DataLab Preparation (Week 7, DataLab I, Thursday)

**2.1a How does LIME generate explanations for model predictions?**

By training a global surrogate model to approximate the predictions of the original model.

By modifying a single input sample and observing changes in the model's output.

By extracting and analyzing the weights of neural networks directly.

[x]By perturbing the input data and training local surrogate models around the prediction.

**2.1b Which of the following statements best describes the applicability of LIME?**

LIME can only be applied to deep learning models.

LIME is specifically designed for regression models and cannot be used with classification models.

[x]LIME is model-agnostic, meaning it can be used to explain the predictions of any machine learning model.

LIME requires modifications to the model architecture to generate explanations.

**2.2c We will use LIME to understand the reasoning of our cats vs. dogs classifier. Follow the steps provided to get started:**

**Choose two different cat images; one of a cat that was classified correctly as a 'cat', and one of a cat that was misclassified as a 'dog'.**

**Use LIME to generate explanations for each image.**

**Compare and contrast the key areas highlighted by LIME in each case.**

**Now, answer the following questions:**

**Are there major differences in the features LIME emphasizes for the correctly classified cat versus the misclassified one?**

**For each cat, can you explain why LIME highlights the features it does? Consider how those features might lead to a 'cat' classification or a potential mix-up.**

**Were there any surprising differences between the LIME explanations? Did this reveal anything about how the model might be biased or prone to errors?**

Wrongly classified:

A black cat lying on a blanket

Description automatically generated

This is the LIME explanation for the mislabelled image:

A black and white image of a map

Description automatically generated

LIME most likely highlights these features because of the cat's colour and the background.

 Correctly classified:

A cat sitting on a window sill

Description automatically generated

This is the LIME explanation for the correctly classified image

A cat with a black background

Description automatically generated

LIME most likely highlights these features because the weights of the pixels were the highest, i.e., these groupings are the most relevant when classifying a cat image.

The differences are quite noticeable in the images. In the wrongly classified instance, the model could not make out the 'cat' because of its colour, which led to confusions. In the correct instance, the cat was clearly noticeable.

These results, although useful, were not particularly shocking. A heavy background, poor lighting, etc. can lead to misclassifications.

**2.2b In this exercise, we will use SHAP to understand the reasoning of our cats vs. dogs classifier. Follow the steps provided to get started:**

**Choose two different images; one of a cat, and one of a dog.**

**Use SHAP to generate explanations for each image.**

**Compare and contrast the key areas highlighted by SHAP in each case.**

**Now, answer the following questions:**

**Can you identify patterns in the regions that are important for different classes?**

**Does this visualization help you understand how the model makes its decisions?**

For the cat image, the regions that are most important are the nose, eyes, ears, and the shape of the face. For the dog image, these are the same, including the shape of the body itself.

The visualisations help me gain insight into how the model makes its decisions because they 'debunk' the blackbox of AI. Additionally, they help me make informed decisions when improving my model.

**2.3b Explain the famous saying: 'Correlation does not imply causation'. Provide an example of how this concept could lead to a possible misinterpretations in our cats vs. dogs image classification task. Write your answer down.**

Sometimes we make connections between certain situations that might seem like they are influencing each other, or together influencing an outcome, based solely on the fact they might have a common denominator. Yet these connections might not be actually relevant to the problem at hand due to the underlying context of said issue. As an example, in our cats vs dogs image classification task this could mean that just because a certain breed of cat tends to be photographed more frequently in outdoor settings, it doesn't necessarily mean that the presence of outdoor scenery in an image is a reliable indicator of whether the image contains a cat or a dog.