Classification

given some data, which of the given x classes does that observation belong to

Why not LR.

Eg → classify { Epilepsy, drug overdose, Stroke }

I can encode the output as 1, 2, 3

this is because giving 1, 2, 3 then first would imply DO = stroke − Epilepsy that is correct so we wont get right answer

for 2 classes we can use a dummy variable $y = \begin{cases} 0 & \text{if stroke} \\ 1 & \text{if drug overdose} \end{cases}$

Why not to use LR

1) wont work for > 2 classes

2) Don't provide meaningful estimates.

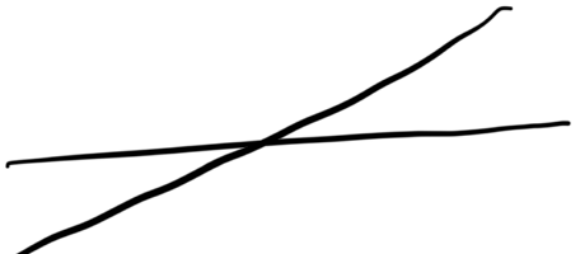
Logistic Regression

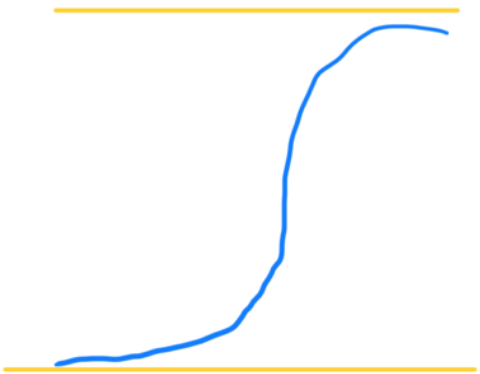
models the probability that Y - belongs to a particular category

eg → Default → balance

Log.R. → P(default = Yes / balance)

P(balance)





Using Linear Regression

Using logistic regression

if we model with

$$p(X) = \beta_0 + \beta_1 X$$

then we can get a -ve prob. & a prob. $P(X) > 1$

to get output between 0 and 1,

$$P(x) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

We use a method Maximum Likelihood

Indicator.

$$\frac{P(x)}{1 - P(x)} = e^{\beta_0 + \beta_1 X}$$

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x$$

↳ Log odds / Logit

Estimating the Regression coefficients

Likelihood function: -

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod (1 - p(x_i))$$

We have z-statistic to measure significance.

$$H_0 : \beta_1 = 0 \Rightarrow p(x) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$$

Multiple Logistic Regression

Multiple logistic Regression

$$\log\left(\frac{P(x)}{1-P(x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

$$P(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots \beta_p x_p}}$$

Using P21 may give a diff co-eff in terms of its sign when compared to p-p in MLR. as other params may also effect it.

{ This is known as Confounding }

effect of outcome of one variable is mixed up with the effect of another variable.

Multinomial Logistic Regression

eg → stroke, drug overdose, epileptic seizure.

$K = 3$; W.L.O.G

we select K^{th} class

$$P(Y = k / X = x) = \frac{e^{\beta_0 + \ldots + \beta_{kp} x_p}}{1 + \sum_{l=1}^{k-1} e^{\beta_{l0} + \beta_{l1} + \ldots \beta_{lp}}}$$

$$\log\left(\frac{P(Y=K/X=x)}{P(Y=K/X=x)}\right) = \beta_{K0} + \beta_{K_1} x_1 + \ldots + \beta_{Kp} x_p$$

Softmax coding → alternate for Multinomial Regression

We treat all K classes symmetrically

$$P(\ldots) \quad \beta_0 + \beta_{k_1} x_1 + \ldots + \beta_{kp} x_p$$

$$P(Y=k \mid X=x) = \frac{e^{\beta_{k0} + \beta_{k1} x_1 + \dots + \beta_{kp} x_p}}{\sum_{l=1}^{K} e^{\beta_{l0} + \beta_{l1} x_1 + \dots + \beta_{lp} x_p}}$$

$$\log\left(\frac{P(Y=k \mid X=x)}{P(Y=k' \mid X=x)}\right) = (\beta_{k0} - \beta_{k'0}) + (\beta_{k1} - \beta_{k'1}) x_1 + \dots + (\beta_{kp} - \beta_{k'p}) x_p$$

Generative model for Classification

$P(Y=k \mid X=x) \longrightarrow$ Bayes's Theorem

→ when there is substantial separation bit the two classes. The parameter estimates for logistic regression are highly unstable.

for $K \geq 2$

$\pi_K \rightarrow$ prior probability.

