MECG15603/MCCG15603
Statistical Learning
MEME19903/MECG11103/MCCG11103
Predictive Modelling
Topic 2b: Supervised Learning: Logistic Regression & NN

Dr Liew How Hui

May 2024

Class Arrangement

- Week 9: Lecture 1:30pm-3:30pm (Logistic Regression). Practical 3:30-4:30pm
- Week 10: Lecture 1:30pm-3:30pm. Practical 3:30-4:30pm
- Week 11: Lecture 1:30pm-3:30pm. Practical 3:30-4:30pm
- Week 12: Lecture 1:30pm-3:30pm. Practical 3:30-4:30pm

Outline

1. Methods of Classification

2. Results Interpretation

3. Models Comparison
   - Compare to Multinomial Logistic Regression
   - Compare to Artificial Neural Network

4. Case Study

Methods of Classification

Problem	Output Y	Arrangement
regression	numerical/quantitative/ continuous	Week 5–Week 8
classification	categorical/qualitative/ discrete of K classes	Week 9–Week 12

Classification problems with $Y \in \{1, 2, \cdots, K\}$ can have a mathematical form

$$Y = (f(\mathbf{X}) + \epsilon \mod K) + 1.$$

Here, ϵ is a random variable generating integers 1 to K.

Methods of Classification (cont)

Problem	Prediction $\hat{Y}$	Performance Measurements
regression	$h(X)$, standard deviation	SSE, MSE, RMSE (root mean square error), R^2, ...
classification	$h(X)$, conditional probability	**contingency table**/ **confusion matrix**, accuracy, kappa, ...

Methods of Classification (cont)

Example 1: Let y_i be the actual observed output and $\hat{y}_i$ be the prediction from a predictive model h for the same inputs $\mathbf{x}_i$.

i	$\hat{y}_i$	y_i
1	A	B
2	B	B
3	A	B
4	A	A
5	B	B

Contingency table

		Observed/Actual	
		A	B
Prediction	A	1	2
	B	0	2

In R:

```
Yhat = c("A","B","A","A","B") # first column
Y    = c("B","B","B","A","B") # second column
table(Yhat, Y)  # Construct contingency table
```

Methods of Classification (cont)

The following supervised learning models for classification problems will be explored:

- Logistic regression models from statistics (Week 9)
- Naive Bayes models (Week 10)
- Tree-based models (Week 11)
- kNN models (Week 12)

They will come out in final exam's Question 4.
They will be applied in the assignment.

Logistic Regression

The *Logistic Regression (LR)* model is a special case of the generalised linear model (GLM) mentioned in Week 7. It is used for **binary classification** and has the form:

$$\begin{aligned} &\ln \frac{\pi}{1-\pi} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \\ &\text{where } \pi = P(Y = 1 | X_1 = x_1, \cdots, X_p = x_p) = \mathbb{E}[Y] \end{aligned} \tag{1}$$

Reference: Wikipedia:Bernoulli Distribution.

The assumption of LR is "the binary data are linearly separable with suitable parameters". Based on this assumption, a test input **x** would get a probability measure.

Logistic Regression (cont)

Rearranging (1) leads to

$$\begin{aligned}&\mathbb{P}(Y=1|X_1=x_1,\cdots,X_p=x_p)\\=&\frac{1}{1+\exp(-(\beta_0+\beta_1x_1+\cdots+\beta_px_p))}\\=&S(\beta_0+\beta_1x_1+\cdots+\beta_px_p)\end{aligned}\tag{2}$$

where the logistic/sigmoid function $S(x)=\frac{1}{1+e^{-x}}=\frac{e^x}{1+e^x}$ has the range $(0,1)$ for $-\infty<x<\infty$.

Using linear algebra, (2) can be expressed in vector form:

$$\mathbb{P}(Y=1|\mathbf{X}=\mathbf{x})=S(\beta^T\tilde{\mathbf{x}})$$

where $\boldsymbol{\beta}=(\beta_0,\cdots,\beta_p)$ and $\widetilde{\mathbf{x}}_j=(1,\mathbf{x}_j)$.

Logistic Regression (cont)

Given an input $\mathbf{x}$, the LR algorithm provides a prediction as follows based on the conditional probability (assuming the cut-off is 0.5):

$$h(\mathbf{x})=\begin{cases}0, & \mathbb{P}(Y=1|X=\mathbf{x})<0.5\\1, & \mathbb{P}(Y=1|X=\mathbf{x})\geq 0.5\end{cases}$$

or based the log-odds (or logit or ‘link’):

$$h(\mathbf{x})=\begin{cases}0, & \beta_0+\beta_1x_1+\cdots+\beta_px_p<0\\1, & \beta_0+\beta_1x_1+\cdots+\beta_px_p\geq 0\end{cases}$$

Logistic Regression (cont)

The coefficients β_i are estimated using MLE: Given data $(\mathbf{x}_i, y_i)$, $i = 1, \cdots, n$, we want find the coefficients β_i so that the **likelihood function** of $\beta_0, \cdots, \beta_p$ is maximised:

$$\begin{aligned} & L(\beta_0, \cdots, \beta_p;\ y_1, \cdots, y_n | \mathbf{x}_1, \cdots, \mathbf{x}_n) \\ = & \prod_{i=1}^{n} \mathbb{P}(Y = y_i | \mathbf{X} = \mathbf{x}_i) \end{aligned} \tag{3}$$

Y is binary and follows a **Bernoulli distribution**.

