4. Classification

- qualitative variable, ~~same~~ also referred to as ~~quantitative~~ categorical variable.
- classification: approaches for predicting qualitative responses.
- classifier: classification technique
  - 1. Logistic regression
  - 2. Linear discriminant analysis.
  - 3. Quadratic discriminant analysis
  - 4. naive Bayes
  - 5. K-nearest neighbors.

4.2 Why not linear regression?

- The categorical data is not ordered If linear regression is applied using coded dummy variables, different set of predictions will be produced for each encoding.
- The above is especially the case when there are more than two categories which don't have any natural ordering.

Reasons:

(a) A regression method cannot accomodate a qualitative response with more than two classes.

(b) a regression method will not provide meaningful estimates of $Pr(Y|X)$ - even for two classes.

ex: $P(y_i|x_0) > 1$ or $P(y_i|x_0) < 0$ for some x_0.

4.3 Logistic regression

Let, $Y = \begin{cases} 1 & ; \text{ category 1} \\ 0 & ; \text{ category 2} \end{cases}$

$P(Y=1|X) = p(X)$: response, X: predictor.

If consider: $p(X) = \beta_0 + \beta_1 X$. → ~~odd~~ probabilities may not make sense

- we must model $p(X)$ using a function that gives outputs between 0 to 1 $\forall x$.

logistic function, $p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$.

$$\text{odds} := \frac{p(x)}{1 - p(x)} = e^{\beta_0 + \beta_1 X} \longrightarrow (0, \infty).$$

- ~~0~~ & ∞ indicate very low & very high probabilities of response variable.

$$\text{log odds / logit} = \log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 X.$$

interpretation: increasing X by one unit changes log odds by β_1.

→ it multiplies the odds by e^{β_1}.

In $p(x)$ due to one unit change in X, depends on the current value of X.

but irrespectively of value of X:

if $\begin{cases} (i)\ \beta_1 > 0 \Rightarrow X \uparrow \text{ then } p(x) \uparrow \\ (ii)\ \beta_1 < 0 \Rightarrow X \uparrow \text{ then } p(x) \downarrow \end{cases}$

Regression coefficients can be estimated by maximising the likelihood function: $(\hat{\beta_0}, \hat{\beta_1})$

$$\ell(\beta_0, \beta_1) = \prod_{i: y_i = 1} p(x_i) \prod_{i: y_i' = 0} (1 - p(x_i))$$

prediction:

$$\hat{p}(x) = \frac{e^{\hat{\beta_0} + \hat{\beta_1} x}}{1 + e^{\hat{\beta_0} + \hat{\beta_1} x}}$$

increase in one unit of x is associated with ~~a~~ increase in log odds of y by $\hat{\beta_1}$.

4.3.4 Multiple logistic regression.

predicting a binary response using multiple multiple predictors

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

where $X = (x_1, x_2, \ldots, x_p)$ are p predictors.

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

⟶ maximise likelihood function to get estimates.

4.3.4 : Multinomial logistic regression.

extension of logistic approach to ~~2 class~~ (no. of classes) > 2 case! : multinomial logistic regression.

select a single class to be baseline : k^{th} class.

$$P(Y=k \mid X=x) = \frac{e^{\beta_{0k} + \beta_{k1}x_1 + \cdots + \beta_{kp}x_p}}{1 + \sum_{l=1}^{K-1} e^{\beta_{l0} + \beta_{l1}x_1 + \cdots + \beta_{lp}x_p}}$$

for $k = 1, \ldots, K-1$ &

$$P(Y=K \mid X=x) = \frac{1}{1 + \sum_{l=1}^{K-1} e^{\beta_{l0} + \beta_{l1}x_1 + \cdots + \beta_{lp}x_p}}$$

for $k = 1, \ldots, K-1$

$$\log\left(\frac{P(Y=k \mid X=x)}{P(Y=K \mid X=x)}\right) = \beta_{k0} + \beta_{k1}x_1 + \cdots + \beta_{kp}x_p.$$

- it is not necessary to treat the k^{th} class as baseline.
- coefficients will be different, but the predictions will be same (the outputs will be same).

4.4. Generative models for classification

Different approach:

(1) we model the distribution of the predictors x separately in each of the response class (i.e. each value of y).

(2) Use Bayes's theorem to find $P(Y=k \mid X=x)$ estimates.

If $x \mid y \sim N$ then model is similar to logistics.

Q. Why not logistic regression?

(1) If there is substantial separation between the two classes: parameter estimates for the logistic regression model - unstable

(2) If distribution of the predictors X is approx normal in each of the classes & sample size is small.

(3) easy extension to more than 2 classes.

π_k → estimated by fraction of observation belonging to k^{th} class.

π_k: prior probability that a randomly chosen observation comes from the k^{th} class.

