CoGrammar

Supervised Machine Learning (Part Two)

SKILLS FOR LIFE

SKILLS BOOTCAMPS

Department for Education

Data Science Lecture Housekeeping

- The use of disrespectful language is prohibited in the questions, this is a supportive, learning environment for all - please engage accordingly. **(FBV: Mutual Respect.)**
- 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 Open Classes. You can submit these questions here: Open Class Questions

Data Science Lecture 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**

Lecture Objectives

- Learning how to make predictions on categorical dependant variables

Classification Problems

★ The linear regression models discussed in the previous tasks assumed that the dependent variable Y is a continuous numerical value. However, it is very common in machine learning problems instead to be dealing with categorical variables.

★ These variables take on distinct non-continuous values which correspond to a particular set of categories. For example, recall the previous compulsory exercise where the nurses could either work in a 'Hospital' or 'Office' position, and these positions were represented as the values 0 and 1.

★ This is an example of a categorical variable, where the category is usually encoded as a numeric value.

Classification Problems

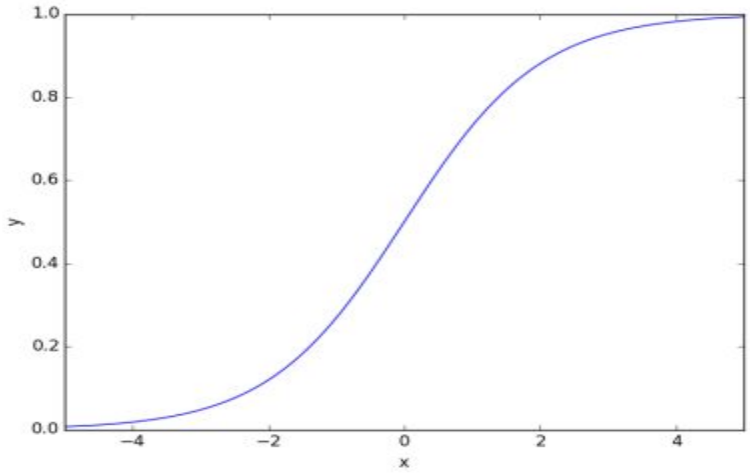
- ★ Predicting categorical variables is called classification. Classification problems are very common, perhaps even more so than problems suited for regression.

- ★ Classification involves predicting the probability of each of the categories for a given observation and assigning the observation to the category with the highest probability.

- ★ For example, a classifier might take an hourly wage and the number of years of practice of an employee and predict the probability that they are a hospital nurse and the probability that they are an office nurse. The model would then typically return either the probabilities or the category with the highest probability.

Logistic Regression

One approach to classification is logistic regression. Logistic regression is a common way of doing binary classification, which is classifying into two categories. It works by using the logistic function. This function, also called the sigmoid function, is an S-shaped curve that maps input values x to output values y.



Logistic Regression

Logistic regression is similar to linear regression, however the output is not continuous along a line, but a value between 0 and 1. That value can then be interpreted as the probability that the instance belongs to a certain category. To give you an example, consider this sample of the Iris dataset:

SepalLength	SepalWidth	PetalLength	PetalWidth	Species
5.3	3.7	1.5	0.2	Iris-setosa
5	3.3	1.4	0.2	Iris-setosa
7	3.2	4.7	1.4	Iris-versicolor
6.4	3.2	4.5	1.5	Iris-versicolor

Logistic Regression

- ★ Our input values are the length and width of both the petal and sepal. The output will be the probability of the instance of being either a setosa or versicolor type.
- ★ The model will calculate the probability that a sample (consisting of a set of these measurements, or ‘features’) belongs to each category using a simple formula: $p(Species) = \beta 0 + \beta 1 SepalLength + \beta 2 SepalWidth + \beta 3 PetalLength + \beta 4 PetalWidth$
- ★ Let’s say we fit the model and it predicts the following output:

Logistic Regression

setosa	versicolor
0.4	0.6
0.7	0.3
0.2	0.8
0.9	0.1

This table shows that the first observation is predicted with a 40% probability to be of the species setosa and with a 60% probability to be the species versicolor. The prediction for the first instance is thus that it's a member of the category versicolor.

