Logistic regression

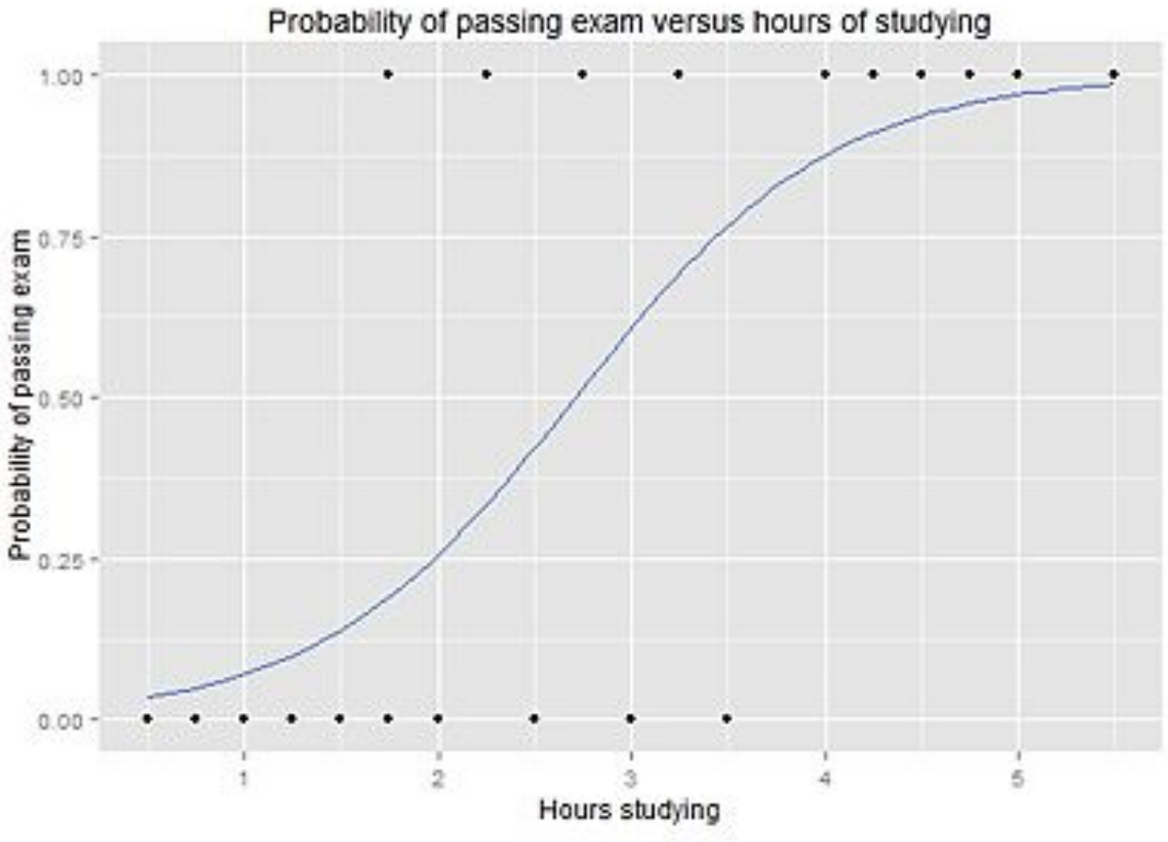
Outline

- logistic regression
- ROC curve

What is Logistic Regression?

- supervised machine learning method that uncovers a linear relationship between a set of independent variables x_i and one dependent binary variable y
  - y, is exponential family of probability distributions

sigmoid: s-shaped curve: that can take any real-valued number and map it into a value between 0 and 1



probability of passing exam=1/(1+e^(-hours studying))

Logistic function

- success: labeled ‘1’, failure:labeled ‘0’
- P: probability of success
- Odd of success

$$Odds = \frac{p}{1-p}$$

- Logistic function

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p = X\beta$$

$$p = \frac{e^{X\beta}}{1+e^{X\beta}} = \frac{1}{1+e^{-X\beta}}$$

- When $e^{X\beta}$ tends towards 0, p tends towards 0.
- When $e^{X\beta}$ tends towards ∞, p tends towards 1.
- p is always within the range of (0,1)

Estimating the coefficients

The best coefficients would result in a model that would predict a value very close to 1 for the success class and a value very close to 0 for the failure class. The intuition for maximum-likelihood for logistic regression is a search procedure seeks values for the coefficients that minimize the error in the probabilities predicted by the model.

