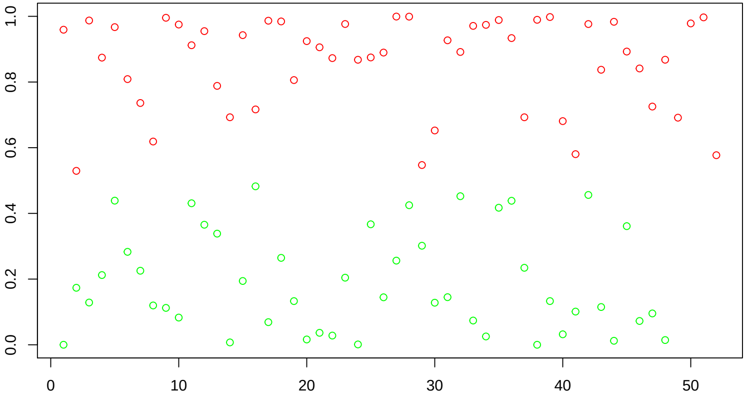
Logistic Regression and Discriminant Analysis

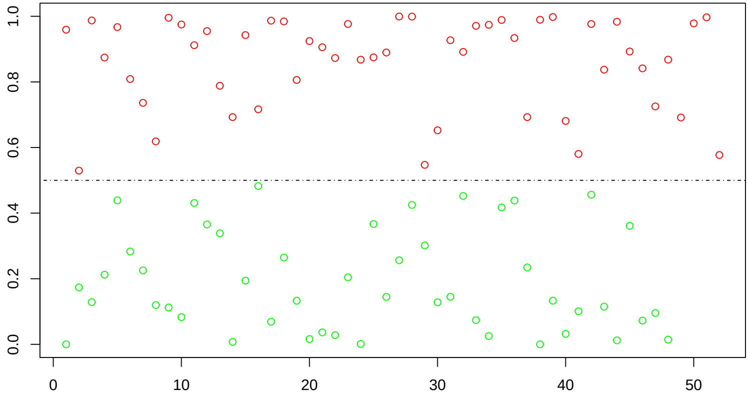
Tathagata Basu

UE de Master 2, AOS1

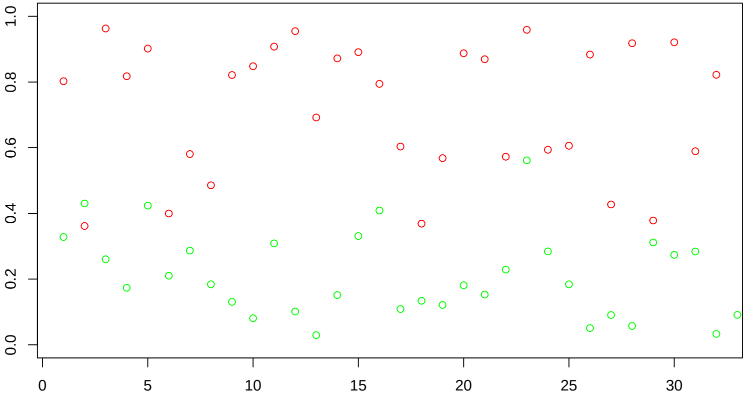
Autumn 2022

Outline I

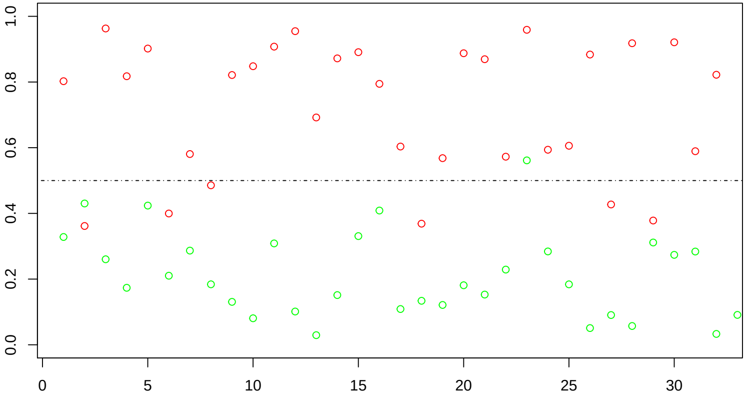

0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50



0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50



0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30


0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30

Classification

Let $C = (c_1, \cdots, c_n)$ such that $c_i \in \{0, 1\}$. Let $a := (a_1, a_2, \ldots, a_p)$ denote attributes.
We define

$$\pi(a) := E(C \mid a) = P(C = 1 \mid a). \tag{1}$$

Therefore, we can say, C is Bernoulli distributed with parameter $\pi(a)$ such that, the likelihood is given by:

$$\mathcal{L}(\pi \mid C, a) = \prod_{i=1}^{n} \pi(a)^{c_i}(1 - \pi(a))^{1-c_i} \tag{2}$$

