Classification

most of the problems would be to classify. be it classifying observations into categories, or predicting the category of a new observation.

examples of classification problems:

1) set of symptoms of a person → which medical condition?
2) online banking service based on IP address, transaction history etc → fraud or not?
3) DNA sequences with & without diseases → which DNA mutations are disease causing?

Reasons why we can't use linear regression for classification

1) regression method can't accommodate a qualitative response with more than 2 classes
(the order would alter the prediction coefficients.
we can only set the order if the difference b/w the classes are clear, & equal. such as low, medium, high)
2) a regression method will not provide meaningful estimates of $P_r(Y|X)$, even with just 2 classes
(the probability value won't be bound b/w 0 & 1)

Logistic model

$p(X) = \beta_0 + \beta_1 X$ - but the value will not be bound to $[0,1]$

to avoid this problem, model $p(X)$ using a function that gives the outputs between 0 & 1 for all values of X

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

the logistic function will always produce an S shaped curve, regardless of the value of X.

$$\frac{p(X)}{p(-X)} \quad \frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$$ - odds ratio $[0,\infty)$

By taking logs at both sides

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

log odds or Logit

increase of one unit of X changes the log odds by β_1
the amount that $p(X)$ changes depends on the current value of X.
if β_1 is +ve, increasing X will increase $\hat{p}(X)$
& vice versa.

maximum likelihood - we seek the estimates for β_0 & β_1 such that the predicted probability $\hat{p}(x_i)$ of a category is as close to the category of individual's observation.

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

multiply the prob of all observations where outcome is 1
outcome is 0

$H_0 : \beta_1 = 0$ (probability doesn't depend on the predictor)
z statistic high → reject the null hypothesis
z statistic of $\beta_1 = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$
z stat high ⇒ p stat low

H_0 implies that $p(x) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$

?? Why z statistic here, but not t statistic?

For qualitative predictors, use dummy variables like 0,1 etc.

confounding - when the coefficient is of one sign when considered alone vs considered with other predictors (correlated)

ex: students on avg are more likely to default

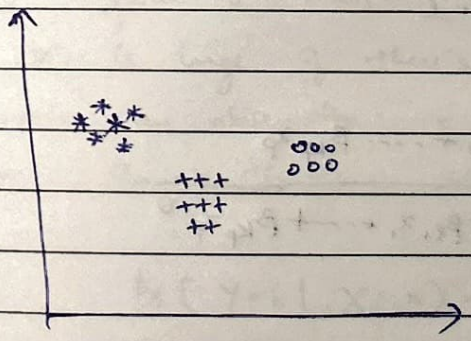
but when only the credit card balance is considered, their prob of default is lower than non students

hence it is important to identify and consider different params before taking a decision from model

Multiple logistic regression:- $\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$

$$p(x) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_p X_p}}$$

Multinomial Logistic Regression



select a single class as baseline then calculate the prob of others classes with reference to this

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

for $k = 1$ to $K-1$

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

$$\log\left(\frac{P_r(Y=k \mid X=x)}{P_r(Y=K \mid X=x)}\right) = \beta_{k0} + \beta_{k1} x_1 + \beta_{k2} x_2 + \cdots \beta_{kp} x_p$$

coefficient estimates will differ based on the chosen baseline class. but the prediction, log odds ratio will be same.
be careful when interpreting the coefficients.
ex: if we set A to be baseline (A, B, C classes), then we can interpret β_B as the log odds of B vs A.

further one unit increase in X_j is associated with a β_{Bj} increase in the log odds of B over A.

if X_j increases by one unit, then

$$\frac{Pr(Y=B \mid X=x)}{Pr(Y=A \mid X=x)} \text{ increases by } e^{\beta_{Bj}}$$

Softmax coding - Rather than selecting a baseline class, we treat all K classes symmetrically and assume for $k = 1, 2, \ldots K$

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

$$= \frac{e^{\beta_{k0} + \beta_{k1}x_1 + \cdots + \beta_{kp}x_p}}{\sum_{l=1}^{K} e^{\beta_{l0} + \beta_{l1}x_1 + \cdots + \beta_{lp}x_p}}$$

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

Generative models for classification

Logistic Regression involves directly modeling $Pr(Y=k \mid X=x)$.

in another approach, we first model the distributions of predictors X separately in each of the response classes (for each value of Y)

then use Baye's theorem to flip these around into estimates for $Pr(Y=k|X=x)$

When the distribution of X within each class is assumed to be normal, it turns out that the model is very similar to logistic regression model.

Why do we need another method?