Evaluating Classification

In the tasks on regression, we evaluated our models by looking at the deviation of the predictions from the gold standard and computing the error. This works for continuous outcomes but not for categorical ones. Outcomes of classifiers are instead evaluated in different ways.

Let’s Breathe!

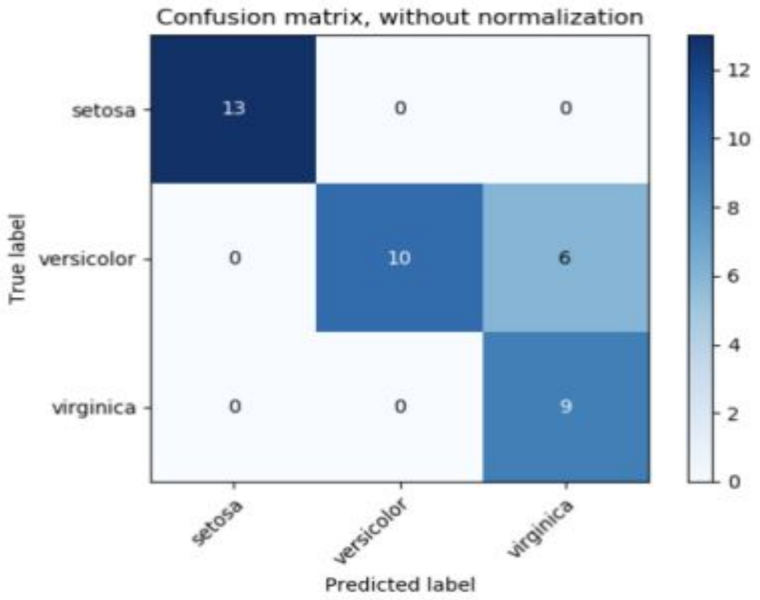
Let’s take a small break before moving on to the next topic.

CoGrammar

Confusion Matrix

A confusion matrix is an NxN table that summarises a classification model's predictions. One axis of a confusion matrix is the label that the model predicted, and the other axis is the actual label. N represents the number of classes. In a binary classification problem N=2. In the complete iris dataset, with three classes, N=3. Here is a sample confusion matrix for the Iris classification problem.

Confusion Matrix



CoGrammar

Confusion Matrix

- ★ The confusion matrix is a great way to see where the weaknesses of the classifier lie. In this example, we see that the model never classifies setosa instances as versicolor or vice-versa, but on 6 occasions it classified versicolor instances as virginica instances, suggesting that those two types are, in the eyes of the system, harder to distinguish.

- ★ While the confusion matrix gives insights into the behaviour of the classifier, we need evaluation metrics to make claims about whether the model did well or not compared to other models. To understand the most common evaluation metrics, let’s go over a story.

Confusion Matrix

A shop owner employs a security guard to keep a lookout for shoplifters. If the guard sees someone slip any merchandise into their pockets, he approaches the person he suspects of shoplifting, retrieves the stolen items, and escorts the shoplifter out of the shop. Unfortunately, looks can be deceptive and the security

guard sometimes makes mistakes and confronts people who may have only been putting their mobile phone back in their pocket. This makes innocent customers very upset and the shop owner gets angry. Practised shoplifters also sometimes deceive the security guard and slip past him unnoticed with stolen goods, which also makes the shop owner angry as he is losing money. The security guard lets regular customers who do not shoplift pass peacefully out of the shop.