Regression model

Let $b := (b_1, b_2, \ldots, b_p)^T$ denote regression coefficients. Then $\pi(a)$ can be defined by:

$$\pi(a) = h\left(a^T b\right) \tag{3}$$

where h acts as a 'link' function.
For logistic regression the link function is given by:

$$h(x) := \frac{\exp(x)}{1 + \exp(x)}. \tag{4}$$

Logistic Regression

Now, replacing our regression model, in the likelihood we get,

$$\mathcal{L}(\pi \mid C, a) = \mathcal{L}(b \mid C, a) = \prod_{i=1}^{n} \left[h\left(a^T b\right)\right]^{c_i} \left[1 - h\left(a^T b\right)\right]^{1-c_i} \tag{5}$$

Then the log likelihood is given by:

$$\log(\mathcal{L}(b \mid C, a)) = \sum_{i=1}^{n} \left(- C_i \left(a_i^T b\right) - \log\left(1 + \exp(a_i^T b)\right) \right). \tag{6}$$

For, maximum likelihood estimates we find

$$\hat{b} := \arg\max_b \{\log(\mathcal{L}(b \mid C, a))\}. \tag{7}$$

Logistic Regression

Now, replacing our regression model, in the likelihood we get,

$$\mathcal{L}(\pi \mid C, a) = \mathcal{L}(b \mid C, a) = \prod_{i=1}^{n} \left[h\left(a^T b\right)\right]^{c_i} \left[1 - h\left(a^T b\right)\right]^{1-c_i} \tag{5}$$

Then the log likelihood is given by:

$$\log(\mathcal{L}(b \mid C, a)) = \sum_{i=1}^{n} \left(-C_i \left(a_i^T b\right) - \log\left(1 + \exp(a_i^T b)\right) \right). \tag{6}$$

For, maximum likelihood estimates we find

$$\hat{b} := \arg\max_b \{\log(\mathcal{L}(b \mid C, a))\}. \tag{7}$$

Regularisation

In high dimensional problems ($p > n$), may contribute to overfitting.

So we can use Bayesian regularisation. A natural choice for prior for b is Gaussian distribution, then our log-posterior becomes

$$\log(P(b \mid C, a) \equiv \log(\mathcal{L}(b \mid C, a)) - \lambda \|b\|^2 \tag{8}$$

By maximising our log-posterior, we can obtain MAP estimates of b.

Motivation I

- Classical logistic regression models are easy but has limited expressiveness.
- We can construct a probabilistic model with Gaussian assumption instead.

Mixture Model

Let $(x_1, \cdots, x_n)$ be a p-dimensional in $\mathbb{R}^p$ and z be corresponding classes such that for $1 \leq i \leq n$, $z_i \in \{1, 2, \cdots, M\}$.

We assume that,

$$x_i \mid z_i = k, \mu_k, \Sigma_k \sim \mathcal{N}(\mu_k, \Sigma_k). \tag{9}$$

Then, we have

$$x_i, z_i = k \mid \theta \sim P(z_i = k) P(x_i \mid z_i = k, \mu_k, \Sigma_k) \tag{10}$$

where $\theta \equiv (\mu_1, \cdots \mu_M; \Sigma_1, \cdots, \Sigma_M)$

Joint Likelihood

The joint likelihood is given by:

$$L(\theta \mid X, Z) = \prod_i P(X = x_i, z = z_i \mid \theta) \tag{11}$$

$$= \prod_i \prod_k \left[P(X = x_i, z = k \mid \theta) P(z_i = k)\right]^{z_{ik}} \tag{12}$$

where, $z_{ik} = \mathbb{I}(z_i = k)$.

So, we have

$$L(\theta \mid X, Z) = \prod_i \prod_k \left[\pi_k f_k(x_i)\right]^{z_{ik}} \tag{13}$$

where $P(z = k) = \pi_k$ and $f_k(x_i) = P(X = x_i, z = k \mid \theta)$.