- when there's substantial separation b/w 2 classes, the parameter estimates for the logistic regression model are unstable
- if the distribution of predictors X is approximately normal in each of the classes & the sample size is small, then these approaches are 111" to logistic regression

π_k - prior probability that a random observation comes from k^{th} class

$f_k(x) = P(X|Y=k)$ - density function of X if an observation from k^{th} class

($f_k(x)$ is large if there's high prob that ~~x=x~~ in k^{th} class, there's an observation i.e $X \approx x$. & 111" for lower prob)

then Baye's theorem states that

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

$p_k(x) = Pr(Y=k|X=x)$ - posterior prob that an observation $X=x$ belongs to class k.

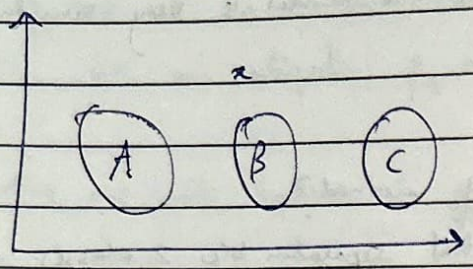
calculating π_k is easy. find out fraction of observations that belong to class k. (prob of class k)

but ~~calculating~~ estimating the density function f_k requires some assumptions

Some techniques - linear discriminant analysis, quadratic discriminant analysis, naive Bayes

Bayes theory: $P(\theta|data) \propto P(data|\theta) \times P(\theta)$

posterior — likelihood — prior



$P(A|x) \propto p(x|A)\, p(A)$

A function $f(x)$ is a density function if

- $f(x) \geq 0$ for all x
- $\int_{-\infty}^{\infty} f(x)\,dx = 1$
- $P(a \leq X \leq b) = \int_a^b f(x)\,dx$

example: if $X \sim N(0,1)$

then $f(0)$ is high → many values near 0

$P(-1 \leq X \leq 1)$ is area under the curve

density function describes how probability is distributed over values of a continuous random variable

Probability mass function

- $p(x) \geq 0$
- $\sum p(x) = 1$
- $P(a \leq X \leq b) = \sum_a^b p(x)$

this is for discrete random variables

example: coin toss

$P(\text{head}) = 0.5$ · $P(\text{tail}) = 0.5$

$$P_r(Y=k \mid X=x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{k} \pi_l f_l(x)} \quad (*)$$

Linear discriminant analysis for $p=1$

assume $f_k(x)$ is normal or Gaussian.

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

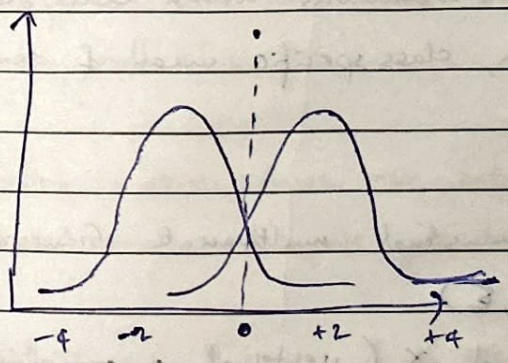
assume $\sigma_1^2 = \cdots = \sigma_k^2$

Substituting (1) in (*) and taking the log

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

if $K=2$ & $\pi_1 = \pi_2$, then Bayes classifier assigns an observation to class 1 if $2x(\mu_1-\mu_2) > \mu_1^2 - \mu_2^2$.

else to class 2



$\mu_1 = -1.25$

$\mu_2 = 1.25$

$\sigma_1^2 = \sigma_2^2 = 1$

assume $\pi_1 = \pi_2 = 0.5$

decision boundary - point at which $\delta_1(x) = \delta_2(x)$ & that's equal to $\frac{\mu_1^2 - \mu_2^2}{2(\mu_1-\mu_2)} = \frac{\mu_1+\mu_2}{2}$

then for the above graph $\frac{-1.25+1.25}{2} = 0$.

if the point is before 0, it is class 1. else class 2.

in reality, to apply Bayesian classifier we need to know

$\mu_1, \mu_2 \cdots \mu_k$, σ^2, $\pi_1 \pi_2 \cdots \pi_k$

linear discriminant analysis estimates these values

$\hat{\mu}_k = \frac{1}{n_k} \sum_{i: y_i = k} x_i$ — avg of all training obs from class k

$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i: y_i = k} (x_i - \hat{\mu}_k)^2$ — weighted avg of sample variances for each of K classes

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

plugging the above

$$\hat{\delta}_k(x) = x \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k) \quad -(3)$$

assign $X = x$ to k for which (3) gives the largest value

LDA classifies assumes that observations within each class come from normal distribution, class specific mean & common variance σ^2

<u>LDA for $p > 1$</u>

to indicate that a p dimensional multivariate Gaussian distn, we write $X \sim N(\mu, \Sigma)$.

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

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

multivariate Gaussian density fn.