Maximum likelihood estimation:

$$p = \frac{e^{X\beta}}{1+e^{X\beta}} = \frac{1}{1+e^{-X\beta}}$$

$$l(\beta_0, \beta_1, ..., \beta_p) = \prod_{i:y_i=1} p(x_i) \prod_{j:y_j=0} (1-p(x_j))$$

- joint probabilities of success and failure within the data set
- include the coefficients parameter
- maximize the function to select the combination of coefficients that produces the highest likelihood for the data set

Estimating the coefficients

For example we have one predictor, we need to calculate β_0 and β_1

$$
\begin{aligned}
\ell(\beta_0, \beta) &= \sum_{i=1}^{n} y_i \log p(x_i) + (1 - y_i) \log(1 - p(x_i)) \\
&= \sum_{i=1}^{n} \log(1 - p(x_i)) + \sum_{i=1}^{n} y_i \log \frac{p(x_i)}{1 - p(x_i)} \\
&= \sum_{i=1}^{n} \log(1 - p(x_i)) + \sum_{i=1}^{n} y_i (\beta_0 + x_i \cdot \beta) \\
&= \sum^{n} -\log 1 + e^{\beta_0 + x_i \cdot \beta} + \sum^{n} y_i (\beta_0 + x_i \cdot \beta)
\end{aligned}
$$

Typically, to find the maximum likelihood estimates we'd differentiate the log likelihood with respect to the parameters, set the derivatives equal to zero, and solve.

$$
\begin{aligned}
\frac{\partial \ell}{\partial \beta_j} &= -\sum_{i=1}^{n} \frac{1}{1 + e^{\beta_0 + x_i \cdot \beta}} e^{\beta_0 + x_i \cdot \beta} x_{ij} + \sum_{i=1}^{n} y_i x_{ij} \\
&= \sum_{i=1}^{n} \left(y_i - p(x_i; \beta_0, \beta)\right) x_{ij} \quad = 0
\end{aligned}
$$

Prediction

Predict the probability of success

$$p(X) = \frac{e^{X\beta}}{1+e^{X\beta}} = \frac{1}{1+e^{-X\beta}}$$

$$\hat{y}_i = \begin{cases} Success\ (1) & \hat{p}_i \geq c \\ Failure\ (0) & \hat{p}_i < c \end{cases}$$

usually c=0.5

Prepare Data

The assumptions made by logistic regression about the distribution and relationships in the data are much the same as the assumptions made in linear regression.
Ultimately in logistic regression we focus on making accurate predictions rather than interpreting the results. As such, we can break some assumptions as long as the model is robust and performs well.

- Binary Output Variable: logistic regression is intended for binary (two-class) classification problems. It will predict the probability of an observation belonging to the success class, which can be snapped into a 0 or 1 classification.
- Remove Noise:if we want lower error in the prediction, we can consider removing outliers and possibly misclassified observation from training data.
- Gaussian Distribution: Logistic regression is a linear algorithm (with a non-linear transform on output). It does assume a linear relationship between the input variables with the output. Data transforms of the input variables that better expose this linear relationship can result in a more accurate model. For example, we can use log, root, Box-Cox and other univariate transforms of the target variables to better expose this relationship.
- Remove Correlated Inputs: Like linear regression, the model can overfit if we have multiple highly-correlated inputs. Consider calculating the correlations between all inputs and removing highly correlated inputs.
- Fail to Converge: It is possible for the expected likelihood estimation process that learns the coefficients to fail to converge. This can happen if there are many highly correlated inputs in the data or the data is very sparse (e.g. lots of zeros in the input data).