$f_K(x) = P(X / Y = K)$

density fⁿ of x, for an obs. coming from K^{th} class.

using Bayes theorem

$$P(Y = K / X = x) = \frac{\pi_K f_K(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

$P_K(x) = P(Y = k / X = x)$

π_K is simple $\rightarrow$ $\frac{\text{no. of obs. in } K \text{ class}}{\text{Total Obs.}}$

finding $f_K(x)$ is tricky. typically we have to make simplifying assumptions.

Discriminant analysis

A) Linear Discriminant Analysis

⇒ assume $p = 1$ {only 1 predictor}

⇒ we assume $f_k(x)$ is Normal

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{(x-\mu_1)^2}{2\sigma_k^2}}$$

⇒ assume $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \cdots = \sigma_k^2 = \sigma^2$ for simplicity.

$$P_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} e^{\left(-\frac{(x-\mu_k)^2}{2\sigma^2}\right)}}{\sum_{i=1}^{k} \pi_i \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{(x-\mu_i)^2}{2\sigma^2}\right)}} \quad \text{— (1)}$$

from StatQuest → LDA reduces the dimensionality and helps project the data points on a single line

$\mu_k \rightarrow$ Mean of k^{th} category.

diff. of means $\swarrow$ $\dfrac{|\mu_1 - \mu_2|^2}{s_1^2 + s_2^2}$ $\swarrow$ $\rightarrow$ maximize this

$\rightarrow$ minimize this

$\swarrow$ spread

$\therefore$ Maximize seperation and minizing the scatter / spread.

this is better than PCA for higher dimensions.

taking log of eq^n 1 and re-arranging

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

eg: for $K = 2$ & if $\pi_1 = \pi_2$

decision boundary $\rightarrow x = \dfrac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \dfrac{\mu_1 + \mu_2}{2}$

In reality we plug estimates for μ_K, μ_K and σ^2

$$\hat{\mu}_K = \frac{1}{n_K} \sum_{i:y_i=k} x_i \quad ; \quad \hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

where n = total number of observations.

$$\hat{\pi}_K = n_K / n$$

$$\hat{\delta}_K(x) = x \cdot \frac{\hat{\mu}_K}{\hat{\sigma}^2} - \frac{\hat{\mu}_K^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_K) \qquad \text{2.2.}$$

discriminant function

assign the observation to the class for which eq^n 2.2 is the largest

Major assumption is that each observation [illegible] normal distribution

comes from normal distr...

LDA : Linear discriminant Analysis for $p > 1$

$$X = (X_1, X_2, \ldots X_p)$$

Multivariate normal distribution.

$$X \sim N(\mu, \Sigma)$$

$E(X) = \mu$ is the mean of X (a vector with p components)

$Cov(X) = \Sigma$ is a $p \times p$ covariance matrix.

$$f(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}\left[(x-\mu)^T \Sigma^{-1} (x-\mu)\right]}$$

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$

Measuring the error → Confusion Matrix for credit default

		True: No	True: Yes	Σ
Preds	No	9644	252	9896
	Yes	23	81	104
		9667	333	

$$\text{Accuracy} = \frac{\text{correct}}{\text{total}} \times 100 = 97.25\%$$

$$\text{error} = 2.75\%$$

but a trivial pred if it always says no. It have an error of 3.33%. so it is only slightly worse.

but of 333 actual defaulters it predicted 252 to not default that is an error of 75.7%

which is unacceptable.

FN ≈ 75.7%

Precision → $\frac{\text{True +ve}}{TP + FP}$

Recall / Sensitivity → $\frac{TP}{TP + FN}$

for the example ≈ 24.2%

Specificity → $\frac{TN}{TN + FP}$ → False that are detected correctly.

F1 score → $\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Recall} + \text{Precision}}$

This large disparity is because LDA is rewarded to minimize the inaccurate preds so it optimizes based on data

the one with majority, but a banking company will prefer LDA correctly identifying defaulters by sacrificing some accuracy on non-defaulters.