Logistic Regression (cont)

According to `https://en.wikipedia.org/wiki/Bernoulli_distribution`, $Y \sim \mathit{Bernoulli}(\pi_{\mathbf{x}} = \mathbb{P}(Y = 1 | \mathbf{X} = \mathbf{x}))$, then the probability mass function of observing $y \in \{0, 1\}$ is

$$\mathbb{P}(y) = (\pi_{\mathbf{x}})^y (1 - \pi_{\mathbf{x}})^{1-y}.$$

The likelihood for the observation $(\mathbf{x}_i, y_i)$ is

$$\mathbb{P}(Y = y_i | \mathbf{X} = \mathbf{x}_i) = \left(\frac{e^{\tilde{\mathbf{x}}_i^T \boldsymbol{\beta}}}{1 + e^{\tilde{\mathbf{x}}_i^T \boldsymbol{\beta}}} \right)^{y_i} \left(1 - \frac{e^{\tilde{\mathbf{x}}_i^T \boldsymbol{\beta}}}{1 + e^{\tilde{\mathbf{x}}_i^T \boldsymbol{\beta}}} \right)^{1-y_i}$$

Logistic Regression (cont)

$$= e^{y_i \widetilde{\mathbf{x}}_i^T \boldsymbol{\beta}} \cdot (1 + e^{\widetilde{\mathbf{x}}_i^T \boldsymbol{\beta}})^{-y_i} \cdot (1 + e^{\widetilde{\mathbf{x}}_i^T \boldsymbol{\beta}})^{-(1-y_i)}$$

where $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_p)$ and $\widetilde{\mathbf{x}}_i = (1, \mathbf{x}_i)$.
Substituting it into (3), we have

$$\begin{aligned} & L(\beta_0, \cdots, \beta_p;\ y_1, \cdots, y_n | \mathbf{x}_1, \cdots, \mathbf{x}_n) \\ = & \prod_{i=1}^{n} (e^{y_i \widetilde{\mathbf{x}}_i^T \boldsymbol{\beta}})(1 + e^{\widetilde{\mathbf{x}}_i^T \boldsymbol{\beta}})^{-1}. \end{aligned}$$

Taking natural log leads to log-likelihood:

$$\ln L = \sum_{i=1}^{n} y_i \widetilde{\mathbf{x}}_i^T \boldsymbol{\beta} - \sum_{i=1}^{n} \ln(1 + e^{\widetilde{\mathbf{x}}_i^T \boldsymbol{\beta}}). \tag{4}$$

Theory (cont)

By Calculus Theory,

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\, L = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \ln L \Rightarrow \frac{\partial}{\partial \boldsymbol{\beta}}(\ln L) = \mathbf{0}$$

i.e.

$$\frac{\partial}{\partial \boldsymbol{\beta}} \left(\sum_{i=1}^{n} y_i \widetilde{\mathbf{x}}_i^T \boldsymbol{\beta} - \sum_{i=1}^{n} \ln(1 + e^{\widetilde{\mathbf{x}}_i^T \boldsymbol{\beta}}) \right) = \mathbf{0}.$$

leading to the nonlinear system:

$$\sum_{i=1}^{n} x_k^{(i)} \left[y_i - \frac{e^{\widetilde{\mathbf{x}}_i^T \boldsymbol{\beta}}}{1 + e^{\widetilde{\mathbf{x}}_i^T \boldsymbol{\beta}}} \right] = 0, \quad k = 0, 1, \cdots, p$$

where $x_0^{(i)}$ is defined to be 1.

Example 2

Consider the following data:

	balance, x_1	default, y
3904	973.9031	No
9146	667.4920	No
5278	1377.4621	No
6930	298.7196	No
7084	919.6724	No
8447	245.3465	No
3024	0.0000	No
8365	1013.2169	Yes
9922	1627.8983	Yes
5210	1711.1691	Yes

Let No=0, Yes=1. The mathematical model for this problem (2) becomes

$$P(Y=1|X_1=x_1)=\frac{1}{1+exp(-(\beta_0+\beta_1 x_1))}$$

Example 2 (cont)

To estimate the coefficients, we need to maximise the likelihood (4):

$$\ln L(\beta_0,\beta_1)=\sum_{i=1}^{10} y_i(\beta_0+\beta_1 x_{1,i})-\sum_{i=1}^{10}\ln(1+e^{\beta_0+\beta_1 x_{1,i}}).$$

i	$x_{1,i}$	y	y_i	$y_i(\beta_0+\beta_1 x_{1,i})$	$\ln(1+e^{\beta_0+\beta_1 x_{1,i}})$
1	973.9031	No	0	0	$\ln(1+\exp(\beta_0+973.9031\beta_1))$
2	667.4920	No	0	0	$\ln(1+\exp(\beta_0+667.4920\beta_1))$
3	1377.4621	No	0	0	$\ln(1+\exp(\beta_0+1377.4621\beta_1))$
4	298.7196	No	0	0	$\ln(1+\exp(\beta_0+298.7196\beta_1))$
5	919.6724	No	0	0	$\ln(1+\exp(\beta_0+919.6724\beta_1))$
6	245.3465	No	0	0	$\ln(1+\exp(\beta_0+245.3465\beta_1))$
7	0.0000	No	0	0	$\ln(1+\exp(\beta_0+0.0000\beta_1))$
8	1013.2169	Yes	1	$\beta_0+1013.2169\beta_1$	$\ln(1+\exp(\beta_0+1013.2169\beta_1))$
9	1627.8983	Yes	1	$\beta_0+1627.8983\beta_1$	$\ln(1+\exp(\beta_0+1627.8983\beta_1))$
10	1711.1691	Yes	1	$\beta_0+1711.1691\beta_1$	$\ln(1+\exp(\beta_0+1711.1691\beta_1))$
				$\sum_{i=1}^{10} y_i(\beta_0+\beta_1 x_{1,i})$	$\sum_{i=1}^{10}\ln(1+e^{\beta_0+\beta_1 x_{1,i}})$