Confusion Matrix

This story highlights two types of error the security guard can make; accusing an innocent person of shoplifting and missing an actual shoplifting event. If we state that:

- **“Shoplifter” is the positive class**
- **“Not a shoplifter” is the negative class**

We can summarise our "shoplifting prediction" model using a 2x2 confusion matrix with the following four possible outcomes:

CoGrammar

Confusion Matrix

True Positive (TP):
- Reality: A person stole an item
- Guard said: "Shoplifter."
- Outcome: The owner is happy

False Positive (FP):
- Reality: Regular customer, not stealing
- Guard said: "Shoplifter."
- Outcome: The owner is angry

False Negative (FN):
- Reality: A person stole an item
- Guard said: "Not a shoplifter."
- Outcome: The owner is angry

True Positive (TP):
- Reality: Regular customer, not stealing
- Guard said: "Not a shoplifter."
- Outcome: The owner is happy

Confusion Matrix

★ Correct predictions occur both for true positive predictions, where the model correctly predicts a positive (shoplifting) event, and true negative predictions, where the model correctly predicts a negative event (not shoplifting).

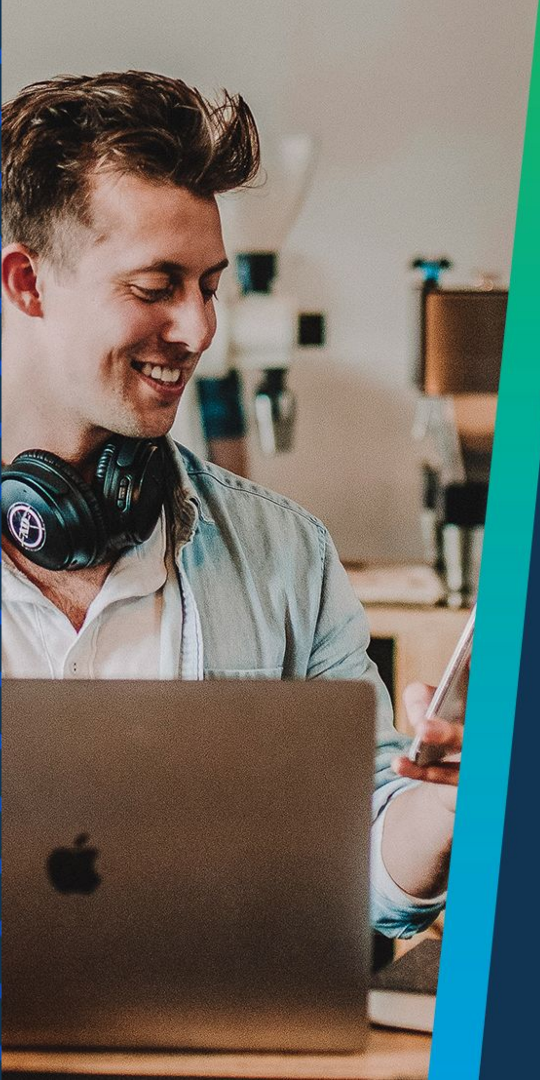
★ Incorrect predictions occur when the model predicts a false positive or a false negative. False-positive predictions occur when the model predicts the positive class, but the reality is the negative class — i.e., a regular customer is accused of shoplifting. False-negative predictions occur when the model predicts the negative class, but the reality is a positive class — i.e., the guard thinks a shoplifter is a regular customer and lets them pass.

Confusion Matrix

Correct predictions occur both for true positive predictions, where the model correctly predicts a positive (shoplifting) event, and true negative predictions, where the model correctly predicts a negative event (not shoplifting). Incorrect predictions occur when the model predicts a false positive or a false negative. False-positive predictions occur when the model predicts the positive class, but the reality is the negative class — i.e., a regular customer is accused of shoplifting. False-negative predictions occur when the model predicts the negative class, but the reality is a positive class — i.e., the guard thinks a shoplifter is a regular customer and lets them pass.

Q & A SECTION

Please use this time to ask any questions relating to the topic, should you have any.



CoGrammar

Thank you for joining us

1. Take regular breaks
2. Stay hydrated
3. Avoid prolonged screen time
4. Practise good posture
5. Get regular exercise

"With great power comes great responsibility"