Advanced Topics in Statistics

~ Classification - Logistic regression ~

2

UNIVERSITY OF EXETER

Introduction.

Classification is a method for predicting qualitative responses. E.g. deciding whether a transaction fraudulent or not, or assigning a diagnosis to a patient based on the observed symptoms.

We call these qualitative variables 'categorical'.

Predicting a qualitative response for an observation can be referred to as classifying that observation, since it involves assigning it to a category or class.

Classification methods often first predict the probability of belonging to a class (and assign a category based on these probabilities). In this sense they also behave like regression methods.

We usually have a set of training data that we use to train the classifier. But we not only want the method to perform well on the set of training data, but also (even more importantly) on the test observations that were not used to train the classifier.

We have seen that the error rate can be used to measure the classifier's performance. The error rate is the fraction of incorrect classifications.

Accuracy.

Another metric that is often used to evaluate the overall performance of the classifier is accuracy.

Accuracy is the proportion of correctly identified observations, thus

$$\text{accuracy} = 1 - \text{error rate}.$$

What we consider good performance however, depends on the problem is questions.

- For example, in medical diagnostics it can be more important to correctly identify those patients who have a certain condition (the 'positives') than to have a higher overall accuracy.
- Another problem with just considering the accuracy can arise when e.g. in a binary classification situation, one class has much less observations than the other class.

  In this case, a naive classifier that assigns the larger class to all the observations can have a high accuracy, even though we wouldn't say it has good performance.

Therefore, while accuracy can be a good metric in many situations, we should also consider other metrics, (e.g. the fraction of correctly identified observations separately for each class) to make sure our method is reliable.

Confusion matrix.

We can use the so-called confusion matrix to construct the metrics of interest.
Each row of the matrix represents the instances in a predicted class, while each column represents the instances in an actual class.

Assume the two classes are positives (1) and negatives (0).

The confusion matrix splits the outcome of the classification algorithm as follows:

		True label	
		1	0
Predicted label	1	a	b
	0	c	d

- a is the number of true positives
- b is number of false positives (also called type I error)
- c is the number of false negatives (also called type II error)
- d is the number of true negatives

Using the confusion matrix we can construct the following metrics (among others):

$$\text{Sensitivity} = \frac{a}{a+c}, \qquad \text{Positive predictive value} = \frac{a}{a+b},$$

$$\text{Specificity} = \frac{d}{b+d}, \qquad \text{Negative predictive value} = \frac{d}{c+d}.$$

In machine learning the sensitivity is also called recall, and the positive predictive value is also called precision.

Example: 1992 US National Election Survey.

Before the 1992 US election voters were asked if they preferred George Bush (Republican) or Bill Clinton (Democrat).

(Here we have excluded voters who preferred any other candidate).

Apart from their candidate preference, the respondents' income level was also recorded on a 5-point scale, where 1 corresponds to 'poor' and 5 corresponds to 'rich'.

Can we say that voters with higher incomes prefer conservative candidates?

We could see if the income is a successful predictor of the voting intentions, i.e. we could try to build a classification algorithm that labels voters 'Republican' (1) or 'Democrat' (0) based on their income level, and see how well this method performs.

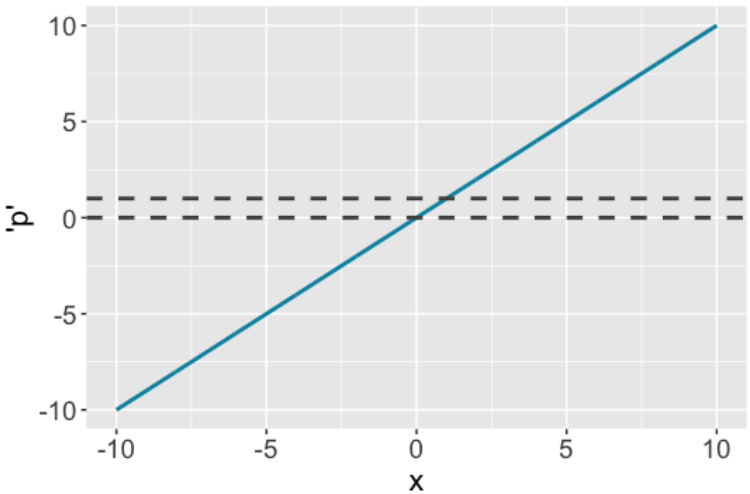
WHY NOT LINEAR?.