Modifying it to create a better classifier

$$P(\text{default} = \text{Yes} \mid X = x) > 0.5$$

We can lower this >50% to 20% if we are concerned by incorrectly predicting the default class

$$P(\text{default} = \text{Yes} \mid X = x) > 0.2$$

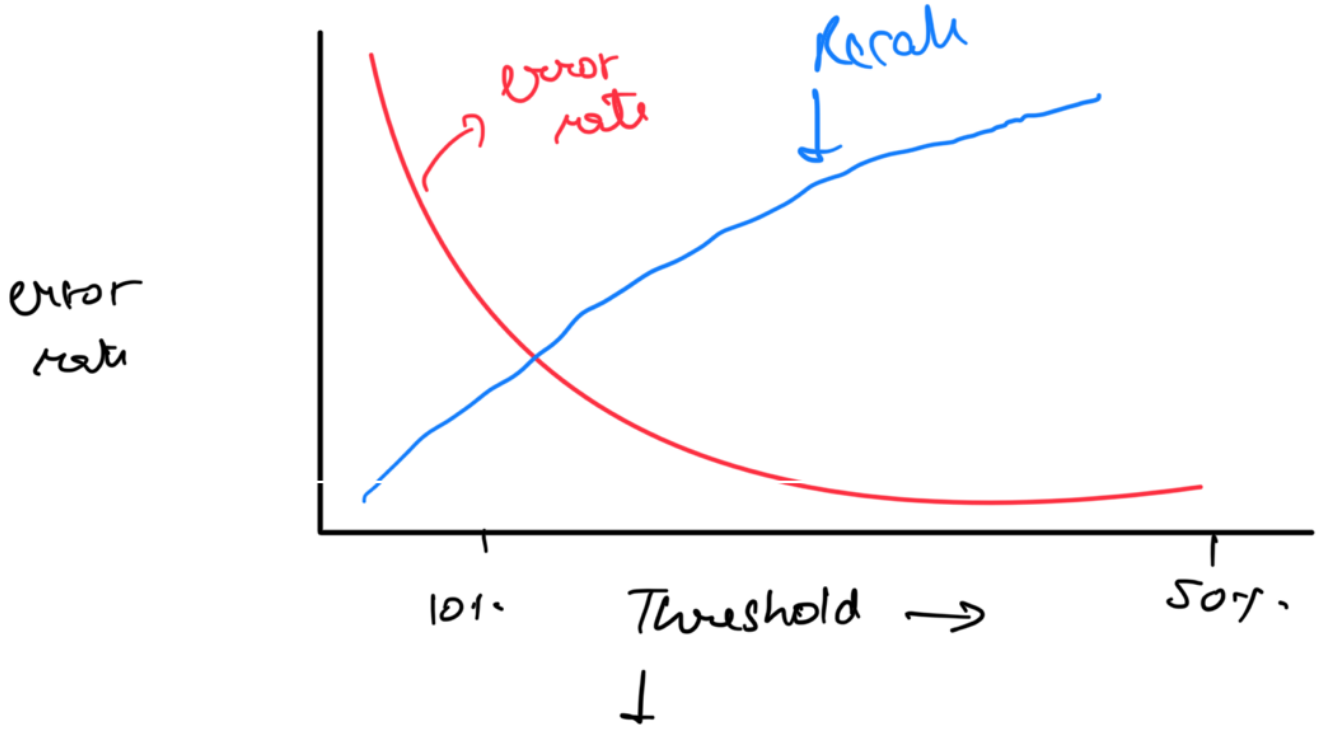
taking 20% we get

Pred \ True	Yes	No
Yes	195	235

No	138	9432
	333	9667

So Recall $= \frac{138}{333} = 41.4\%$

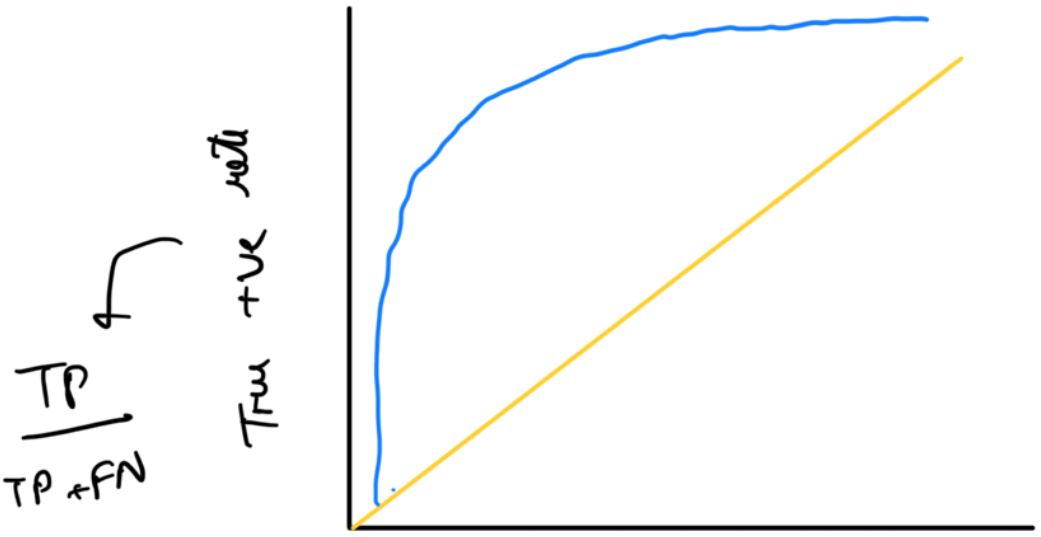
error rate increased to 3.73%.



$P(\text{default} = \text{Yes} \mid X = x) \geq \text{Threshold}$

deciding the threshold value requires domain knowledge

ROC → Receiver operating Characteristic



area under curve of ROC shows accuracy closer to 1 is best.

by chance classifier will have 0.5

Quadratic discriminant Analysis

→ more flexible

→ covariance matrix

Naive Bayes Classifier

$$p_K(x) = P(Y = K \mid X = x)$$

assumption → within the k^{th} class, the p predictors are independent

$$f_K(x) = f_{K1}(x_1) \cdot f_{K2}(x_2) \cdots \times f_{Kp}(x_p)$$

$\pi_K \rightarrow$ prior of being in class K

$f_{Kp}(x_p) =$ if it was in class K then p of x_p

$$P(Y = K \mid X = x) = \frac{\pi_K \times f_{K1}(x_1) \cdots \times f_{Kp}(x_p)}{\sum_{l=1}^{K} \pi_l \cdot f_{l1}(x_1) \times \cdots \times f_{lp}(x_p)}$$

Bike share Data

Poisson Regression

$$P(Y=K) = \frac{e^{-\lambda}\lambda^{K}}{K!} \qquad K \in N \cup \{0\}$$