Example 2 (cont)

We want to find β_0 and β_1 which maximises $\ln L$.
In optimisation theory, this is the same as minimising $-\ln L$.
Using an initial guess $\beta_0 = 1$ and $\beta_1 = 0$, we are able to find the estimate below using conjugate gradient method:

$$\beta_0 = -5.949928639, \quad \beta_1 = 0.004692262$$

This is very close the estimate obtained using R's glm():

$$\beta_0 = -6.092158330, \quad \beta_1 = 0.004784055$$

The discrepencies are due to numerical formulation.

Theory (cont)

So far, the theory assumes all inputs to be numeric so that we can evaluate

$$\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p.$$

When the inputs x_i are categorical, we need to introduce 'dummy variables' to convert each categorical data to numeric data — **One-Hot Encoding for Categorical Data**.

Example 3: The categorical data (left) to 'dummy varibles' (right):

gender	bloodtype
0	A
1	B
1	AB
0	B
0	O

gender1	bloodtypeAB	bloodtypeB	bloodtypeO
0	0	0	0
1	0	1	0
1	1	0	0
0	0	1	0
0	0	0	1

Outline

1. Methods of Classification

2. Results Interpretation

3. Models Comparison
   - Compare to Multinomial Logistic Regression
   - Compare to Artificial Neural Network

4. Case Study

Results Interpretation

After we obtain the estimate of the coefficients from the likelihoood function:

$$\frac{\partial}{\partial\boldsymbol{\beta}}(\ln L) = \mathbf{0} \Rightarrow \hat{\boldsymbol{\beta}} = \operatorname*{argmax}_{\boldsymbol{\beta}} L,$$

The computed results may not fit the data well and we need to answer the following questions.

1. Does the model explain the data?
2. How does each individual predictor influence the response?

Results Interpretation (cont)

(1) Does the model explain the data?

The statistician's answer, reflected in glm()'s output is to compare

- Null deviance = 2(LL(**Saturated Model**) - LL(**Null Model**)) on `df = df_Sat - df_Null`
- Residual deviance = 2(LL(**Saturated Model**) - LL(**Proposed Model**)) on `df = df_Sat - df_Proposed`

The **Saturated Model** is a model that assumes each data point has its own parameters (which means we have n parameters to estimate.)

Results Interpretation (cont)

The **Null Model** assumes the exact "opposite", in that is assumes one parameter for all of the data points, which means we only estimate 1 parameter.

The **Proposed Model** assumes we can explain the data points with p parameters + an intercept term, so we have $p + 1$ parameters.

If the Null Deviance is really small, it means that the Null Model explains the data pretty well. Likewise for the Residual Deviance. Usually, when null Deviance is much larger than residual deviance, the linear model may explain the data. For prediction purposes, we use the contingency table instead.

Results Interpretation (cont)

(2) How does each individual predictor influence the response?

To answer the question, we analyse the influence of individual predictor to the response using the hypothesis:

$$H_0 : \beta_i = 0 \quad \text{vs} \quad H_1 : \beta_i \neq 0.$$

The Z-statistic of β_i characterises the above hypothesis:

$$Z = \frac{\hat{\beta}_i - 0}{SE(\hat{\beta}_i)}$$

Results Interpretation (cont)

The **square error** in the Z-statistic:

$$SE(\hat{\beta}_i) = [[\mathcal{I}(\beta)]^{-1}]_{(i+1),(i+1)}$$

is the square root of the $(i+1)$-th diagonal element of the inverse matrix of the $(p+1) \times (p+1)$ **information matrix**:

$$\mathcal{I}(\boldsymbol{\beta}) = \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\left(\sum_{i=1}^n y_i \widetilde{\mathbf{x}}_i^T \boldsymbol{\beta} - \sum_{i=1}^n \ln(1 + e^{\widetilde{\mathbf{x}}_i^T \boldsymbol{\beta}})\right) = \sum_{i=1}^n \sigma_i^2 \mathbf{x}_i \mathbf{x}_i^T$$

where $\sigma_i^2 = S(\mathbf{x}_i^T \beta) \cdot (1 - S(\mathbf{x}_i^T \beta))$;

Results Interpretation (cont)

When the number of samples "n" is large, the Z-statistic approaches the normal distribution

$$\frac{\hat{\beta}_i - 0}{SE(\hat{\beta}_i)} \sim Normal(0,1),$$

according to `https://en.wikipedia.org/wiki/Wald_test`.

A $(1-\frac{\alpha}{2}) \times 100\%$ confidence interval for β_i, $i = 1, \cdots, p$, can be estimated as

$$\hat{\beta}_i \pm Z_{1-\alpha/2} SE(\hat{\beta}_i).$$

A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter. In this case, $\alpha = 0.05$ and $Z_{1-\alpha/2} \approx 1.96$, therefore, the 95% confidence interval for β_i takes the form

$$[\hat{\beta}_i - 1.96 \cdot SE(\hat{\beta}_i),\ \hat{\beta}_i + 1.96 \cdot SE(\hat{\beta}_i)]. \tag{5}$$

Results Interpretation (cont)

The interception β_0 is typically not of interest and it only for fitting data to the model.

