Linear Discriminant Analysis

- LDA can be used for multi-class classification unlike logistic regression.
- LDA is more stable than logistic regression when N is small, and when classes are well separated.

Model

LDA tries to approximate the Baye's classifier (the classifier that has the lowest error rate) presented below:

$$P(Y=k \mid X=\boldsymbol{x}) = \frac{\overbrace{P(Y=k)}^{\text{(prior)}} \cdot \overbrace{P(X=\boldsymbol{x} \mid Y=k)}^{\text{(likelihood)}}}{\sum_{l=1}^{K} P(Y=l) \cdot P(X=\boldsymbol{x} \mid Y=l)}$$

where, k represents a class; K is the total number of classes

$\boldsymbol{x}$ represents the feature vector

$P(Y=k \mid X=\boldsymbol{x})$ is the conditional probability that $\boldsymbol{x}$ belongs to class k.

$P(Y=k)$ is the marginal probability of the response. This is also known as the prior.

$P(X=\boldsymbol{x} \mid Y=k)$ is the probability density function of class K. It gives the likelihood of $\boldsymbol{x}$ given the distribution of K.

Fitting the Model

We need two quantities to come up with the Baye's classifier:

1. $P(Y=k)$: To compute the prior, we simply compute the fraction of data points present in each class. (prior)
   Sometimes, priors could also be computed based on some information given to us.

2. $P(X=x \mid Y=k)$:
   We assume this density function to be a Gaussian distributed (univariate in case of single predictor; multivariate in case of multiple predictors). Number of distributions $= K$.

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

i.e. observations from k^{th} class are drawn from $N(\mu_k, \Sigma_k)$ where $x, \mu_k, \Sigma_k \in \mathbb{R}^p$.

Estimating $P(X \mid Y)$ means estimating Gaussian means estimating μ_k and Σ_k:
μ_k is mean of feature vectors for class k.
Σ_k is covariance matrix of features for class k. It is assumed to be equal for all classes, i.e. $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k = \Sigma$

Now, given the prior ($P(Y)$) and density functions (Gaussian with μ_K and Σ_K, we can compute class K for which $P(Y|X)$ is largest.

Computing $\underset{K}{\text{argmax}}$ (Baye's classifier) = computing $\underset{K}{\text{argmax}}$ (discriminant function)

$$\Rightarrow \quad K = \underset{K}{\text{argmax}}\ \delta(x)$$

where $\delta(x)$ (discriminant function) $= x^T \Sigma^{-1} \mu_K - \frac{1}{2} \mu_K^T \Sigma^{-1} \mu_K + \log p(Y=K)$

The discriminant function is linear in x, hence the name "linear" discriminant analysis.

Assumptions

All assumptions lie in estimating the density function $P(X|Y)$.

1. Each class has Gaussian distribution — huge assumption.
2. All classes shares covariance matrix Σ. This is addressed by quadratic discriminant analysis.

Quadratic Discriminant Analysis

- Model is same as LDA.
- Fitting is same as LDA except that QDA does not assume covariance matrices for each class to be the same i.e. $\Sigma_1 \neq \Sigma_2 \neq \cdots \neq \Sigma_K$.

The above assumption leads to change in the discriminant function which now becomes quadratic in $\boldsymbol{x}$:

$$K = \arg\max_K \delta(K)$$

where, $$\delta(K) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_K)^T \boldsymbol{\Sigma}_K^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_K) + \log p(Y = K)$$

↑ quadratic function

↑ each class has a different Σ

- Assumption: The only assumption is that the density functions are Gaussian.

Note: LDA is better than QDA where the data has linear decision boundary, and when there are more predictors (per observation).

QDA is better when decision boundary is non-linear but needs large N to estimate each Σ properly.

LDA - high bias, less variance

QDA - less bias, high variance

- When true decision boundary is linear : LDA, log reg > QDA, KNN
- When X are normally distributed : LDA, QDA > log reg