Lecture 10

Extensions of logistic regression

In the last lecture, we considered single variable logistic regression. Today, we will look at extending the model to more than one explanatory variable. Recall the single-variable model we considered in the last class, predicting whether a car was American made or not based on mpg. Let's look at our summary output again.

Call:
glm(formula = am ~ mpg, family = "binomial", data = mtcars)

Deviance Residuals:
Min	1Q	Median	3Q	Max
-1.5701	-0.7531	-0.4245	0.5866	2.0617

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-6.6035	2.3514	-2.808	0.00498 **
mpg	0.3070	0.1148	2.673	0.00751 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 43.230 on 31 degrees of freedom
Residual deviance: 29.675 on 30 degrees of freedom
AIC: 33.675

Number of Fisher Scoring iterations: 5

Suppose that we wanted to extend our model to use hp (horsepower) in addition to mpg.

Call:
glm(formula = am ~ mpg + hp, family = "binomial", data = mtcars)

Deviance Residuals:
Min	1Q	Median	3Q	Max
-1.41460	-0.42809	-0.07021	0.16041	1.66500

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-33.60517	15.07672	-2.229	0.0258 *
mpg	1.25961	0.56747	2.220	0.0264 *
hp	0.05504	0.02692	2.045	0.0409 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 43.230 on 31 degrees of freedom
Residual deviance: 19.233 on 29 degrees of freedom
AIC: 25.233

Number of Fisher Scoring iterations: 7

In comparison with the previous model, the P-values of the coefficients have increased, but they remain significant. The residual deviance has decreased and so has the AIC, telling us that our model has improved. Model selection methods, such as stepwise, forward and backward selection can also be applied to multiple logistic regression.

The logistic regression model uses a logistic function (sigmoid function) to transform a linear combination of the independent variables into a probability score between 0 and 1. The model estimates the coefficients (also called weights or parameters) of the independent variables, which represent the strength and direction of their relationship with the dependent variable.

The logistic regression equation can be expressed as follows:

$$P(y = 1|x_1, x_2, \ldots, x_n) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}}$$

where:

$P(y = 1|x_1, x_2, \ldots, x_n)$ is the predicted probability of the dependent variable y being 1 (i.e., the positive outcome) given the values of the independent variables $x_1, x_2, \ldots, x_n$.

β_0 is the intercept or bias term, which represents the predicted log-odds of the outcome variable when all the independent variables are zero.

$\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients or weights of the independent variables $x_1, x_2, \ldots, x_n$, which indicate the change in the log-odds of the outcome variable for a unit change in the corresponding independent variable, holding all other variables constant.

To estimate the coefficients, the logistic regression model uses maximum likelihood estimation, which seeks to find the values of the coefficients that maximize the likelihood of observing the given set of data. The model is evaluated based on its ability to correctly classify the observed data into the binary outcomes, using measures such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic (ROC) curve.

Null deviance is a statistical term used in logistic regression analysis that refers to the measure of the overall goodness-of-fit of a logistic regression model that includes only the intercept or the constant term, but no independent variables (also known as the "null model"). In other words, the null deviance represents the amount of variability in the dependent variable that can be explained by the intercept alone.

The null deviance is computed as the difference between the deviance of the null model and the deviance of a model that assumes a perfect fit to the data. The deviance is a measure of how well the model fits the data, and it is calculated as minus two times the log-likelihood ratio of the model, compared to the perfect fit model. A smaller deviance value indicates a better fit of the model to the data.

The null deviance is useful for assessing the significance of the independent variables included in the logistic regression model. It is compared to the residual deviance, which is the difference between the deviance of the full model (with all independent variables) and the deviance of the null model, to determine if the inclusion of the independent variables significantly improves the model fit.

If the residual deviance is substantially smaller than the null deviance, it suggests that the independent variables have a significant impact on the outcome variable and contribute to the model fit. However, if the residual deviance is similar to the null deviance, it implies that the independent variables do not improve the model fit significantly and the null model may be adequate.

Residual deviance is a statistical term used in logistic regression analysis to evaluate the goodness-of-fit of a logistic regression model that includes one or more independent variables. It measures the amount of unexplained variability or lack of fit of the model to the observed data, after taking into account the effects of the independent variables.

The residual deviance is calculated as the difference between the deviance of the full logistic regression model (with all independent variables) and the deviance of the null model (with no independent variables). The deviance is a measure of the goodness-of-fit of the model and is calculated as minus twice the log-likelihood ratio of the model, compared to the saturated model (a model that perfectly fits the data).

A smaller residual deviance value indicates a better fit of the model to the data, and a larger value indicates a poorer fit. The residual deviance can be used to test the significance of the independent variables in the model, by comparing the residual deviance to the null deviance. If the residual deviance is significantly smaller than the null deviance, it indicates that the model with the independent variables provides a better fit to the data than the null model.

The residual deviance can also be used to calculate other goodness-of-fit measures, such as the Pearson chi-square statistic and the likelihood ratio test statistic. These measures can be used to assess the overall fit of the model, the adequacy of the model specification, and the potential for overfitting or underfitting the data.

