Welcome to the CoGrammar Tutorial: Multiple Linear and Logistic Regression

The session will start shortly...

Questions? Drop them in the chat. We'll have dedicated moderators answering questions.

CoGrammar

Data Science Session Housekeeping

- The use of disrespectful language is prohibited in the questions, this is a supportive, learning environment for all - please engage accordingly. **(Fundamental British Values: Mutual Respect and Tolerance)**
- No question is daft or silly - **ask them!**
- There are **Q&A sessions** midway and at the end of the session, should you wish to ask any follow-up questions. Moderators are going to be answering questions as the session progresses as well.
- If you have any questions outside of this lecture, or that are not answered during this lecture, please do submit these for upcoming Academic Sessions. You can submit these questions here: **Questions**

Data Science Session Housekeeping cont.

- For all **non-academic questions**, please submit a query: www.hyperiondev.com/support

- Report a **safeguarding** incident: www.hyperiondev.com/safeguardreporting

- We would love your **feedback** on lectures: **Feedback on Lectures**

Skills Bootcamp
8-Week Progression Overview

Fulfil 4 Criteria to Graduation

✅ Criterion 1: Initial Requirements

Timeframe: First 2 Weeks
Guided Learning Hours (GLH): Minimum of 15 hours
Task Completion: First four tasks

Due Date: 24 March 2024

✅ Criterion 2: Mid-Course Progress

60 Guided Learning Hours

Data Science - **13 tasks**
Software Engineering - **13 tasks**
Web Development - **13 tasks**

Due Date: 28 April 2024

CoGrammar

Skills Bootcamp Progression Overview

✅ Criterion 3: Course Progress

Completion: All mandatory tasks, including Build Your Brand and resubmissions by study period end
Interview Invitation: Within 4 weeks post-course
Guided Learning Hours: Minimum of 112 hours by support end date (10.5 hours average, each week)

✅ Criterion 4: Demonstrating Employability

Final Job or Apprenticeship Outcome: Document within 12 weeks post-graduation
Relevance: Progression to employment or related opportunity

CoGrammar

SKILLS FOR LIFE
SKILLS BOOTCAMPS

Department for Education

CoGrammar

Tutorial: Multiple Linear and Logistic Regression

May 2024

Learning objectives

- ❖ Understand and implement **Multiple Linear Regression** and **Logistic Regression** models

- ❖ Using Python **scikit-learn** library for **regression** and **classification** tasks

CoGrammar

Multiple Linear Regression

Recap

Multiple Linear Regression

Extension of simple linear regression, uses **several explanatory (independent) variables** (x_i) to predict the outcome of **one response (dependent) variable** (y).

$$\boxed{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_n x_n + \varepsilon}$$

y = **Output/Response/Dependent** variable

$x_1, x_2, x_3, \ldots, x_n$ = Various **Feature/Explanatory/Independent** variables

β_0 = y-intercept (constant term)

$\beta_1, \beta_2, \beta_3, \ldots, \beta_n$ = slope coefficient for each explanatory variable

ε = Model's **error** term (also called **residuals**)

CoGrammar

Evaluation Metrics

- **Mean Squared Error (MSE), Root mean squared error (RMSE) :**
  - MSE = **average squared difference** between the **predicted** and **actual** values**. RMSE** is root of MSE.
- **Mean Absolute Error (MAE)**
  - Sum of **absolute errors** between **predicted** and **actual** values.
- **R-squared (R^2) score** (coefficient of determination):
  - R^2 = **proportion of variance** in the target (dependent) variable that can be **explained by the independent variables/model.**

A **lower MSE and MAE** indicates better model performance.

An R^2 value **closer to 1** indicates a better fit of the model to the data.

CoGrammar

Feature Scaling

Normalisation	Standardisation
Rescales values to a range between 0 and 1	Centers data around mean and scales to standard deviation of 1
Useful when data distribution is unknown or **not Gaussian**	Useful with **Gaussian** data distribution
Sensitive to outliers	**Less sensitive to outliers**
Retains shape of original distribution	Changes shape of original distribution
May not preserve relationships between data points	Preserves relationships between data points
MinMaxScaler() (x – min)/(max – min)	**StandardScaler()** (x – mean)/standard deviation

Which of the following is an example of a regression problem?

1. Distinguishing different soil types based on their physical and chemical properties.
2. Simulating soil organic carbon and total nitrogen relationship
3. Both A and B
4. Neither A nor B

Which of the following is an example of a regression problem?

