hence this learning algorithm is called Linear Discriminant Analysis.

Quadratic Discriminant Analysis

In QDA we assume that the observations are drawn from Multivariate Gaussian Distribution with class specific means and covariance matrices, i.e., $X \sim \mathcal{N}(\mu_k, \Sigma_k)$. (In case of LDA, we assumed that covariance matrix is common to all the K classes). In this case the estimate of Covariance matrix differs from the one in case of LDA (Eq (5)), as follows:

$$\hat{\Sigma}_k = \frac{\sum_{i=1}^{n_k} 1\{y_i = k\}(x_i - \hat{\mu}_{y_i})(x_i - \hat{\mu}_{y_i})'}{n_k}, \; n_k = \sum_{i=1}^{n} 1\{y_i = k\}, \; \forall k = 1(1)10 \tag{9}$$

The resultant discriminant function is of the form:

$$\begin{aligned} \hat{\delta}_k(x) &= -\frac{1}{2}(x - \hat{\mu}_k)'\hat{\Sigma}_k^{-1}(x - \hat{\mu}_k) + \log \hat{\pi}_k \\ &= -\frac{1}{2}x'\hat{\Sigma}_k^{-1}\hat{x} + x'\hat{\Sigma}_k^{-1}\hat{\mu}_k - \frac{1}{2}\hat{\mu}_k'\hat{\Sigma}_k^{-1}\hat{\mu}_k + \log \hat{\pi}_k \end{aligned} \tag{10}$$

Note that the discriminant function is actually a *Quadratic* function in x, and gives rise to quadratic decision boundaries, hence called Quadratic Discriminant Analysis.