Evaluating goodness of fit

Fitted/Residual Deviance: similar to RSS

$$D_{\text{fitted}} = -2 \ln \frac{\text{likelihood of fitted model}}{\text{likelihood of the saturated model}}.$$

- saturated model: has a parameter fit every observations in the dataset, leading to overfitting
- higher fitted deviance: a bad fit
- Fitted Deviance is calculated based on comparing the saturated model against the model at hand, it is a measure of how well fits the data in general.

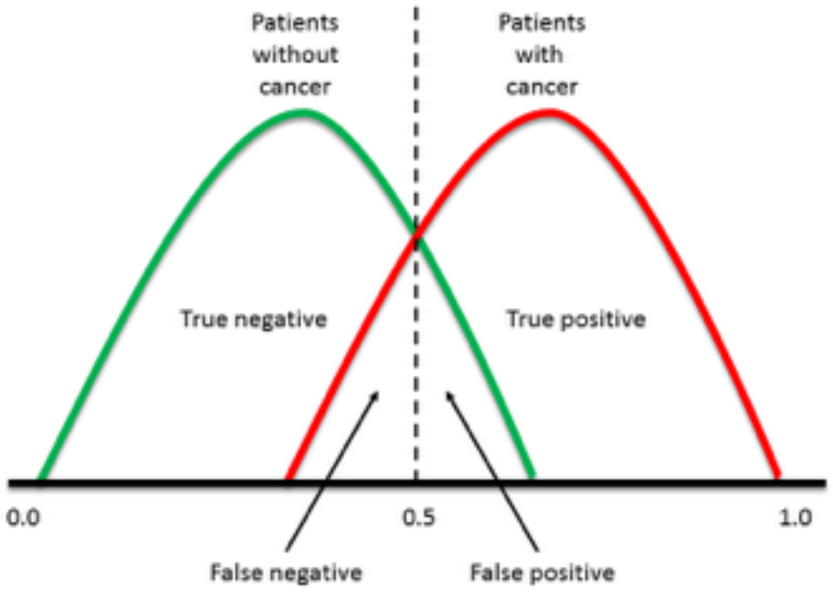
McFadden's Pseudo R^2_L : similar to adjusted R^2

$$D_{\text{null}} = -2 \ln \frac{\text{likelihood of null model}}{\text{likelihood of the saturated model}} \qquad R^2_{\text{L}} = \frac{D_{\text{null}} - D_{\text{fitted}}}{D_{\text{null}}}$$

- null model: only includes intercept term (with no predictors)
- higher R^2_L : a better fit
- Null Deviance is calculated based on comparing the saturated model against the model that only includes intercept ; it is a way of assessing the overall maximum deviance because it compares the most complicated model to the most simple model.

Classification model assessment

- Use thresholded to decide the classification. If a classifier produces a score between 0.0 (definitely negative) and 1.0 (definitely positive), it is common to consider anything over 0.5 as positive.
- Any threshold applied to a dataset (in which PP is the positive population and NP is the negative population) is going to produce true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN)
- Accuracy = (1 – Error) = (TP + TN)/(PP + NP) = Pr(C), the probability of a correct classification.
- Sensitivity = TP/(TP + FN) = TP/PP = the ability of the test to detect disease in a population of diseased individuals.
- Specificity = TN/(TN + FP) = TN / NP = the ability of the test to correctly rule out the disease in a disease-free population.