1. Distinguishing different soil types based on their physical and chemical properties.
2. **Simulating soil organic carbon and total nitrogen relationship**
3. Both A and B
4. Neither A nor B

In the multiple linear regression equation $y = \beta_0 + \beta_1 x + \beta_2 x_2 + \varepsilon$, what does β_2 represent?

1. The intercept of the second independent variable.
2. The intercept of the second dependent variable.
3. The slope of the second independent variable.
4. The slope of the second dependent variable.

In the multiple linear regression equation $y = \beta_0 + \beta_1 x + \beta_2 x_2 + \varepsilon$, what does β_2 represent?

1. The intercept of the second independent variable.
2. The intercept of the second dependent variable.
3. **The slope of the second independent variable.**
4. The slope of the second dependent variable.

When the independent variables are correlated with one another in a multiple regression analysis, this condition is called:

1. Linearity
2. Multicollinearity
3. Homoscedasticity
4. Normality

When the independent variables are correlated with one another in a multiple regression analysis, this condition is called:

1. Linearity
2. **Multicollinearity**
3. Homoscedasticity
4. Normality

CoGrammar

In the context of linear regression, what does the term "residual" refer to?

1. The difference between the predicted and actual values
2. The correlation between the independent and dependent variables
3. The statistical significance of the regression coefficients
4. The proportion of variance in the target variable explained by the model

In the context of linear regression, what does the term "residual" refer to?

1. **The difference between the predicted and actual values**
2. The correlation between the independent and dependent variables
3. The statistical significance of the regression coefficients
4. The proportion of variance in the target variable explained by the model

CoGrammar

In a multiple regression analysis, if the model provides a poor fit, this indicates that:

1. The sum of squares for error will be large
2. The standard error of estimate will be large
3. The multiple coefficient of determination will be close to zero
4. All of the above

In a multiple regression analysis, if the model provides a poor fit, this indicates that:

1. The sum of squares for error will be large
2. The standard error of estimate will be large
3. The multiple coefficient of determination will be close to zero
4. **All of the above**

A scatter plot between the residuals and predicted values in linear regression shows a relationship between them. Which statement is true?

1. Since there is a relationship means our model is not good
2. Since there is a relationship means our model is good
3. Cannot judge
4. None of these

A scatter plot between the residuals and predicted values in linear regression shows a relationship between them. Which statement is true?

1. **Since there is a relationship means our model is not good**
2. Since there is a relationship means our model is good
3. Cannot judge
4. None of these

CoGrammar

Logistic Regression

Recap

CoGrammar

Logistic Regression

- **Linear regression** models make **predictions** for the datasets for which dependent variables have **continuous numerical values**.

- **Logistic Regression**
  - **supervised learning** algorithm
  - **classification** algorithm
  - dependent variables are **distinct, non-continuous, categorical**

- **Classification** - predicting **probability** of **categorical variables** for a given observation and assigning the observation to the category with the highest probability.

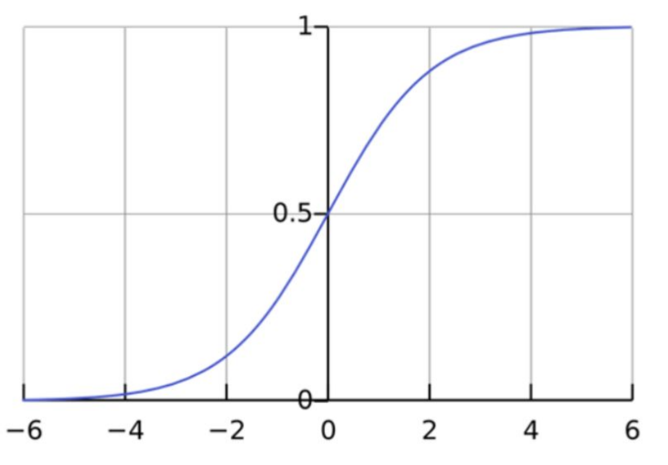
CoGrammar

Logistic function

- Logistic regression: statistical model that uses the **logistic (logit) function**, as the equation between x and y (also called **sigmoid function** or **S-shaped curve**).

- Returns only values between 0 and 1 for the dependent variable, irrespective of the values of the independent variable.