For β_i where $i = 1, 2, ..., p$, we have the analysis:

- When Z-statistic is large, *p-value* is small.
  ⇒ null hypothesis should be rejected (when p-value is less than some significance level, e.g. α=5%).
  ⇒ X is associated with Y and is a significant predictor.
- When Z-statistic is small, *p-value* is large.
  ⇒ null hypothesis should not be rejected (when (when p-value $> \alpha = 0.05$).
  ⇒ X and Y is most likely not related and X is probably an unimportant predictor to Y.

Results Interpretation (cont)

As mentioned in Week 7, a logistic regression model is a special case of GLM where the link function is logit. In R, this is specified using the option ‘family=binomial’:

```
lr.fit = glm(Y ~ ., data=D, family=binomial)
```

Here `binomial` uses `logit` link (for logistic CDF) by default. Other link options for `binomial` are ‘probit’, ‘cauchit’, (corresponding to normal and Cauchy CDFs respectively) ‘log’ and ‘cloglog’ (complementary log-log).

Example 4:

```
library(ISLR2)
lr.fit = glm(default ~ balance, data=Default, family=binomial)
print(summary(lr.fit))
```

Results Interpretation (cont)

Example 4: (cont)

```
Call:
glm(formula = default ~ balance, family = binomial, data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2697  -0.1465  -0.0589  -0.0221   3.7589

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.065e+01  3.612e-01  -29.49   <2e-16 ***
balance      5.499e-03  2.204e-04   24.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1596.5  on 9998  degrees of freedom
AIC: 1600.5

Number of Fisher Scoring iterations: 8
```

Results Interpretation (cont)

Example 4: (cont)
(a) Write down the mathematical formula of the logistic regression model.

Solution

$$\mathbb{P}(Y=1|X)=\frac{1}{1+\exp(-(-10.65+0.0055\,\texttt{balance}))}$$

(b) Predict the default probability for an individual with a balance of (i) \$1000, (ii) \$2000.
Exercise.

Results Interpretation (cont)

One reason for the popularity of LR in practice is due to the interpretability of β_i using the notion `https://en.wikipedia.org/wiki/Odds_ratio`. The **odds ratio** (OR) is the ratio between two odds:

$$\text{OR}=\frac{\frac{\mathbb{P}(Y=1|X_i=b)}{\mathbb{P}(Y=0|X_i=b)}}{\frac{\mathbb{P}(Y=1|X_i=a)}{\mathbb{P}(Y=0|X_i=a)}}=\frac{\exp(\cdots+\beta_i\cdot b+\cdots)}{\exp(\cdots+\beta_i\cdot a+\cdots)}=\exp(\beta_i(b-a)).$$

The odds (in the OR) are the ratio of the probabilities of two complementing events:

$$\frac{\mathbb{P}(Y=1|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=0|\mathbf{X}=\mathbf{x})}=\frac{\mathbb{P}(Y=1|\mathbf{X}=\mathbf{x})}{1-\mathbb{P}(Y=1|\mathbf{X}=\mathbf{x})}=\exp(\tilde{\mathbf{x}}^T\beta). \qquad (6)$$

Results Interpretation (cont)

Taking the logarithm of both sides of (6), we obtain (1):

$$\ln \frac{\mathbb{P}(Y=1|\mathbf{X}=\mathbf{x})}{1-\mathbb{P}(Y=1|\mathbf{X}=\mathbf{x})} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

The LHS is called the *log-odds* or *logit*, which is linear in X.

For a 1 unit increment in X_i leads to

$$\begin{aligned}\beta_i > 0 \Rightarrow \mathit{logit} > 0 \Rightarrow \mathit{OR} > 1 \Rightarrow \mathit{odds}(X_i+1) > \mathit{odds}(X_i) \Rightarrow \\ \mathbb{P}(Y=1|X_i+1) > \mathbb{P}(Y=1|X_i) \text{ (higher prob for } X_i+1) \\ \beta_i < 0 \Rightarrow \mathit{logit} < 0 \Rightarrow \mathit{OR} < 1 \Rightarrow \mathit{odds}(X_i+1) < \mathit{odds}(X_i) \Rightarrow \\ \mathbb{P}(Y=1|X_i+1) < \mathbb{P}(Y=1|X_i) \text{ (lower prob for } X_i+1)\end{aligned}$$

Qualitative Predictors

So far the predictors are all assumed numeric. When a predictor (or factor) is **qualitative**, we need to introduce **dummy variable(s)**: For example, the predictor "gender" has two levels 0 (male) and 1 (female), a new variable below is created

$$\text{gender1} = \begin{cases} 1, & \text{if gender} = 1 \\ 0, & \text{if gender} = 0 \end{cases}$$

Therefore, the logistic model is

$$\begin{aligned}&\mathbb{P}(Y=1|\mathbf{X}=\mathbf{x}) \\ =&\frac{1}{1+\exp(-(\beta_0+\cdots+\beta_i\text{gender1}+\cdots))}\end{aligned}$$

Results Interpretation (cont)

The linear algebra theory associated with qualitative predictors are more complex but the result interpretation of the qualitative predictors is also related to the odds ratio, but now, of the the dummy variable(s), for example, "gender1":

$$\text{OR} = \frac{\frac{\mathbb{P}(Y=1|\text{gender}=1)}{\mathbb{P}(Y=0|\text{gender}=1)}}{\frac{\mathbb{P}(Y=1|\text{gender}=0)}{\mathbb{P}(Y=0|\text{gender}=0)}} = \frac{\exp(\cdots + \beta_i + \cdots)}{\exp(\cdots + 0 + \cdots)} = \exp(\beta_i)$$

Note that 0=male, 1=female, we have

β_i	OR	Relative probability of $\mathbb{P}(Y = 1\vert\text{gender} = 1)$	Probability to be classified into Class 1
Positive	> 1	Higher	female $>$ male
Negative	< 1	Lower	male $>$ female

Results Interpretation (cont)

Example 5:
Consider the ISLR2's **Default** data. Use R to work on the influence of the `student` predictor on the output `default`.