Estimation I

a) π_k
Lagrangian:

$$\mathcal{L}(\theta) = \ln L(\theta \mid \dots) - \lambda \left(\sum_k \pi_k - 1 \right)$$

$$\begin{cases} \frac{\partial \mathcal{L}(\theta)}{\partial \pi_k} = \sum_i \frac{z_{ik}}{\pi_k} - \lambda \\ \frac{\partial \mathcal{L}(\theta)}{\partial \lambda} = 1 - \sum_k \pi_k \end{cases}$$

$$\left.\begin{array}{l} \frac{\partial \mathcal{L}(\theta)}{\partial \pi_k} = 0 \Rightarrow \pi_k = \frac{1}{\lambda} \sum_i z_{ik} \\ \frac{\partial \mathcal{L}(\theta)}{\partial \lambda} = 0 \Rightarrow \sum_k \pi_k = 1 \end{array}\right\} = \frac{1}{\lambda} \sum_i \sum_k z_{ik} = 1 \Leftrightarrow \lambda = \sum_{i,k} z_{ik} = n$$

Estimation II

$$\Rightarrow \mathcal{L}(\theta) = 0 \Leftrightarrow \hat{\pi}_k = \frac{1}{n} \sum_i z_{ik}$$

b) μ_k

$$\frac{\partial \ln L(\theta)}{\partial \mu_k} = \frac{\partial}{\partial \mu_k} \left[\sum_{i,k} z_{ik} \ln f_k(x_i \theta) \right] \tag{14}$$

$$= C_1 \frac{\partial}{\partial \mu_k} \left[\sum_{i,k} z_{ik} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) \right] \tag{15}$$

$$= C_2 \sum_i z_{ik} (x_i - \mu_k) \tag{16}$$

Matrix Differentiation

Let $M = [m_{ij}]_{p \times p}$ be a $p \times p$ matrix such that

$$\frac{\partial f(M)}{\partial M} = \begin{bmatrix} \frac{\partial f(M)}{\partial m_{11}} & \cdot & \frac{\partial f(M)}{\partial m_{1p}} \\ \vdots & & \\ \frac{\partial f(M)}{\partial m_{p1}} & \cdots & \frac{\partial f(M)}{\partial m_{pp}} \end{bmatrix}. \tag{17}$$

Then the following holds

- $\frac{\partial |M|}{\partial M} = |M|(M^{-1})^T$
- $\frac{\ln f(M)}{\partial M} = \frac{1}{f(M)} \frac{\partial f(M)}{\partial M}$
- $\frac{\partial \mathrm{tr}(MA)}{\partial M} = A^T$

Estimation III

c)Σ_k

Using the rule of matrix differentiation, we can show that

$$\hat{\Sigma}_k = \frac{\sum_i z_{ik} \hat{B}_{ik}}{\sum_i z_{ik}} \tag{18}$$

where

$$B_{ik} = (x_i - \mu_k)(x_i - \mu_k)^T. \tag{19}$$

$$\hat{B}_{ik} = (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T. \tag{20}$$

Prediction

Reminder

$$P(Z \mid X) = \frac{P(X \mid Z)P(Z)}{P(X)} = \frac{P(X \mid Z)P(Z)}{\sum_z P(X \mid Z = z)P(Z = z)}$$

$$\Rightarrow P(Z = k \mid X = x) = \frac{\pi_k f_k(x)}{\sum_k \pi_k f_k(x)} \texttt{ for any } x$$

$\Rightarrow$ our posterior probability estimates are

$$\hat{P}(Z = k \mid x) = \frac{\hat{\pi}_k f_k(x \mid \hat{\mu}_k, \hat{\Sigma}_k)}{\sum_\ell \hat{\pi}_\ell f_\ell(x \mid \hat{\mu}_\ell, \hat{\Sigma}_\ell)}$$

Bayesian Regularisation

In some cases, we might not have enough data to train our model. Instead, we can take help of Bayesian methods. We can use

- Multivariate Gaussian for μ_k so that $\mu_k \sim \mathcal{N}(\mu_{kp}, \Sigma_k / K_{kp})$
- Inverse Wishart for Σ_k so that $\Sigma_k \sim IW(\nu_{kp}, \Lambda_{kp})$

Both of these are conjugate priors and give us simple closed form expressions.

Posterior Estimates

Recall $\hat{B}_{ik} = (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$. Then we have the following posterior estimates

$$\hat{\mu}_k = \frac{\sum_i z_{ik} x_i + K_{kp}\mu_{kp}}{\sum_i z_{ik} + K_{kp}} \tag{21}$$

$$\hat{\Sigma}_k = \frac{\sum_i z_{ik}\hat{B}_{ik} + K_{kp}(\hat{\mu}_{kp} - \mu_{kp})(\hat{\mu}_{kp} - \mu_{kp})^T + \Lambda_{kp}^{-1}}{\sum_i z_{ik} + \nu_{kp} + p + 2} \tag{22}$$