- Also model equations between **multiple independent variables** and **one dependent variable.**



Sigmoid function

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

CoGrammar

Categorical Encoding

Label Encoding	One-hot Encoding
Categorical feature is ordinal	Categorical feature is not ordinal
Categorical values are labeled into numeric values by assigning each category to a unique number	A column with categorical values is split into a binary vector, creating new binary columns for each category.
Categories are converted into unique numeric values. Fewer computations	Add more columns and will be computationally heavy
Unique information	Redundant information
Different integers are used to represent data	Only 0 and 1 are used to represent data

CoGrammar

Evaluation Metrics

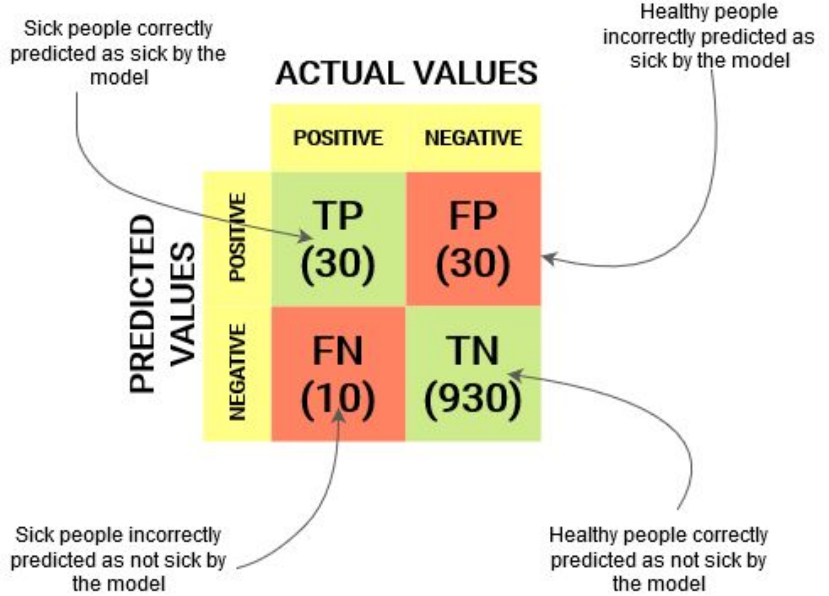
Confusion matrix, Accuracy

Precision, Recall, F1 score

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$



		ACTUAL VALUES	
		POSITIVE	NEGATIVE
PREDICTED VALUES	POSITIVE	TP (30)	FP (30)
	NEGATIVE	FN (10)	TN (930)

Sick people correctly predicted as sick by the model

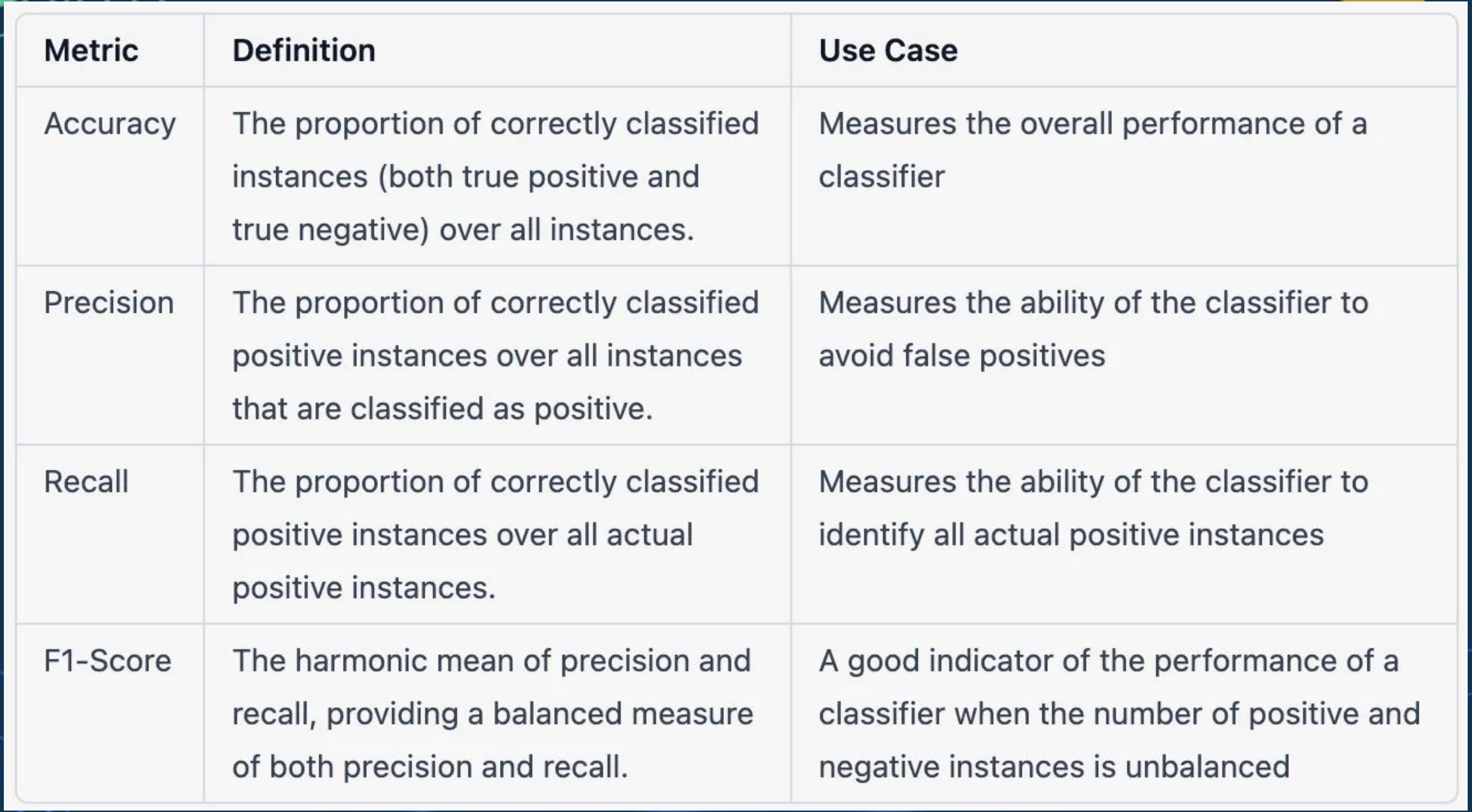
Healthy people incorrectly predicted as sick by the model

