Model-Based Predictive Models

LDA & QDA

BAYES FORMULA

- The Bayes theorem gives us the following formula to compute the probability that the record belongs to class Ci:

$$P(C_i|X_1,\ldots,X_p) = \frac{P(X_1,\ldots,X_p|C_i)P(C_i)}{P(X_1,\ldots,X_p|C_1)P(C_1) + \cdots + P(X_1,\ldots,X_p|C_m)P(C_m)}.$$

Where

Ci : classes of interest

$X_1, X_2, \ldots X_p$: Variables which co-exist with Classes of interest

Bayes Theorem

$$P(C_i|X_1,\ldots,X_p) = \frac{P(X_1,\ldots,X_p|C_i)P(C_i)}{P(X_1,\ldots,X_p|C_1)P(C_1)+\cdots+P(X_1,\ldots,X_p|C_m)P(C_m)}.$$

- $P(C_i)$ are called prior probabilities. We can find them by dividing the incidences of occurrence of C_i by total number of observations.
- In place of $P\left(X_1, X_2, \ldots X_p \middle| C_i\right)$, we can also write a continuous function like probability density function of normal distribution as $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$

$$P(C_i|X_1) = \frac{P(C_i)\frac{1}{\sigma_i\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu_i}{\sigma_i})^2}}{\sum_{i=1}^{p} P(C_i)\frac{1}{\sigma_i\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu_i}{\sigma_i})^2}} \quad \ldots\ldots (\mathrm{I})$$

Sane's STATS Academy of Statistics

LDA – Univariate

- We can estimate the parameters (μ_k, σ_k^2) from the data and use the expression (I) as classifying the observation to that class i for which $P(C_i|X_1, X_2, \dots X_p)$ will be maximum. But we have a better approach than this by solving this expression to $\delta_i(x)$ given below.
- We assume here $\sigma_i = \sigma$, a constant for all the classes.
- We can solve expression (I) by taking log of terms of both sides which finally results into the following expression

$$\boxed{\delta_i(x) = x\frac{\mu_i}{\sigma^2} - \frac{\mu_i^2}{2\sigma^2} + \log(\mathrm{P}(C_i))}$$

- Observe here that the function $\delta_i(x)$ is linear function in x. Hence the term Linear Discriminant Analysis.
- Each test observation is assigned to that class i, for which $\delta_i(x)$ is maximum. This function is a one-dimensional form of linear discriminating function.

Sane's STATS Academy of Statistics

Multivariate LDA

- The operations done on one variable can be extended to multiple variables and the expression $\delta_i(x)$ can be written as

$$\boxed{\delta_i(\bar{x}) = x^T \Sigma^{-1} \mu_i - \frac{1}{2}\mu_i^T \Sigma^{-1} \mu_i + \log(P(C_i))}$$

Where

Σ : Covariance Matrix

x: vector of variables x_i

μ_i: Mean of variable x_i

Note: We assume that the covariance matrix is same for all the classes

Sane's STATS Academy of Statistics

Multivariate QDA

- In Quadratic Discriminant Analysis, we assume that the covariance matrix $\sum$ is different for each class i.
- $\sum_i$: Covariance matrix for class i.
- Hence the discriminating function changes to

$$\boxed{\delta_k(\bar{x}) = -\frac{1}{2}(x-\mu_i)^T \sum\nolimits_i^{-1}(x-\mu_i) + \log(P(C_i))}$$

Assumptions of LDA & QDA

- Predictors are all numeric
- Predictors have a multivariate normal distribution
- LDA : All the variances and covariances for all the classes are same
- QDA : All the variances and covariances for each class is different

LDA & QDA in Python

- LDA & QDA in Python can be performed with the function LinearDiscriminantAnalysis and QuadraticDiscriminantAnalysis from the **sklearn.discriminant_analysis** respectively.

Example: Satellite Imaging

- Consider the dataset Satellite.csv
- The dataset consists of the multi-spectral values of pixels in 3x3 neighborhoods in a satellite image, and the classification associated with the central pixel in each neighborhood. The aim is to predict this classification, given the multi-spectral values.
- Variable **classes** is the target(response) variable.

```
In [75]: df.head()
Out[75]:
   x.1  x.2  x.3  x.4  x.5  x.6  x.7  x.8  x.9  x.10     ...      x.28  x.29  \
0   92  115  120   94   84  102  106   79   84   102     ...       104    88
1   84  102  106   79   84  102  102   83   80   102     ...       100    84
2   84  102  102   83   80  102  102   79   84    94     ...        87    84
3   80  102  102   79   84   94  102   79   80    94     ...        79    84
4   84   94  102   79   80   94   98   76   80   102     ...        79    84

   x.30  x.31  x.32  x.33  x.34  x.35  x.36    classes
0   121   128   100    84   107   113    87  grey soil
1   107   113    87    84    99   104    79  grey soil
2    99   104    79    84    99   104    79  grey soil
3    99   104    79    84   103   104    79  grey soil
4   103   104    79    79   107   109    87  grey soil

[5 rows x 37 columns]
```

Questions?