Solution

The R script to fit the logistic model is listed below.

```
library(ISLR2)
lr.fit = glm(default ~ student, data=Default,
  family=binomial)
print(summary(lr.fit))
```

Results Interpretation (cont)

Example 5: (cont)

```
Call:
glm(formula = default ~ student, family = binomial, data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.2970  -0.2970  -0.2434  -0.2434   2.6585

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.50413    0.07071  -49.55  < 2e-16 ***
studentYes   0.40489    0.11502    3.52 0.000431 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 2908.7  on 9998  degrees of freedom
AIC: 2912.7

Number of Fisher Scoring iterations: 6
```

Results Interpretation (cont)

Example 5: (cont)

Use the analysis results from R to answer the following questions.

(a) Find the odds ratio of default for a student with a non-student. Explain.

(b) Predict the probability of default for (i) student (ii) non-student.

Hint: (i) $\mathbb{P}(Y = 1|student = Yes)$; (ii) $\mathbb{P}(Y = 1|student = No)$

Classroom discussion.

Results Interpretation (cont)

When a qualitative predictor X_i has $K > 2$ levels, $(K-1)$ **dummy variables** X_i.level2, $\cdots$, X_i.level$K$ are introduced to the logistic regression model

$$\mathbb{P}(Y=1|\mathbf{X}) = \frac{1}{1+\exp(-(\beta_0+\cdots+\beta_i^{(2)}x_i.\text{level2}+\cdots+\beta_i^{(K)}x_i.\text{level}K+\cdots))}$$

where

$$x_i.\text{level}k = \begin{cases} 1, & x_i = \text{level } k, \\ 0, & \text{otherwise}, \end{cases} \quad k = 2,\cdots,K.$$

The introduction of $K-1$ dummy variables is called the *"nearly" one-hot encoding*, where the reference variable is implicit. In a **one-hot encoding** all dummy variables are kept.

Outline

1. Methods of Classification
2. Results Interpretation
3. Models Comparison
   - Compare to Multinomial Logistic Regression
   - Compare to Artificial Neural Network
4. Case Study

Models Comparison

Unlike the multiple linear regression (OLS) which has the F-statistic to compare (by contrasting) how well models match the data, The GLM, in particular, the logistic regression model only has AIC (C_p, BIC, etc.) for matching model and data.

In the practical, we are going to do manual subsets selection rather than using the `regsubsets` from the leaps library.

Models Comparison (cont)

As mentioned in Week 7, to compare two trained GLM (LR in particular), we need to look at the deviance (or the closely related AIC, C_p, BIC, etc.).

This is accomplished using `anova(small.m, large.m)`, where `small.m` is a model with less features, `large.m` is a model with more features. If the deviances' likelihood ratio test (LRT, same as χ^2 test and Rao test) has a p-value less than 0.05, we can conclude that `large.m` is much better at capturing the data than the `small.m`.

Models Comparison to Multinomial LR

A general K-level qualitative response cannot be handled by the LR model. `https://en.wikipedia.org/wiki/Multinomial_logistic_regression` (or Softmax regression) is a generalisation of the LR model:

$$\begin{cases} \ln \dfrac{\mathbb{P}(Y=2|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=1|\mathbf{X}=\mathbf{x})} = \boldsymbol{\beta}_2 \cdot \mathbf{x} \\ \ln \dfrac{\mathbb{P}(Y=3|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=1|\mathbf{X}=\mathbf{x})} = \boldsymbol{\beta}_3 \cdot \mathbf{x} \\ \qquad\qquad\qquad \cdots\cdots \\ \ln \dfrac{\mathbb{P}(Y=K|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=1|\mathbf{X}=\mathbf{x})} = \boldsymbol{\beta}_K \cdot \mathbf{x} \end{cases}$$

Models Comparison (cont)

After some algebra, we have

$$\begin{aligned} \mathbb{P}(Y=1|\mathbf{X}=\mathbf{x}) &= \frac{1}{1+\sum_{i=2}^{K} e^{\boldsymbol{\beta}_i\cdot\mathbf{x}}} \\ \mathbb{P}(Y=j|\mathbf{X}=\mathbf{x}) &= \frac{e^{\boldsymbol{\beta}_j\cdot\mathbf{x}}}{1+\sum_{i=2}^{K} e^{\boldsymbol{\beta}_i\cdot\mathbf{x}}}, \quad j=2,\cdots,K. \end{aligned} \tag{7}$$

This model requires more data than LR, so when we have little data, this model won't work.

Models Comparison (cont)

An implementation of Multinomial LR is available in the `nnet` package:

```
multinom(formula, data, weights, subset, na.action,
         contrasts = NULL, Hess = FALSE, summ = 0,
         censored = FALSE, model = FALSE, ...)
```

When $K = 2$, the multinomial LR is just the usually logistic regression model and we will explore this in the practical.