Sick people incorrectly predicted as not sick by the model

Healthy people correctly predicted as not sick by the model

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

Evaluation Metrics



Metric	Definition	Use Case
Accuracy	The proportion of correctly classified instances (both true positive and true negative) over all instances.	Measures the overall performance of a classifier
Precision	The proportion of correctly classified positive instances over all instances that are classified as positive.	Measures the ability of the classifier to avoid false positives
Recall	The proportion of correctly classified positive instances over all actual positive instances.	Measures the ability of the classifier to identify all actual positive instances
F1-Score	The harmonic mean of precision and recall, providing a balanced measure of both precision and recall.	A good indicator of the performance of a classifier when the number of positive and negative instances is unbalanced

Logistic regression assumes a:

1. Linear relationship between continuous predictor variables and the outcome variable.
2. Linear relationship between continuous predictor variables and the logit of the outcome variable.
3. Linear relationship between continuous predictor variables.
4. Linear relationship between observations.

Logistic regression assumes a:

1. Linear relationship between continuous predictor variables and the outcome variable.
2. **Linear relationship between continuous predictor variables and the logit of the outcome variable.**
3. Linear relationship between continuous predictor variables.
4. Linear relationship between observations.

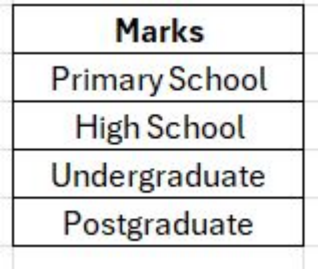
Which of the following is true?

1. Linear Regression errors values has to be normally distributed but in case of Logistic Regression it is not the case
2. Logistic Regression errors values has to be normally distributed but in case of Linear Regression it is not the case
3. Both Linear Regression and Logistic Regression error values have to be normally distributed
4. both Linear Regression and Logistic Regression error values have not to be normally distributed

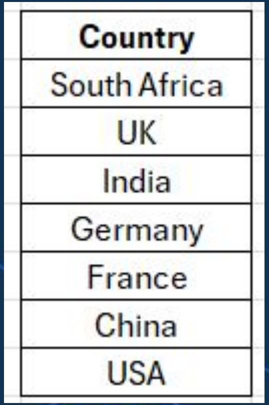
Which of the following is true?

1. **Linear Regression errors values has to be normally distributed but in case of Logistic Regression it is not the case**
2. Logistic Regression errors values has to be normally distributed but in case of Linear Regression it is not the case
3. Both Linear Regression and Logistic Regression error values have to be normally distributed
4. both Linear Regression and Logistic Regression error values have not to be normally distributed

Which encoder should be used for the two categorical columns (ideally)?