$f_k(x) = P(X|Y=k)$ denote the density function of X for an obs that comes from the k^{th} class.

by Bayes thm:

$$P(Y=k|X=x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

(π_k, $f_k(x)$ → estimates; whole expression → Bayes classifier.)

Bayes classifier.

Let $p_k(x) \equiv P(Y=k|X=x)$

:= This is the posterior probability that an observation $X=x$ belongs to the k^{th} class, given the predictor value for that observation.

Instead of ~~code~~ Estimating LHS, we estimate RHS.

Three classifiers that use different estimates of $f_k(x)$ to approximate the Bayes classifier:

(1) Linear discriminant analysis.

(2) quadratic discriminant analysis.

(3) Naive Bayes.

(1) Linear Discriminant Analysis for $p=1$

$p=1 \Rightarrow$ only one predictor.

x classified to k^{th} class s.t $p_k(x)$ is max.

Assumption: $f_k(x)$ is Gaussian.

(1) $f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left\{ -\frac{(x-\mu_k)^2}{2\sigma_k^2} \right\}$

(2) $\sigma_1^2 = \sigma_2^2 = \cdots \sigma_K^2 = \sigma^2$

$$\therefore p_k(x) = \frac{\pi_k \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_k)^2}{2\sigma^2}\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu_2)^2}{2\sigma^2}\right)}$$

maximize:

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

boundary: $\delta_1(x) = \delta_2(x)$ & $\pi_1 = \pi_2$.

$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}$$

discriminant function:

$$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)$$

Ests: estimates

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{y_i = k} x_i$$

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

$$\hat{\pi}_k = n_k / n$$

"linear" as $\delta_k(x)$ is linear function of x.

Linear discriminant analysis for $p > 1$.

Assumption: $x = (x_1, x_2, \ldots, x_p)$ drawn $N(\underset{\sim}{\mu}, \Sigma)$

$\underset{\sim}{\mu}$: $(\mu_1, \mu_2, \ldots, \mu_p)$

Σ: common variance covariance matrix

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

- observations in the k^{th} class are drawn from a multivariate Gaussian distribution $N(\mu_k, \Sigma)$

→ k^{th} class s.t. $\delta_k(x)$ is maximum:

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

- Boundary: for $k \neq l$.

$$\delta_k(x) = \delta_l(x)$$

$$x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k = x^T \Sigma^{-1} \mu_l - \frac{1}{2} \mu_l^T \Sigma^{-1} \mu_l$$

- sensitivity & specificity are measures of performance of LDA.

4.4.3 Quadratic Discriminant Analysis.

- $X \sim N(\mu_k, \Sigma_k)$ (i.e. variance covariance is not common for all classes)
- observations drawn from each class follow Gaussian distribution.

- The Bayes classifier assigns an observation $X = x$ to the class for which $\delta_k(x)$ is max where:

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

$$= -\frac{1}{2}x^T \Sigma_k^{-1} x + x^T \Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k^T \Sigma_k^{-1}\mu_k - \frac{1}{2}\log|\Sigma_k| + \log \pi_k.$$

→ As first term is quadratic form, QDA gets its name.

* How to decide LDA or QDA?

- No. of parameters to be estimated: for p many predictors.

In QDA: $K \times \left[p + \binom{p}{2}\right]$

- K: classes
- p: means for each x_j
- $\binom{p}{2}$: variance covariance matrix entries.

$$= K\left[p + \frac{p(p-1)}{2}\right]$$

$$= K\left[\frac{2p + p(p-1)}{2}\right]$$

$$= K\left[\frac{p(p+1)}{2}\right]$$

$$= \frac{Kp(p+1)}{2}$$ → to be estimated parameters.

In LDA: $K \times p \times 1$

- K: classes
- p: means
- 1: σ^2 is same for all classes.

⇒ LDA is much less flexible classifier than QDA. and so has substantially lower variance ⇒ improved prediction performance

- If $\sigma^2 = \sigma_K^2$ assumption is not good then high bias.
- QDA: if dataset is very large. - else LDA.

4.4 Naive Bayes classifier.

- Assumption:
within the k^{th} class, the p predictors are independent

mathematically:

$$f_k(x) = f_{k1}(x_1) * f_{k2}(x_2) \times \cdots \times f_{kp}(x_p)$$

where: f_{kj} : is the density function of the j^{th} predictor among observations in the k^{th} class.

- used when n is not very large compared to p.

$$p(Y=k|X=x) = \frac{\pi_k \times f_{k1}(x_1) \times f_{k2}(x_2) \times \cdots \times f_{kp}(x_p)}{\sum_{l=1}^{K} \pi_l \times f_{l1}(x_1) \times f_{l2}(x_2) \times \cdots \times f_{lp}(x_p)} \quad k = 1, \cdots, K$$

- To estimate the one-dimensional density functions: f_{kj}

(1) X_j is quantitative $\quad X_j | Y=k \sim N(\mu_{jk}, \sigma_{jk}^2)$

- Assumptn: within each class, the j^{th} predictor is drawn from a univariate normal distribution.