Models Comparison (cont)

In Python, the “Logistic Regression” is actually a generalisation to the **elastic net** instead of the LR we discussed:

```
class sklearn.linear_model.LogisticRegression(penalty=’l2’, *,
  dual=False, tol=0.0001, C=1.0, fit_intercept=True,
  intercept_scaling=1, class_weight=None, random_state=None,
  solver=’lbfgs’, max_iter=100, multi_class=’auto’, verbose=0,
  warm_start=False, n_jobs=None, l1_ratio=None)
```

When $C = \infty$, it approaches the LR. The LR and multinomial LR are properly implemented in Python as `Logit` and `MNLogit` in `statsmodels.discrete.discrete_model`.

Models Comparison (cont)

Feed-forward Artificial Neural Networks (ANN) or multi-layer perceptrons (MLP), "include" LR and multinomial LR as special cases.

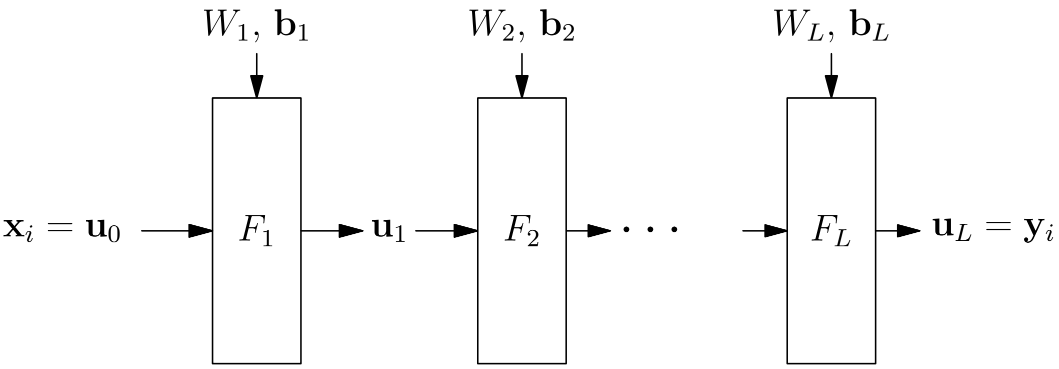
A multi-layer feed-forward ANN with input $\mathbf{x}_i \in \mathbb{R}^p$ and output is $\mathbf{y}_i \in \mathbb{R}^m$:

$$\begin{aligned} \mathbf{u}_1 &= F_1(W_1\mathbf{u}_0 + \mathbf{b}_1), \quad \mathbf{u}_0 = \mathbf{x}_i \\ \mathbf{u}_2 &= F_2(W_2\mathbf{u}_1 + \mathbf{b}_2) \\ &\cdots \\ \hat{\mathbf{y}}_i = \mathbf{u}_L &= F_L(W_L\mathbf{u}_{L-1} + \mathbf{b}_L). \end{aligned} \tag{8}$$

where L is the number of layers of ANN (with $L-1$ hidden layers).

Models Comparison (cont)

Horizontal pictorial representation:



Models Comparison (cont)

The algorithm to estimate the parameters W_ℓ and $\mathbf{b}_\ell$ for the layer $\ell = 1, \ldots, L$ is the improvement of back-propagation algorithm:

1. $t = 0$;
2. Using the guess parameters $W_\ell^{(t)}$, $\mathbf{b}_\ell^{(t)}$, calculate all the intermediate states
$$\mathbf{u}_\ell^{(t)} = F_\ell(W_\ell^{(t)}\mathbf{u}_{\ell-1}^{(t)} + \mathbf{b}_\ell^{(t)})$$
and the output $\hat{\mathbf{y}}_i$;

Models Comparison (cont)

3. The output layer
$$\delta_L = \hat{\mathbf{y}}_i - \mathbf{y}_i$$
4. Back-Propagation (roughly): For ℓ from L to 1, do
$$\begin{aligned}\delta_{\ell-1} &= \frac{\partial F_\ell}{\partial W_\ell}(\mathbf{u}_{\ell-1}^{(t)})\delta_\ell \\ W_\ell^{(t+1)} &= W_\ell^{(t)} + \alpha \times \mathbf{u}_{\ell-1}^{(t)} \times \delta_{\ell-1}\end{aligned}$$
5. $t = t + 1$ and go to step 2.

Models Comparison (cont)

When $L = 1$, we obtain a `https://en.wikipedia.org/wiki/Perceptron`:

$$\mathbf{y} = \mathbf{u}_1 = F_1(W_1\mathbf{x}_i + \mathbf{b}_1). \tag{9}$$

We can see that when $m = 1$, $F_1(x) = S(x)$, we obtain the LR. When $m = K - 1$ ($K \geq 2$), we obtain the multinomial LR (which is how `nnet::multinom` was implemented).

Models Comparison (cont)

When $L = 2$, we obtain an ANN with a single hidden-layer.

$$\begin{aligned} \mathbf{u}_1 &= F_1(W_1\mathbf{x}_i + \mathbf{b}_1) \\ \mathbf{y} = \mathbf{u}_2 &= F_1(W_2\mathbf{u}_1 + \mathbf{b}_2). \end{aligned} \tag{10}$$

This is implemented in R's `nnet` package as

```
nnet(x, y, weights, size, Wts, mask,
     linout = FALSE, entropy = FALSE, softmax = FALSE,
     censored = FALSE, skip = FALSE, rang = 0.7, decay = 0,
     maxit = 100, Hess = FALSE, trace = TRUE, MaxNWts = 1000,
     abstol = 1.0e-4, reltol = 1.0e-8, ...)
```

Outline

1. Methods of Classification
2. Results Interpretation
3. Models Comparison
   - Compare to Multinomial Logistic Regression
   - Compare to Artificial Neural Network
4. Case Study

Case Study 1: Simple Model Comparison

Example 6: Given the info of a fitted model below.

```
Call: glm(formula=default~balance+income+student, family=binomial,
          data=Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4691  -0.1418  -0.0557  -0.0203   3.7383

Coefficients: Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1571.5  on 9996  degrees of freedom
AIC: 1579.5
Number of Fisher Scoring iterations: 8
```

Case Study 1 (cont)

Discuss the results involving the coefficients, odds and significance of each variable.