Some common model diagnostics for logistic regression include:

Residual analysis: This involves examining the residuals or errors in the model to check if they are normally distributed, have constant variance, and are independent. Deviations from these assumptions may indicate problems with the model specification, such as omitted variables or nonlinear relationships.

Influence analysis: This involves identifying influential observations or outliers that may have a disproportionate effect on the estimated coefficients or predicted probabilities. These observations can

be identified using measures such as Cook's distance or leverage, and may require further investigation or data cleaning.

Multicollinearity analysis: This involves examining the correlations between the independent variables in the model to check for multicollinearity or high correlation between predictors. High multicollinearity can make it difficult to interpret the individual effects of each variable and can lead to unstable or biased coefficient estimates.

Goodness-of-fit tests: These tests check whether the model fits the data well, using measures such as the deviance, Pearson chi-square statistic, or Hosmer-Lemeshow test. A significant p-value indicates that the model does not fit the data well and may require further modifications.

Model selection and comparison: This involves comparing different models with different sets of independent variables or specifications to determine the best fitting model. This can be done using techniques such as stepwise regression, AIC or BIC criteria, or cross-validation.

Hypothesis tests on dimensionality reduction in logistic regression are used to determine the significance and usefulness of reducing the number of independent variables in the model. These tests can help to identify which variables are most important for predicting the outcome variable, and whether including additional variables improves the model fit.

One common hypothesis test for dimensionality reduction in logistic regression is the likelihood ratio test. This test compares the fit of the full model, with all independent variables, to the fit of a reduced model, with some of the independent variables removed. The likelihood ratio test statistic is calculated as the difference in deviance between the full and reduced models, multiplied by -2, and follows a chi-square distribution with degrees of freedom equal to the number of variables removed.

If the p-value of the likelihood ratio test is less than the significance level (usually 0.05), it indicates that the full model is significantly better than the reduced model, and that removing some of the independent variables would result in a loss of predictive power. Conversely, if the p-value is greater than the significance level, it indicates that the reduced model is not significantly worse than the full model, and that reducing the number of variables may simplify the model without sacrificing too much predictive accuracy.

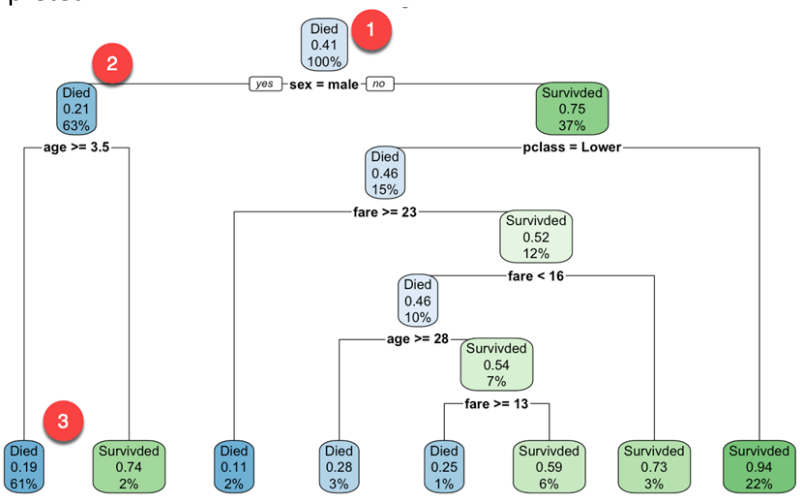
Another hypothesis test for dimensionality reduction in logistic regression is the **Wald test**. This test compares the coefficients of the full model to the coefficients of a reduced model, where some of the variables are set to zero. The Wald test statistic is calculated as the ratio of the squared difference between the two coefficients, divided by the estimated variance of the coefficient, and follows a standard normal distribution under the null hypothesis of no difference.

If the p-value of the Wald test is less than the significance level, it indicates that the coefficient for the removed variable is significantly different from zero, and that including this variable improves the model fit. Conversely, if the p-value is greater than the significance level, it indicates that the coefficient for the removed variable is not significantly different from zero, and that the variable can be safely removed from the model.

Classification methods are a type of supervised machine learning algorithm that learn to assign objects or instances to predefined categories or classes based on their characteristics or features. Here are some common classification methods used in machine learning:

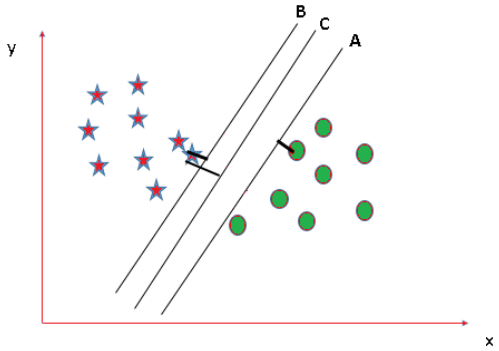
Logistic Regression: This is a statistical technique that uses a logistic function to model the probability of a binary response variable, based on one or more independent variables. Logistic regression can handle both linear and nonlinear relationships between the predictors and the response, and can be extended to multiclass classification using techniques such as one-vs-rest or softmax regression.