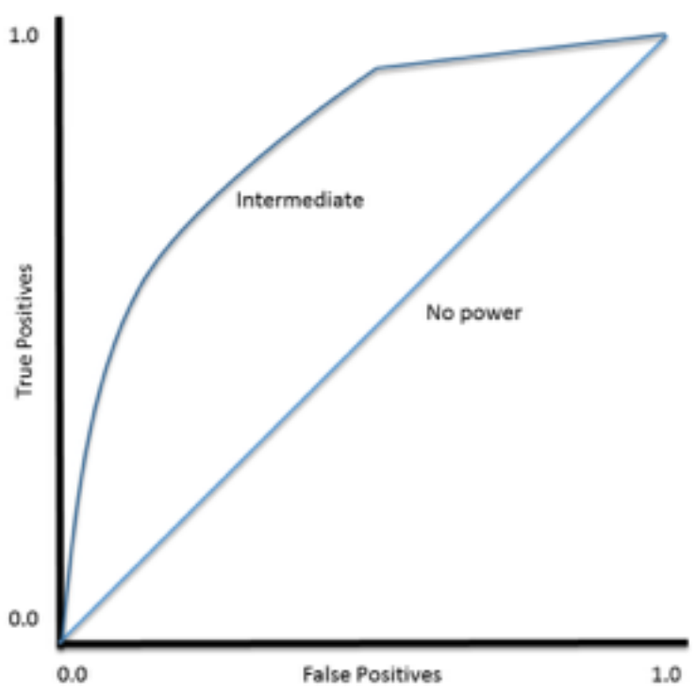


	Test Positive	Test Negative	Total
Patient Diseased	160	40	200
Patient Healthy	29940	69860	99800
Total	30100	69900	100000

- Accuracy= (TP + TN)/(PP + NP) =(160+69860)/(200+99800)= 70.0%
- Sensitivity = TP/(TP + FN) = 160 / (160 + 40) = 80.0%
- Specificity = TN/(TN + FP) = 69,860 / (69,860 + 29,940) = 70.0%
- The test will correctly identify 80% of people with the disease, but 30% of healthy people will incorrectly test positive.
- True positive rate is 80%, false positive rate is 30%.

ROC (receiver operating characteristic) curve

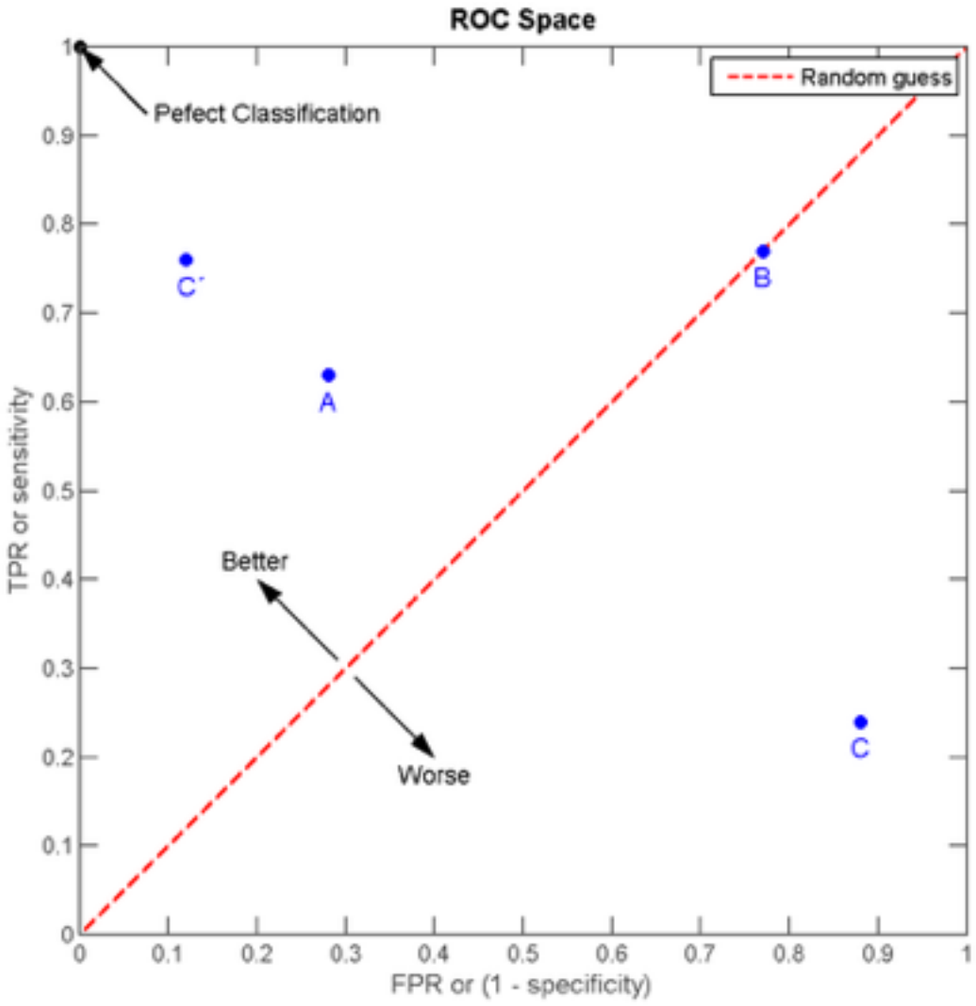
- One way to overcome the problem of having to choose a cutoff is to start with a threshold of 0, so that every case is considered as positive. We correctly classify all the positive cases, and incorrectly classify all the negative cases. Then we move the threshold over every value between 0 and 1, progressively decreasing the number of false positives and increasing the number of true positives.
- ROC curve: plot sensitivity (true positive rate) vs 1-specificity (false positive rate)
- ROC curve: select a threshold for a classifier which maximizes the true positives and minimizes the false positives
- AUC (the area under the curve): assess the performance of the classifier
- AUC is usually bigger than 0.5