- $Cov(X_i, X_j) = 0 \; (i \neq j) \quad \sigma_{jk}^2 \neq \sigma_{ik}^2$.
(only different variances).

(2) Estimating $f_{kj}(x_j)$ as the fraction of the training observations in the k^{th} class that belong to the same histogram bin as x_j.
OR Kernel density estimator.

(3) X_j is qualitative: proportion of training observations for the j^{th} predictor corresponding to each class.

4.5 Comparison of classification methods.

4.5.1 An analytical comparison.

~~k: general class that~~

k: a class that needs to be assigned to new observation if the condition is satisfied.

K: baseline.

maximizes:

$$\log\left[\frac{P(Y=k|X=x)}{P(Y=K|X=x)}\right] \quad \text{for } k = 1, 2, \cdots, K.$$

① LDA: $X_k \sim N_k(\mu_k, \Sigma)$.

$$\log\left[\frac{P(Y=k|X=x)}{P(Y=K|X=x)}\right] = \log\left(\frac{\pi_k f_k(x)}{\pi_K f_K(x)}\right)$$

$$= \log\left(\frac{\pi_k \cdot \exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)\right)}{\pi_K \exp\left(-\frac{1}{2}(x-\mu_K)^T \Sigma^{-1}(x-\mu_K)\right)}\right)$$

$$= \log\left(\frac{\pi_k}{\pi_K}\right) - \frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k) + \frac{1}{2}(x-\mu_K)^T \Sigma^{-1}(x-\mu_K)$$

$$= a_k + \sum_{j=1}^{p} b_{kj} x_j$$

LDA: $a_k + \sum_{j=1}^{p} b_{kj} x_j$ is

Q.DA: $a_k + \sum_{j=1}^{p} b_{kj} x_j + \sum_{j=1}^{p}\sum_{l=1}^{p} c_{kjl} x_j x_l$

Naive Bayes: $a_k + \sum_{j=1}^{p} g_{kj}(x_j)$

↳ form of generalized linear models.

(1) LDA is special case of QDA with $c_{kjl} = 0 \quad \forall j, l, k$.

(2) LDA is special case of Naive Bayes with ~~$b_{kj} x_j$~~

$g_{kj}(x_j) = b_{kj} x_j$

(3) ~~Proof of LDA~~

$f_{kj}(x_j) = N(\mu_{kj}, \sigma_j^2)$

then $g_{kj}(x_j) = b_{kj} x_j$

where $b_{kj} = \frac{(\mu_{kj} - \mu\mu_{kj})}{\sigma_j^2}$

Then naive Bayes is special case of LDA with Σ restricted to diagonal matrix with j^{th} diagonal element equal to σ_j^2.

(4) Neither QDA nor ~~Bayes~~ naive Bayes is a special case of the other.

- As QDA has $c_{kjl} x_j x_l$ terms : it has potential to be more accurate in settings where interactions among the predictors are important in discriminating between classes.
- choice of method will depend will depend on (1) the distribution of the predictors
(2) values of n & p.

For multi-nomial logistics regression:

$$\log\left[\frac{P(Y=k \mid X=x)}{P(Y=K \mid X=x)}\right] = \beta_{k0} + \sum_{j=1}^{p} \beta_{kj} x_j$$

↳ identical to LDA.

- KNN is completely non parametric approach: no assumptions are made about the shape of decision boundary.

(1) Dominates LDA & logistic when the decision boundary is highly non-linear, provided $n \gg p$.

(2) $n \gg p$, tends to reduce the bias while incurring a lot of variance.

(3) n is not very large, p is not very small then QDA.

(4) KNN does not tell us which predictors are important.

~~Empirical Comparisons~~.

Poisson regression: $Y \sim pois(\lambda)$ · Y: counts.

neither quantitative nor qualitative

$$\log(\lambda(X_1, X_2, \dots, X_p)) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p.$$

$$\rightarrow \lambda(X) = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p}.$$

Estimation:

$$L(\beta) = \prod_{i=1}^{n} \frac{e^{-\lambda(x_i)} (\lambda(x_i))^{y_i}}{y_i!}$$

where $\lambda(x_i) = e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}$.

Generalized Linear Model.

$$X \rightarrow y \rightarrow E(Y|X) \rightarrow \eta(Y|X) = \beta_0 + \sum \beta_j X_j$$

* Link function (η) applies transformation to $E(Y|X_1, \dots, X_p)$ so that the transformed mean is a linear function of predictors.

$\eta(\mu) = \mu \rightarrow$ linear

$\eta(\mu) = \log\left(\frac{\mu}{1-\mu}\right) \rightarrow$ logistic

$\eta(\mu) = \log \mu \rightarrow$ poisson.