Decision Trees: This is a tree-based algorithm that recursively splits the data into smaller subsets based on the most informative features, in order to maximize the homogeneity or purity of the resulting subgroups. Decision trees can handle both categorical and numerical variables, and can be easily visualized and interpreted.



Random Forests: This is an ensemble method that combines multiple decision trees trained on different subsets of the data and different subsets of the features, in order to reduce overfitting and improve accuracy. Random forests can handle high-dimensional data with correlated features, and can provide estimates of feature importance and variable interactions.

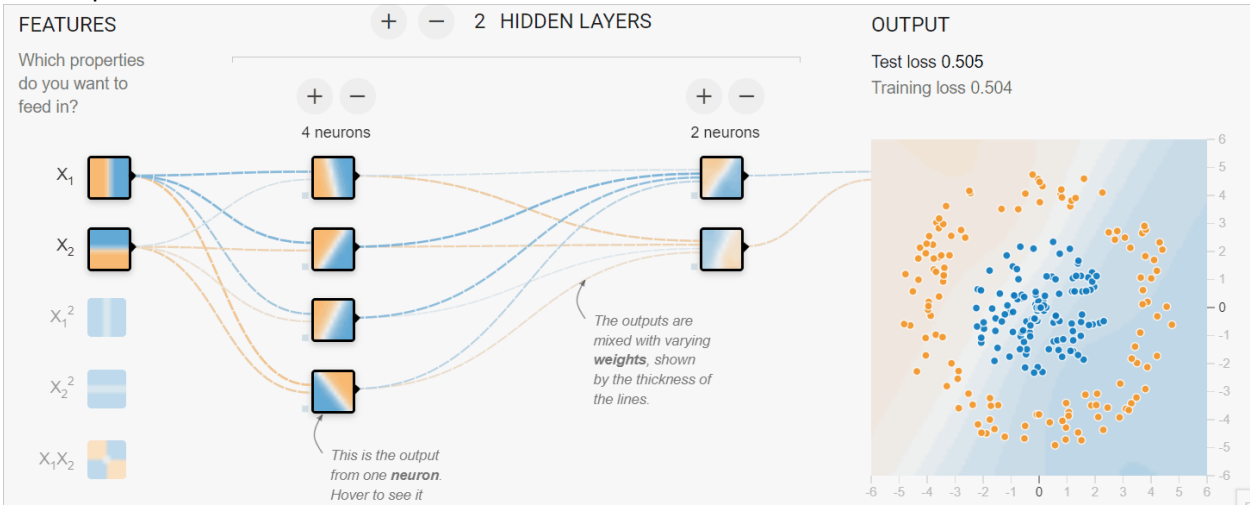
Support Vector Machines (SVMs): This is a linear or nonlinear algorithm that learns a decision boundary or hyperplane that maximally separates the classes in the feature space. SVMs can handle both binary and multiclass classification, and can incorporate kernel functions to handle nonlinearity and high-dimensional data.



Naive Bayes: This is a probabilistic algorithm that uses Bayes' theorem to calculate the probability of each class given the observed features. Naive Bayes assumes that the features are conditionally independent given the class, and can handle high-dimensional and sparse data with fast training and prediction times.

Type	Swim	Wings	Green	Sharp teeth
Cat	450/500	0	0	500/500
Parrot	50/500	500/500	400/500	0
Turtle	500/500	0	100/500	50/500

Neural Networks: This is a class of algorithms that simulate the structure and function of the human brain using interconnected nodes or neurons, and learn to approximate complex nonlinear functions through supervised or unsupervised training. Neural networks can handle various types of data and can achieve high accuracy on large and complex datasets, but can be computationally expensive and difficult to interpret.



This list is not exhaustive. The choice of method depends on the specific problem, the type and amount of data available, and the desired trade-offs between accuracy, interpretability, and computational efficiency.

References:

1. https://faculty.ksu.edu.sa/sites/default/files/probability_and_statistics_for_engineering_and_the_sciences.pdf
2. https://stats.oarc.ucla.edu/r/dae/logit-regression/
3. https://www.geeksforgeeks.org/logistic-regression-in-r-programming/
4. https://www.statology.org/null-residual-deviance/
5. http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/
6. https://rpubs.com/Saskia/520216

7. https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-12/issue-1/Dimension-reduction-based-significance-testing-in-nonparametric-regression/10.1214/18-EJS1414.pdf
8. https://www.javatpoint.com/r-classification
9. https://techvidvan.com/tutorials/classification-in-r/
10. https://www.guru99.com/r-decision-trees.html
11. https://www.geeksforgeeks.org/classifying-data-using-support-vector-machinessvms-in-r/
12. https://www.edureka.co/blog/naive-bayes-in-r/
13. https://www.datacamp.com/tutorial/neural-network-models-r