Solution

Coefficients: $\beta_0 = -10.8690$, $\beta_1 = 0.0057$, $\beta_2 = 3.033 \times 10^{-6}$, $\beta_3 = -0.6468$.

Significance: Based on the *p*-value, we find that `balance` and `student` are significant while `income` is probably insignificant (according to the default $\alpha = 0.05$).

Odds: The odds of the default increases with the balance and income but students has a lower odds compare to non-students.

Case Study 1 (cont)

Performing an ANOVA from NULL model to full model using χ^2-test, we obtain

```
Analysis of Deviance Table

Model 1: default ~ 1
Model 2: default ~ student
Model 3: default ~ student + balance
Model 4: default ~ student + balance + income
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      9999     2920.7
2      9998     2908.7  1    11.97 0.0005416 ***
3      9997     1571.7  1  1337.00 < 2.2e-16 ***
4      9996     1571.5  1     0.14 0.7115139
```

“Model 3” is the best model.

Case Study 2

Example 7: Given the following results from the analysis of credit card applications approval dataset using logistic regression model.

```
glm(formula=Approved˜., family=binomial, data=d.f.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6796  -0.5477   0.2681   0.3316   2.4501

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    3.1379649  0.5744168   5.463 4.68e-08 ***
Maleb         -0.1758676  0.3229541  -0.545   0.5861
Age            0.0001318  0.0142338   0.009   0.9926
Debt           0.0042129  0.0298740   0.141   0.8879
YearsEmployed -0.1023132  0.0582368  -1.757   0.0789 .
PriorDefaultt -3.6614227  0.3659226 -10.006  < 2e-16 ***
Employedt     -0.2500687  0.4013495  -0.623   0.5332
CreditScore   -0.1098142  0.0644360  -1.704   0.0883 .
ZipCode        0.0011958  0.0009540   1.253   0.2100
Income        -0.0004544  0.0001966  -2.311   0.0209 *
---
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 625.90  on 454  degrees of freedom
Residual deviance: 294.33  on 445  degrees of freedom
  (27 observations deleted due to missingness)
AIC: 314.33
```

Case Study 2 (cont)

Example 7: (cont)
where the output `Approved` is either positive (represented as 0) and negative (represented as 1) and the features

- `Male` is categorical with `a`=Female, `b`=Male;
- `PriorDefault` is categorical with `f`=false, `t`=true;
- `Employed` is categorical with `f`=false, `t`=true;
- `Age`, `Debt`, `YearsEmployed`, `CreditScore`, `ZipCode`, `Income` are continuous variables.

Case Study 2 (cont)

(i) Write down the mathematical expression of the logistic model for the given data with the coefficient values rounded to 4 decimal places.

(i) Solution

The logistic model is

$$\mathbb{P}(\texttt{Approved} = 1|\mathbf{X}) = \frac{1}{1 + e^{-(3.1380 + \mathbf{w}^T\mathbf{X})}}$$

$$\begin{aligned}\mathbf{w}^T\mathbf{X} = &-0.1759\,\texttt{Male} + 0.0001\,\texttt{Age} + 0.0042\,\texttt{Debt} - 0.1023\,\texttt{YearsEmployed}\\ &- 3.6614\,\texttt{PriorDefault} - 0.2501\,\texttt{Employed} - 0.1098\,\texttt{CreditScore}\\ &+ 0.0012\,\texttt{ZipCode} - 0.0005\,\texttt{Income}\end{aligned}$$

Case Study 2 (cont)

(ii) By calculating the probability of the credit card application being approved for a male of age 22.08 with a debt of 0.83 unit who has been employed for 2.165 years with no prior default and is currently unemployed, has a credit score 0 and a zip code 128 with income 0, find the **probability** of credit card applications approval and determine if the approval is positive or negative (using the cut-off of 0.5).

Case Study 2 (cont)

(ii) Solution

First, we calculate

$$\begin{aligned}\mathbf{w}^T\mathbf{X} &= -0.1759\,(1) + 0.0001\,(22.08) + 0.0042\,(0.83) - 0.1023\,(2.165) \\ &\quad - 3.6614\,(0) - 0.2501\,(0) - 0.1098\,(0) \\ &\quad + 0.0012\,(128) - 0.0005\,(0) \\ &= -0.2380855\end{aligned}$$

The probability of the credit card application being ‘negatively’ approved,

$$\mathbb{P}(\texttt{Approved} = 1|\mathbf{X}) = \frac{1}{1+\exp(-(3.1380 - 0.2380855))} = 0.9478$$

Since the probability is more than 0.5, the approval is **negative**.