$$E(Y) = \lambda$$

$$\log(\lambda(x_1, \ldots x_p)) = \beta_0 + \beta_1 x_1 + \ldots \beta_p x_p$$

(3) — $$\lambda(x_1, x_2, \ldots x_p) = e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}$$

{so that $\lambda(x_1, \ldots x_p)$ takes non-ve values}

$$l(\beta_0, \beta_1, \ldots \beta_p) = \prod_{i=1}^{n} \frac{e^{-\lambda(x_i)} \lambda(x_i)^{y_i}}{y_i!} \quad \text{— (4)}$$

where $\lambda(x_i) = e^{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}}$ {from eq^n 3}

Under poisson regression

$\text{var}[Y] = E[Y] = \lambda$

Generalised Linear Models

all use predictors $X_1, X_2, \ldots X_p$

Method	distribution
Linear Regression	Normal
Logistic Regression	Bernoulli
Poisson Regression	Poisson

Expected Value

$$\text{LiR} \quad E(Y/X_1, \ldots X_p) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

$$\text{LoR} \quad E(Y/X_1, \ldots, X_p) = \frac{e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \ldots \beta_p X_p}}$$

$$\text{PoR} \quad E(Y/X_1 \ldots X_p) = \lambda(X_1 \ldots X_p)$$

$$= e^{\beta_0 + \beta_1 X_1 + \ldots \beta_p X_p}$$

In general

$$\eta\left[E(Y/x_1,\ldots x_p\right] = \beta_0 + \beta_1 x_1 + \ldots \beta_p x_p$$

LiR $\quad \eta(u) = u$

LoR $\quad \eta(u) = \log\left(\frac{u}{1-u}\right)$

PoR $\quad \eta(u) = \log(u)$

These are GLM Generalised Linear Modeling (GLM)