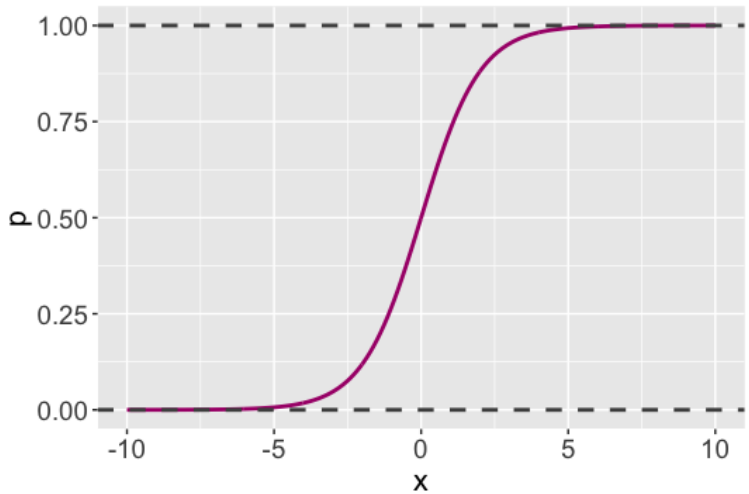
To build a classifier we will be using logistic regression to predict the probability of favouring the Republican candidate using the income category of the voter as an input.

Why logistic regression and not linear regression?

$$''p = \beta_0 + \beta_1 x''$$

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





- Using linear regression would give us negative, and greater-than-one probabilities.
- The log-odds scale ($\text{logit}(p) = \beta_0 + \beta_1 x$) of the logistic regression however restricts the outcome to the interval $(0, 1)$.

BINARY CLASSIFICATION WITH LOGISTIC REGRESSION.

Let X denote the explanatory variable, and assume the two classes are 0 and 1.

1. First we use the training data to estimate the coefficients β_0 and β_1 in the expression
$$P(Y = 1|X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}.$$
This involves fitting a generalised linear model to the available observations:
```
fit <- glm(y~x,family=binomial,data=df)
```

2. In order to predict the class of an observation $\widetilde{X}$ (saved in the `test` data frame), we then first choose a threshold p^* (the default is $p^* = 0.5$).

3. Then using the estimated coefficients, we compute $p = P(Y = 1|\widetilde{X})$ for this observation that is, the probability that $\widetilde{X}$ belongs to class 1:
```
a=predict(fit,newdata=test,type="response")
```

4. Finally, we assign a class $\widetilde{Y}$ to the observation $\widetilde{X}$ such that
$$\widetilde{Y} = \begin{cases} 1, & \text{if } p > p^*, \\ 0, & \text{otherwise.} \end{cases}$$
```
b=ifelse(a>pstar,1,0)
```

US Election Survey - Logistic regression.

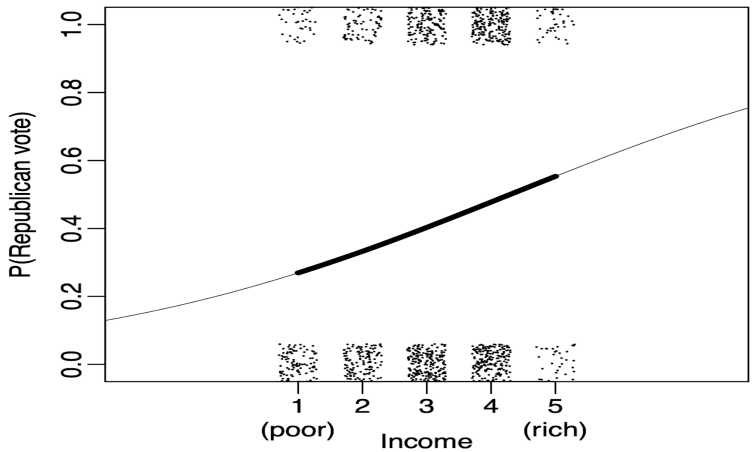
Fitting the glm to the US election data gives the following outcome. Note that both the intercept and the `income` variable are highly significant.

```
## Call:
## glm(formula = vote ~ income, family = binomial, data = nes)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2699  -1.0162  -0.8998   1.2152   1.6199
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.3017     0.1828  -7.122 1.06e-12 ***
## income        0.3033     0.0551   5.505 3.69e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Dispersion parameter for binomial family taken to be 1)
```

As the income level increases the probability of favouring the conservative candidate also increases.

More precisely, with a unit increase in income level the *log odds* of the voter favouring the conservative candidate increases by about 0.3. *But does this mean that we will be able to correctly classify the voters?*

US Election Survey - Prediction.



So how well can we predict the observed data?