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

in the case of $p > 1$, LDA assumes that observations in k^{th} class are drawn from multivariate Gaussian distribution $N(\mu_k, \Sigma)$, μ_k is class specific mean vector,

Σ is covariance matrix thats common to all classes.

Performing some algebra, we get-

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

assign to class k for which the above formula gives largest value.
above is the matrix equivalent of ② 2 pages ago.

as before, we need to estimate $\mu_1, \mu_2 \ldots \mu_k, \pi_1, \pi_2 \ldots \pi_k, \Sigma$

lookout for low error rates: it can be because
- higher the p:n ratio - overfitting
- when the observations in a class itself is low, a useless classifier which predicts wrong will have low error rate (only 3% ppl default. so a classifier that predicts no one defaults will have error rate 3%.)
- aka null classifier

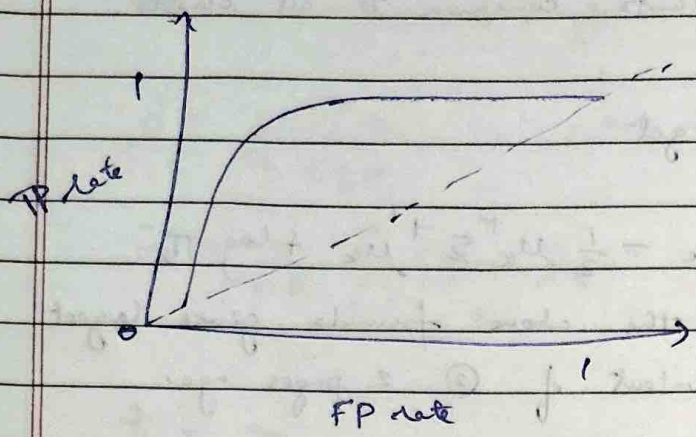
confusion matrix shows the categories of how many were correctly predicted and not; we get to know false +ve & false -ve

Sensitivity - true +ve % all +ve : TP / TP+FN

Specificity - true -ve % all -ve : TN / TN+FP

in a Bayes classifier, we can change the threshold to accomodate ~~the risk~~ to handle the risk thats caused by FP & FN according to business / problem

Roc curve - displays 2 types of errors for all possible thresholds, a good Roc curve will hug the top left corner.
overall performance of the classifier is given by area under the Roc curve



----- → no information classifier, useless

		True Class		
Predicted class		−	+	
	−	TN	FN	N*
	+	FP	TP	P*
		N	P	

FP rate	FP / N	Type 1 error, 1 − specificity
TP rate	TP / P	1 − type II error, power, recall, sensitivity
Pos pred value	TP / P*	precision, 1 − false discovery proportion
Neg pred value	TN / N*	

OMG.

Quadratic discriminant analysis (QDA)

Unlike LDA, here it's assumed that each class has its own covariance matrix Σ_k

$X \sim N(\mu_k, \Sigma_k)$

$\delta_k(x)$ = (Some huge formula)

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

Why LDA & QDA? ⇒ bias-variance tradeoff

bias - error from wrong assumptions in model
high bias = underfitting
straight line fitted on curved line

variance - over sensitive
high variance = overfitting
wiggly line that passes through every data point

- when there are p predictors, estimating covariance matrix requires $p(p+1)/2$ parameters
for QDA when Σ is different for each class, $Kp(p+1)/2$ parameters
- so, if we assume a single covariance matrix in LDA, it becomes linear in x - only Kp linear coefficients.
LDA is much less flexible,
- if the assumption that K classes share same covariance matrix is false, then LDA can suffer from high bias.
- prefer
LDA - Relatively fewer training observations
QDA - training set is huge

Naive Bayes

In LDA & QDA the assumption that f_k is a multivariate normal random variable distribution, class specific mean μ_k, single & class specific covariance matrix (Σ & Σ_k) were made. & the problem of estimating K p-dimensional density fns were replaced into estimating K p-dimensional vectors & 1 & K pxp covariance matrices

in Naive Bayes we make assumption that within k^{th} class, the p predictors are independent

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

estimating a p dimensional density f_k is hard because we should consider marginal distribution of each predictor, also the joint distribution

but by assuming p covariates are independent in each class, we eliminate the need of association b/w predictors.

in reality its not independent. but it works. so lalalalalala

$$Pr(Y=k \mid X=x) = \frac{\pi_k \times f_{k1}(x_1) \times f_{k2}(x_2) \times f_{k3}(x_3) \times \cdots \times f_{kp}(x_p)}{\sum_{l=1}^{K} \pi_l \times f_{k1}(x_1) \times f_{k2}(x_2) \times \cdots \times f_{kp}(x_1)}$$

to estimate 1 D density function f_{kj} using training data $x_{1j}, x_{2j}, \cdots x_{nj}$, we have the options

