STATS 235: Modern Data Analysis
Classification Models– Logistic Regression

Babak Shahbaba

Department of Statistics, UCI

Logistic regression model

- When dealing with binary outcome variables, we assume the response variable, y_i, has a Bernoulli distribution (or Binomial if $n_i > 1$),

$$y_i|\mu_i \sim \mathrm{Bernoulli}(\mu_i)$$

- In this case, a common link function to connect the mean of the response variable to a set of predictors, x_i, is the *logit* function defined as follows:

$$g(\mu_i) = \log(\frac{\mu_i}{1-\mu_i}) = \log[\frac{P(y_i = 1|x_i, \beta)}{1 - P(y_i = 1, \beta|x_i)}] = x_i\beta$$

where $\beta = (\beta_0, \beta_1, ..., \beta_p)$.

- Note that

$$\mu_i \quad = \quad P(y_i = 1|x_i, \beta) = \frac{\exp(x_i\beta)}{1 + \exp(x_i\beta)}$$

Logistic regression model

- The likelihood function for the general case where $n_i > 1$ is defined in terms of β as follows:

$$
\begin{aligned}
p(y|\mu) &\propto \prod_{i=1}^{n} \mu_i^{y_i}(1-\mu_i)^{n_i-y_i} \\
p(y|\beta) &\propto \prod_{i=1}^{n} \left(\frac{\exp(x_i\beta)}{1+\exp(x_i\beta)}\right)^{y_i} \left(\frac{1}{1+\exp(x_i\beta)}\right)^{n_i-y_i}
\end{aligned}
$$

Logistic regression model

- The score function for this model is as follows:

$$u_j(\beta) = \sum_{i=1}^{n}[y_i - n_i \frac{\exp(x_{i\beta})}{1+\exp(x_i\beta)}]x_{ij}$$

  where $u(\beta)$ is a $p+1$ vector.

- The Fisher information is

$$i_{jk}(\beta) = \sum_{i=1}^{n} n_i x_{ij} x_{ik} \frac{\exp(x_{i\beta})}{[1+\exp(x_i\beta)]^2}$$

Maximum likelihood estimation (MLE)

- We can use Newton's method to find the MLE, $\hat{\beta}$.
- As usual, $cov(\hat{\beta}) = [i(\hat{\beta})]^{-1}$ asymptoticly.
- The standard error for each β is obtained by taking the square root of the corresponding diagonal element of $cov(\hat{\beta})$.

Interpretation

- To interpret β, notice that $\log[\frac{P(y_i=1|x_i,\beta)}{1-P(y_i=1,\beta|x_i)}]$ is the log of odds for the outcome of interest, $y_i = 1$.
- The intercept β_0 is therefore the log of odds when all predictors are set to zero (note that this might not make sense in some cases).
- Or we can say, $\exp(\beta_0)$ is the odds when all predictors are set to zero.
- $\exp(\beta_j)$ on the other hand is how much the odds multiplicatively increases for one unit increase in x_j when all other predictors are fixed.
- Or we can say, $\exp(\beta_j)$ is the odds ratio for subjects with $X_j = x_j + 1$ compared to subjects with $X_j = x_j$ when all other predictors are fixed.
- Positive β_j indicates that the odds increases as x_j increases (everything else fixed), where is for negative estimate of β_j the odds decreases as x_j increases (everything else fixed).

Model selection

- Similarly to linear regression analysis, modeling binary response variables involves many decisions regarding the type of model.
- The main decision is to choose a set of predictors to include in the model.
- In statistical inference, the variables should be selected through a proper hypothesis testing procedure.
- However, in the absence of a clear hypothesis (which is commonly the case in machine learning problems with a large number of predictors), we could use Akaike information criterion (AIC) or Bayesian information criterion (BIC) to choose a model.

Model selection

- If our objective for building a logistic regression model is to predict the values of response variable for future cases, it makes more sense to select the model that would help us in prediction.
- For this purpose, we could build the model on one part of the data, i.e., the *training set*, fine-tune it on another part, i.e., the *validation set*, and test it on the third part, i.e., the *test set*.
- Alternatively, we could use *cross-validation* or *leave-one-out* procedure.

Predictive power

- When apply our model to the test set, we need a good measurement for evaluating the predictive power of the model; that is, how well our model can identify the correct class (0 or 1) for future observations.

- A common measure for predictive power is *accuracy rate*, which is defined as the percentage of the times the correct class (0 or 1 in this case) is predicted for future observations (or observations in the test set).

$$acc = \frac{\sum_{i=1}^{n_t} I(\hat{y}_i = y_i)}{n_t}$$

  where n_t is the number of observations in the test set, y_i is the true class, and $\hat{y}_i$ is the predicted class for i^{th} observation in the test set. The index i here is for test cases.

- Instead of accuracy rate, we could also use error rate, which is defined as the percentage of the times the wrong class is predicted.

Predictive power

- Note that the output of logistic regression models are in fact between 0 and 1, which are interpreted as probabilities.
- Therefore, we need to set an appropriate cutoff to obtain $\hat{y}$ as a binary prediction.
- In general, the cutoff depends on the loss function; that is, the cost of predicting the class as 0, when the true class is 1, and vice versa.
- In most practical problems, the costs of misclassifying 0 as 1 and 1 as 0 are not the same.
- For 0-1 loss function, we assign a test case to the class with the highest probability; that is, we set the cutoff at 0.5.

Predictive power

- Instead of averaging over all predictions, it might be more informative to separate the types of error.

- One common approach for doing this is to present the results in a *classification table* (a.k.a, *confusion matrix*)

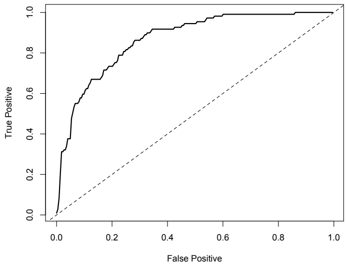
		Predicted class 0	Predicted class 1
True class	0	True Negative	False Positive
	1	False Negative	True Positive

- Based on this table, we have

$$\begin{aligned} \text{Sensitivity} &= P(\hat{y} = 1 | y = 1) \\ \text{Specificity} &= P(\hat{y} = 0 | y = 0) \end{aligned}$$

ROC

- Receiver Operating Characteristic (ROC) curve allows for simultaneous consideration of sensitivity and specificity without setting an arbitrary cut-off.
- The curve plots sensitivity (true positive) as a function of 1-specificity (false positive).



ROC

- Each point on the curve corresponds to a specific value of the cutoff.
- A more accurate model will have an ROC curve further away from the diagonal line (random model) with perfect prediction corresponding to the (0, 1) point.
- The Area Under the ROC Curve (AUC) is used as a summary statistic to compare models. For a perfect model, the AUC is 100%.

Decision boundary

- Note that for a logistic regression model as a classifier, the decision boundary is a hyperplane since the boundary is where $P(y = 1|x, \beta) = P(y = 0|x, \beta)$.

- Therefore, at the boundary we have

$$log(\frac{P(y = 1|x, \beta)}{1 - P(y = 1|x, \beta)}) = x\beta = 0$$

- Where $\{x|x\beta = 0\}$ is a hyperplane.

- For two dimensional spaces, the above hyperplane is of course a straight line.