Using the threshold $p^* = 0.5$ we get the following confusion matrix

	Observed $= 1$	Observed $= 0$
Predicted $= 1$	39	37
Predicted $= 0$	463	683

This gives an accuracy of 0.59, and we can also compute the following metrics

$$\text{Sensitivity} = \frac{39}{502} = 0.08, \qquad \text{Positive predictive value} = \frac{39}{76} = 0.51,$$

$$\text{Specificity} = \frac{683}{720} = 0.95, \qquad \text{Negative predictive value} = \frac{683}{1146} = 0.6.$$

US Election Survey - Comments.

A binary classifier usually uses $p^* = 0.5$ as a threshold.

- This is because most classification algorithms aim to estimate the Bayes classifier, which assigns an observation to the class for which the probability of belonging to that class is the highest.
- However we could choose any p^* such that $0 < p^* < 1$. The actual value will depend on the aim of our analysis.

For example, our classifier not only has a low accuracy, bu it also fails to correctly identify the majority of Republican voters (notice that the sensitivity is very low).

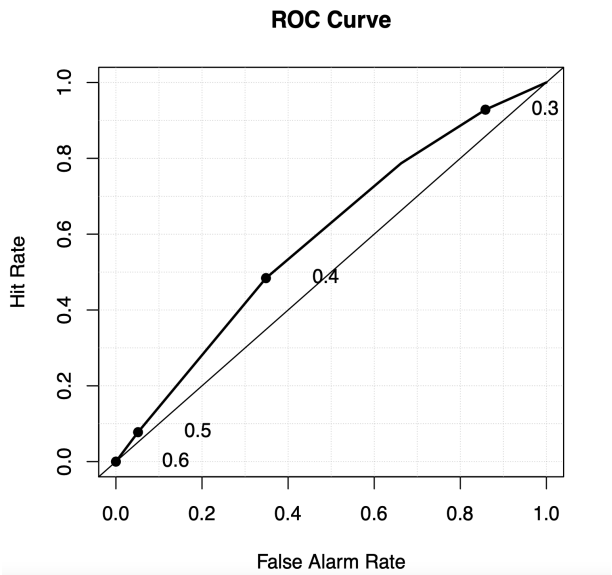
- Thus, if our primary aim was to correctly classify those voters who prefer the Republican candidate, we could try to decrease the threshold.
- Decreasing the threshold however will most likely result in some decrease in specificity.
- So there is always a trade-off between sensitivity and specificity.
- A ROC (Receiver Operator Characteristic Curve) can help in deciding the best threshold value.

ROC CURVE.

A Receiver Operating Characteristic (ROC) curve plots the sensitivity (hit rate) against $1 -$ specificity (false alarm rate) for a range of probability thresholds.

Each point on the curve corresponds to a different threshold. We can use the provided information to select the best threshold for the trade-off we want to make.

For this we need to compare the cost of failing to detect positives vs the cost of raising false alarms.



High threshold usually corresponds to

- High specificity
- Low sensitivity

Low threshold usually corresponds to

- Low specificity
- High sensitivity

Without having a strong preference, one common approach is choosing the cut-off point closest to the $(0, 1)$ corner of the ROC plane.

AREA UNDER THE CURVE.

The Area Under the Curve (AUC) of the ROC plot provides a measure of the overall performance of the classifier across all possible thresholds (how much the model is capable of distinguishing between classes).

The ideal ROC curve hugs the top left corner, so the larger the AUC the better.

- AUC $= 0.5$ indicates no discriminatory power.

  That is, on the given dataset, the method performs no better than the classifier that randomly assigns observations to one of the two classes.

- AUC $= 1$ indicates perfect discriminatory power.

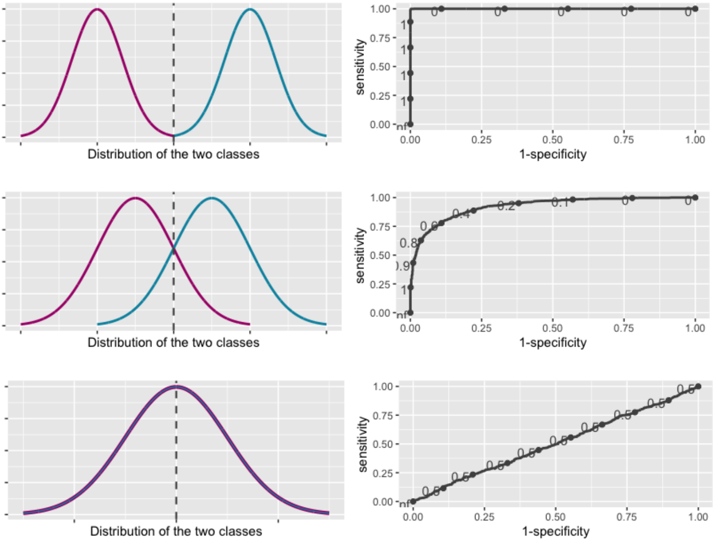
As a rule of thumb:

AUC	Discriminatory power
≥ 0.9	Outstanding
$0.8 - 0.9$	Excellent
$0.7 - 0.8$	Acceptable

For our simple classifier the AUC is 0.59 so, regardless of the choice of cut-off point, the method has a poor discriminatory power.