Case Study 2 (cont)

(ii) Advice for Final Exam

If you are not able to write the formula $\mathbf{w}^T\mathbf{X}$ in the previous correctly, try to for a table based on the given information:

Feature, X_i	Value, x_i	$\beta_i x_i$ (one-hot)
Male	male=b	-0.1758676×1
Age	22.08	0.0001318×22.08
Debt	0.83	0.0042129×0.83
YearsEmployed	2.165	-0.1023132×2.165
PriorDefault	no=f	$0 = -3.6614227 \times 0$
Employed	unemployed=f	$0 = -0.2500687 \times 0$
CreditScore	0	0
ZipCode	128	0.0011958×128
Income	0	0
	Sum:	-0.237906427
	$\beta_0 + \mathbf{w}^T\mathbf{X}$	$3.1379649 - 0.237906427 = 2.900058473$

Case Study 2 (cont)

(iii) Calculate the odds ratio for the approval being negative with the prior default to be true against the prior default to be false. Infer the likelihood of getting a negative approval based on the prior default.

(iii) Solution

The odds ratio for the approval with respect to prior default is

$$\frac{\frac{\mathbb{P}(\texttt{Approved}=1|\texttt{PriorDefault}=t)}{1-\mathbb{P}(\texttt{Approved}=1|\texttt{PriorDefault}=t)}}{\frac{\mathbb{P}(\texttt{Approved}=1|\texttt{PriorDefault}=f)}{1-\mathbb{P}(\texttt{Approved}=1|\texttt{PriorDefault}=f)}} = \frac{\exp(-3.6614227\times 1)}{\exp(-3.6614227\times 0)} = 0.02569593$$

Someone with a prior default has a lower likelihood to get a negative approval compare to someone without a prior default.

Case Study 3

Example 9:
(a) The human resource department would like to determine potential employees for promotion. You have collected some data from previous employee promoting records as described below:

exp	Number of years of experience working in the company
sal_mth	Average monthly salary in last 12 months
sal_yr	Yearly salary in last 12 months
pjt	Is there any project involved? [Yes; No]
dpmt	Department [A; B; C; D]
emp_id	Employee ID
promote	Is the employee getting promoted? [Yes=1; No=0]

Case Study 3 (cont)

A logistic regression has been constructed to predict the promotion of an employee. Table Q2(a) shows parts of the results of the logistic regression.

	Coefficient	P-value
Intercept	0.0035	$<$ 2e-16
exp_yr	0.7124	$<$ 2e-16
sal_mth	-0.0212	0.0057
sal_yr	-0.0363	0.0086
pjt_Yes	0.0330	0.2479
dpmt_B	1.0447	0.0002
dpmt_C	-1.5318	6.87e-05
dpmt_D	2.1539	0.0017
emp_id	-0.0279	0.5245

Table Q2(a)

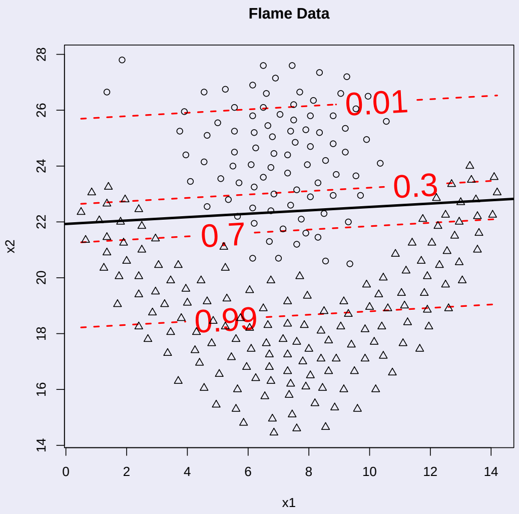
Case Study 3 (cont)

1. Write the logistic regression model that compute the probability that an employee get promoted, $\mathbb{P}(Y = 1)$.
2. Calculate the odds and compare the probability of promotion for employee with 7 years of working experience and an employee with 2 years of working experience.
3. Calculate the odds and compare the probability of promotion for employee in different departments. Arrange the probability of promotion of department from lowest to highest.

Case Study 4

ROC Example

For the "flame" data, the "boundary" of the classifier is shown in the left figure below as the solid line:



Case Study 4 (cont)

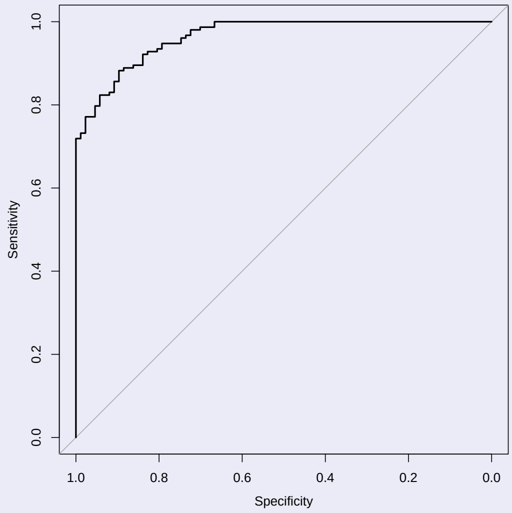
ROC Example continue

The dashed lines correspond to different "cut-off" 0.01, 0.3, 0.7 and 0.99. The ROC curve can be understood as the result of varying the "cut-off" and calculating the "sensitivity" (TPR) and "specificity" mentioned in Topic 1. If we calculate out, we have

	0.01		0.3		0.7		0.99	
Predicted	1	2	1	2	1	2	1	2
1	19	0	64	6	79	23	87	80
2	68	153	23	147	8	130	0	73
	TPR = 0.2184	FPR = 0	0.7356	0.0392	0.9080	0.1503	1	0.5229

Case Study 4 (cont)

ROC Example continue



Practical for LR

`prac_cls1.R`

Start reading the assignment and exploring the data in the assignment based on what you have learned from the Week 1 to Week 9 practicals.