The best possible prediction method would yield a point in the upper left corner or coordinate (0,1) of the ROC space, representing 100% sensitivity (no false negatives) and 100% specificity (no false positives). The (0,1) point is also called a perfect classification. A random guess would give a point along a diagonal line (the so-called line of no-discrimination) from the left bottom to the top right corners (regardless of the positive and negative base rates). An intuitive example of random guessing is a decision by flipping coins. As the size of the sample increases, a random classifier's ROC point tends towards the diagonal line. In the case of a balanced coin, it will tend to the point (0.5, 0.5).

ROC (receiver operating characteristic) curve

The diagonal divides the ROC space. Points above the diagonal represent good classification results (better than random); points below the line represent bad results (worse than random).



A			B			C			C′		
TP=63	FP=28	91	TP=77	FP=77	154	TP=24	FP=88	112	TP=76	FP=12	88
FN=37	TN=72	109	FN=23	TN=23	46	FN=76	TN=12	88	FN=24	TN=88	112
100	100	200	100	100	200	100	100	200	100	100	200
TPR = 0.63			TPR = 0.77			TPR = 0.24			TPR = 0.76		
FPR = 0.28			FPR = 0.77			FPR = 0.88			FPR = 0.12		

- The result of method A clearly shows the best predictive power among A, B, and C.
- The result of B lies on the random guess line (the diagonal line), and it can be seen in the table that the accuracy of B is 50%.
- When C is mirrored across the center point (0.5,0.5), the resulting method C′ is even better than A. This mirrored method simply reverses the predictions. Although the original C method has negative predictive power, simply reversing its decisions leads to a new predictive method C′ which has positive predictive power. When the C method predicts positive or negative, the C′ method would predict negative or positive, respectively. In this manner, the C′ test would perform the best.

Advantages and Disadvantages

Advantages

- Logistic Regression performs well when the dataset is linearly separable.
- Logistic regression is less prone to over-fitting but it can overfit in high dimensional datasets. We should consider Regularization (L1 and L2) techniques to avoid over-fitting in these scenarios.
- Logistic Regression not only gives a measure of how relevant a predictor (coefficient size) is, but also its direction of association (positive or negative).
- Logistic regression is easier to implement, interpret and very efficient to train.

Disadvantages

- Main limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables. In the real world, the data is rarely linearly separable. Most of the time data would be a jumbled mess.

- If the number of observations are lesser than the number of features, Logistic Regression should not be used, otherwise it may lead to overfit.