ROC (and AUC) is useful when comparing different classifiers, since they take into account all possible thresholds.

ROC EXAMPLES.



MULTIPLE LOGISTIC REGRESSION.

If we have more than one potential predictors, $\boldsymbol{X} = (X_1, \ldots, X_p)$, and the class is denoted by Y, then the log-odds (logit) of belonging to class 1 can be written as

$$\log\left(\frac{P(Y=1|\boldsymbol{X})}{1-P(Y=1|\boldsymbol{X})}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$$

which then gives

$$p = P(Y=1|\boldsymbol{X}) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}.$$

1. In the case of multiple predictors, we again use logistic regression to estimate the coefficients $\beta_0, \beta_1, \ldots, \beta_p$, keeping significance in mind.

2. Once we have the estimated coefficients, $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, and have chosen a cut-off point p^*, we can assign a label to any new observation.

3. If the new observation has explanatory variables $\widetilde{\boldsymbol{X}} = (\widetilde{X}_1, \ldots, \widetilde{X}_p)$, then the probability of it belonging to class 1 is

$$p = P(Y=1|\widetilde{\boldsymbol{X}}) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1\widetilde{X}_1+\cdots+\hat{\beta}_p\widetilde{X}_p}}{1+e^{\hat{\beta}_0+\hat{\beta}_1\widetilde{X}_1+\cdots+\hat{\beta}_p\widetilde{X}_p}}.$$

4. If $p > p^*$, we assign the observation $\widetilde{\boldsymbol{X}}$ to class 1; and to class 0 otherwise.

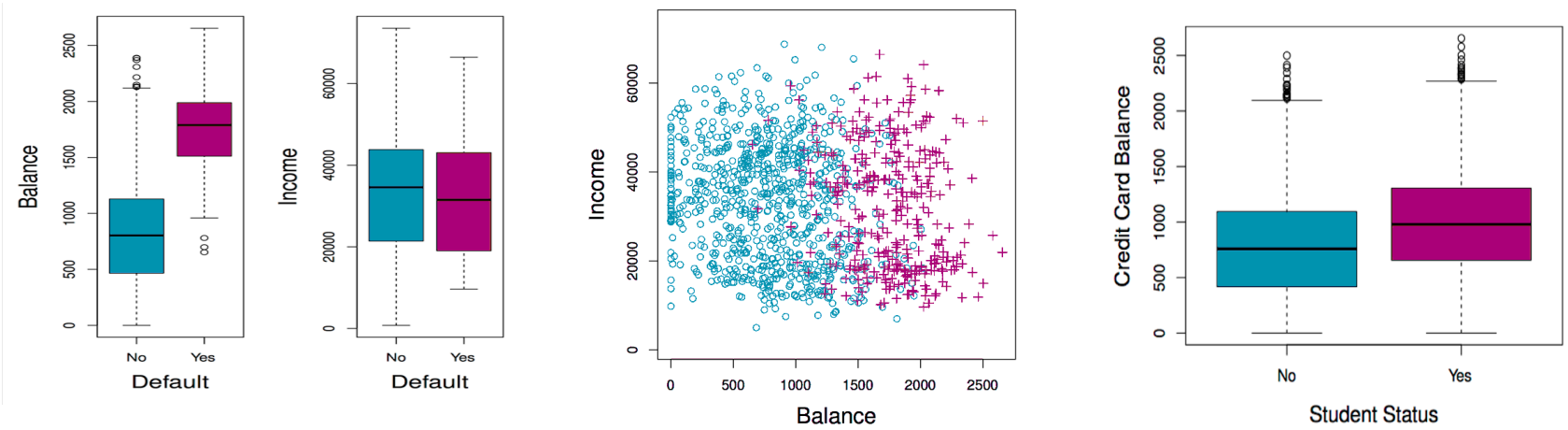
Default data.

Assume we would like to be able to predict customers that are likely to default on their credit card.

We have the following explanatory variables, $\boldsymbol{X}$:

- Annual Income (Quantitative)
- Monthly credit card balance (Quantitative)
- Student status: yes or no (Qualitative)

The label Y that we are trying to predict is a categorical variable: Default (Yes or No).