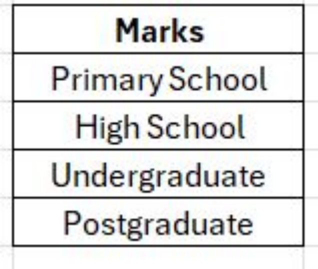
Marks
Primary School
High School
Undergraduate
Postgraduate



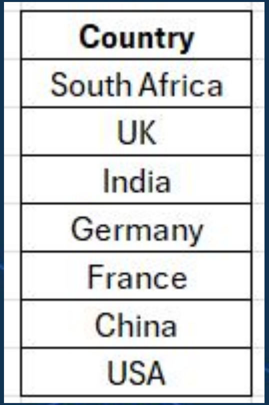
Country
South Africa
UK
India
Germany
France
China
USA

1. LabelEncoder for Marks and Country
2. OneHotEncoder for Marks and Country
3. LabelEncoder for Marks and OneHotEncoder for Country
4. OneHotEncoder for Marks and LabelEncoder for Country

Which encoder should be used for the two categorical columns (ideally)?



Marks
Primary School
High School
Undergraduate
Postgraduate



Country
South Africa
UK
India
Germany
France
China
USA

1. LabelEncoder for Marks and Country
2. OneHotEncoder for Marks and Country
3. **LabelEncoder for Marks and OneHotEncoder for Country**
4. OneHotEncoder for Marks and LabelEncoder for Country

Which of the following statements are true?

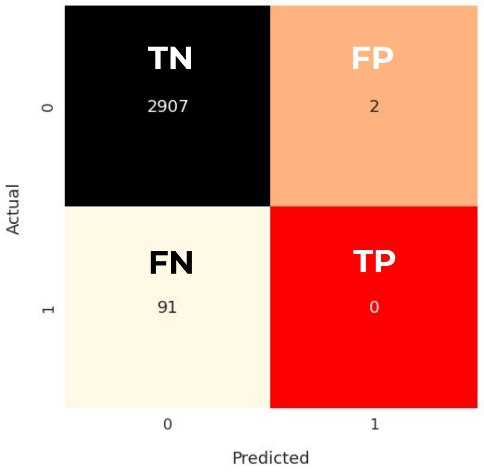
1. Precision measures the accuracy of positive predictions.
2. Recall Precision measures the accuracy of positive predictions.
3. Precision measures the completeness of positive predictions.
4. Recall measures the completeness of positive predictions.

CoGrammar

Which of the following statements are true?

1. **Precision measures the accuracy of positive predictions.**
2. Recall Precision measures the accuracy of positive predictions.
3. Precision measures the completeness of positive predictions.
4. **Recall measures the completeness of positive predictions.**

Confusion matrix for credit card debt model. Default status is 0 (did not default) and 1 (defaulted). Which statements are True?



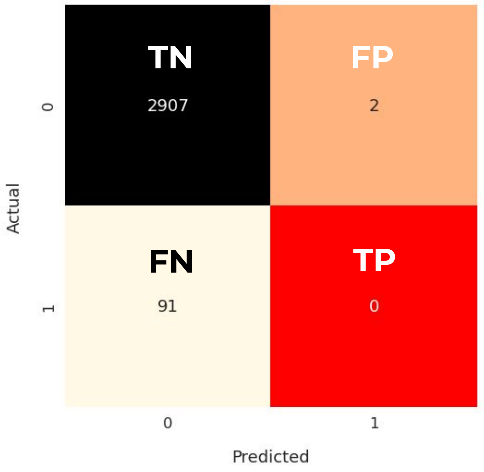
	0	1
0	TN 2907	FP 2
1	FN 91	TP 0

1. Accuracy of the model is 97% (=2907/(2907+2+91), so the model has a great performance.
2. Accuracy of the model is zero, so the model has a poor performance.
3. The precision, recall and F1-score is zero, so the model has an excellent performance.
4. The precision, recall and F1-score is zero, so the model performs poorly.

Note: TP = 0

CoGrammar

Confusion matrix for credit card debt model. Default status is 0 (did not default) and 1 (defaulted). Which statements are True?



Only partly true

1. Accuracy of the model is 97% (=2907/(2907+2+91)), ~~so the model has a great performance~~.
2. Accuracy of the model is zero, so the model has a poor performance.
3. The precision, recall and F1-score is zero, so the model has an excellent performance.
4. **The precision, recall and F1-score is zero, so the model performs poorly.**

Note: TP = 0

CoGrammar

Questions and Answers

CoGrammar

Thank you for attending

SKILLS FOR LIFE
SKILLS BOOTCAMPS

Department for Education

CoGrammar