- $X_j \mid Y = K \sim N(\mu_{jk}, \sigma^2_{jk})$ (if X_j is quantitative)
univariate normal distribution
like to QDA, but the covariance matrix is diagonal since the predictors are independent
- non parametric estimate for f_{kj} (if X_j is quantitative)
plot histograms with same width for each class
check which bin ~~the~~ it falls to, compare & choose the ~~class which has highest density for value~~ probability density of the b
OR kernel density for - smoothened histogram, then evaluate the curve at x_j to get the density
- count the proportion of training observations for (qualitative) j^{th} predictor corresponding to each class

~~Sometimes Naive Bayes has slightly less overall test rate~~
when p is larger or n is smaller, Naive Bayes will perform better than LDA or QDA.

Comparison of classification methods

Analytical comparison

- LDA is a special case of QDA
  (LDA is restricted version of QDA with $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k$
- Any classifier with a linear decision boundary is a special case of Naive Bayes.
  Surprisingly this means LDA is a special case of Naive Bayes
- Naive Bayes is a special case of LDA with Σ restricted to be a diagonal matrix with j^{th} diagonal element equal to σ_j^2
- Neither QDA nor Naive Bayes is a special case of other.
  Naive Bayes produces a more flexible fit which is additive.
  If the interactions among the predictors are important in discriminating b/w classes QDA is better.

$$\log\left(\frac{P_r(Y=k \mid X=x)}{P_r(Y=K \mid X=x)}\right) = a_k + b_{kj} x_j \quad \text{– LDA}$$

like logistic regression, assumes log odds of posterior prob is linear in x.

$$= a_k + \sum_{j=1}^{p} b_{kj} x_j + \sum_{j=1}^{p} \sum_{l=1}^{p} c_{kjl} x_j x_l \quad \text{– QDA}$$

assumes log odds of posterior prob is quadratic in x

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

takes the form of generalised additive model

- Choice of method depends on true distribution of predictors in each of K classes & the bias variance tradeoff (n, p numbers size)

- LDA will outperform logistic regression when the normality assumption holds true
- KNN - non parametric. will perform good when the decision boundary is highly non linear and n is very large and p is small
- when n is modest & p is not very small & decision boundary is non linear, QDA > KNN
- unlike logistic regression, KNN doesn't tell which predictors are important

Empirical comparison

refer textbook for scenarios & the graph of error rates

when the true decision boundaries are linear, LDA & logistic regression will perform well

moderately non linear - QDA & Naive Bayes

more complicated - non parametric approach such as KNN

but the level of smoothness must be chosen carefully

Generalized linear models

* Linear regression on bikeshare data
  - the output is not necessarily an integer
  - it's could be negative numbers
  - data is heteroscedastic. as the mean increases, variance also increases. thus not suitable for linear regression
  - $Y = \sum_{j=1}^{p} X_j \beta_j + \epsilon$

    here ϵ is assumed to be mean zero term, σ^2 variance as constant which is violation of assumptions of linear model

* Poisson regression

$$Pr(Y=k) = \frac{e^{-\lambda} \lambda^k}{k!} \quad \text{for } k = 1, 2, \ldots$$

Y takes on non negative integers

$\lambda = E(Y) = Var(Y)$ larger the mean - larger the variance

Poisson distn is typically used to model counts

if we model Y as poisson distn with $E(Y) = \lambda = 5$,

then $P(Y = 0) = \frac{e^{-5} 5^0}{0!} = 0.0067$

probability of finding no users (counts) at that hour/condn

we expect the mean $\lambda = E(Y)$ to vary

$$\log(\lambda(X_1, X_2 \cdots X_p)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

$$\lambda(X_1, X_2 \cdots X_p) = e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p} \quad -①$$

to estimate the coefficients $\beta_0, \beta_1 \cdots \beta_p$ use the max likelihood estn

$$l(\beta_0, \beta_1 \cdots \beta_p) = \prod_{i=1}^{n} \frac{e^{-\lambda(x_i)} \lambda(x_i)^{y_i}}{y_i!}$$ where λ - ①

important distinctions

- an increase in X_j by one unit is associated with a change in $E(Y) = \lambda$ by a factor of e^{β_j}
- unlike the assumptions on linear regression, here the mean & variance relationship are handled
- no negative predictions

[over dispersion - variance much higher than mean. inflates Z values
topic to be checked later]

* we assume Y follows

normal (Gaussian) distn	in linear regression
Bernoulli distn	logistic regression
Poisson distn	Poisson regression

these 3 distributions come from exponential family.
other dists in this are Gamma dist, negative binomial dist

we can perform regression by modeling the response Y as coming from the exponential family, the transform the mean so that it is a linear fn of predictors.
this regression approach is called Generalised Linear Model