Default data - Multiple logistic regression.

Fitting a full generalised linear model to the data gives the following outcome.

Model 1	Coefficient	Std. Error	p-value
`Intercept`	-10.87	0.49	<0.0001
`balance`	0.0057	0.0002	<0.0001
`income`	0.0030	0.0082	0.71
`student[Yes]`	-0.65	0.24	0.0062

Notice that the `income` is not significant. Dropping that from the model gives

Model 2	Coefficient	Std. Error	p-value
`Intercept`	-10.75	0.37	<0.0001
`balance`	0.0057	0.0002	<0.0001
`student[Yes]`	-0.72	0.15	<0.0001

Now all the variables are highly significant.

We can use the above model fit to predict the probability of default for a new observation. For example, a student with credit card balance of £1500 has an estimated probability of default of

$$\hat{p} = \frac{e^{-10.75+1500\times 0.0057-0.72}}{1+e^{-10.75+1500\times 0.0057-0.72}} = 0.051.$$

Default data - Simple logistic regression.

The choice of threshold for classification depends on what the bank's aim is with the analysis.

Instead, here we will see how the estimated coefficients change if we use just one predictor from the previous model.

That is, we will also fit two single-predictor models, one with just the `balance` and the other one with just the student status as the explanatory variable, which leads to the following outcomes.

Model 3	Coefficient	Std. Error	p-value
`Intercept`	-10.65	0.36	<0.0001
`balance`	0.0055	0.0002	<0.0001

Model 4	Coefficient	Std. Error	p-value
`Intercept`	-3.5	0.071	<0.0001
`student[Yes]`	0.41	0.12	0.0004

Notice that the coefficient estimate for the balance hasn't changed much, however the coefficient estimate for student status became positive, when student status is the only predictor (in the previous model $\hat{\beta} = -0.72$).

Using the coefficient estimate we can compute that the probability of default for a student is 0.043, while the probability of default for a non-student is just 0.29.

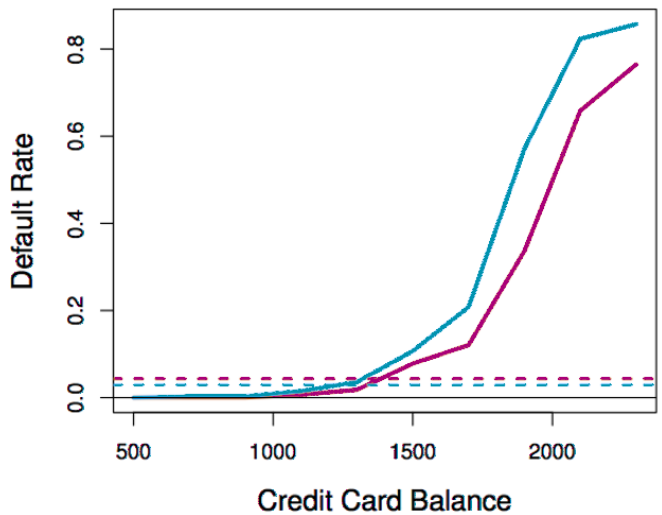
Default data confounding.

The change in the student effect shows the dangers of using just one predictor when other predictors might also be relevant, especially when there is correlation among predictors.

This phenomenon is called confounding.

Overall what we can conclude is that

- Students are riskier than non-students if no information about the credit card balance is available.
- However, a student is less risky than a non-student with the same credit card balance.



Students tend to hold higher levels of debt, which is associated with higher probability of default.

So even though individual students with a given credit card balance tend to have a lower probability of default than a non-student with the same balance, students on the whole also tend to have higher credit card balances. This results in higher overall rate for students than non-students.

EXTENSIONS.

Logistic regression can be extended to multi-class scenarios, or can be made more flexible by using higher-order terms in the regression.

1. Multinomial logistic regression is the extension of the two class logistic regression models we have discussed to cases when we have more than two categories.

   However in practice this method is barely used. Other classification methods are preferred for multiple-class classification problems.

2. By default, the logistic regression is linear on the log-odds scale, and produces a linear decision boundary for classification.

   However we often need something more flexible than a linear decision boundary.

   In these cases polynomial logistic regression can be used to create more flexible, polynomial decision boundaries.

   This can be achieved by using explanatory variables like $\{X_1, X_2, X_1 \cdot X_2, X_1^2, X_2^2\}$.

   That is, we can multiply features, square them, or we can increase the degree of the polynomial even further, depending on the model we want to fit.

   But we have to keep in mind that as the degree of polynomial increases, overfitting could become a problem.