DS-UA 202, Responsible Data Science, Spring 2022

Homework 3: Transparency and Interpretability
Due at 11:59pm on Thursday, May 5, 2022

Objectives

This assignment consists of written problems and programming exercises on transparency and explanations.

The programming part of this assignment focuses on explaining classifiers, and on doing feature selection based on a set of explanations. You will use the open-source SHAP library to complete the programming portion of this assignment. We encourage you to carefully read the paper describing SHAP, which we discussed in class.

After completing this assignment, you will:

1. Understand how to use SHAP to generate locally interpretable explanations of classification decisions on a text corpus.
2. Learn how to use local explanations for feature selection, ultimately improving classification accuracy on a text corpus.
3. Understand the relationship between transparency and data protection, by reasoning about the benefits and the risks of automated decision making to different stakeholders.

You must work on this assignment individually. If you wish to clarify any parts of this assignment, please post on the Discussions forum on Brightspace. For any other questions, please email all instructors.

Grading

The homework is worth 65 points, or 10% of the course grade. Your grade for the programming portion will be significantly impacted by the quality of your written report for that portion. In your report, you should explain your observations carefully.

You are allotted 2 (two) late days over the term, which you may use on a single homework, or on two homeworks, or not at all. If an assignment is submitted at most 24 hours late: one day is used in full. If it's submitted between 24 and 48 hours late: two days are used in full.

Submission instructions

Provide written answers to Problems 1, 2, and 3 in a **single PDF file**. (It is recommended that you Google Docs to prepare this PDF, but you may instead use Word or LaTeX). Provide code in answer to Problem 3 in a **Google Colaboratory notebook**. Both the PDF and the notebook

should be turned in as Homework 3 on Brightspace. Please clearly label each part of each question.

Problem 1 (10 points): Online Job Ads

Consider a hypothetical job search website that uses a machine learning system to determine which job openings to show to which users. The system uses historical employment data, and also collects interaction data from its own users: which users were shown which job openings, and which users clicked. The service also receives data from employers, detailing which users were invited to job interviews, and which were hired.

(a) **(6 points)** Give **three distinct reasons** why gender disparities might arise in the operations of such a system.

(b) **(4 points)** Suppose that the job search service decides to increase the number of times it presents job openings in STEM to women. To do so, the service observes that STEM job experience (in years) is **positively associated** with the likelihood that a user clicks on an advertised STEM job opening: the more years of experience, the more likely a user is to click. Consider the following intervention:

> *Pre-process the training dataset, replacing the value of the "job experience" feature for women with the best (highest) possible value for the feature in the dataset*.

(i) Under what conditions will this intervention **increase** the number of times job openings in STEM are shown to women?

(ii) Under what conditions will this intervention **fail to increase** the number of times job openings in STEM are shown to women?

Problem 2 (20 points): AI Ethics: Global Perspectives

In this part of the assignment, you will watch a lecture from the AI Ethics: Global Perspectives course and write a memo (500 words maximum) reflecting on issues raised in the lecture. You can watch either:

- "Content Moderation in Social Media and AI" (watch the lecture)
- "Indigenous Data Sovereignty" (watch the lecture)
- "Do Carebots Care? The Ethics of Social Robots in Care Settings" (watch the lecture)
- "The Intersection of AI and Consumer Protection" (watch the lecture)

If you have not already registered, please register for the course at

https://aiethicscourse.org/contact.html, specify “student” as your position/title, “New York University” as your organization, and enter the course number, DS-UA 202, in the message box.

In your memo, you should discuss the following:

- Identify the stakeholders. In particular, which organizations, populations, or groups could be impacted by the data science issues discussed in the lecture? How could the data science application benefit the population(s) or group(s)? How could the population(s) or group(s) be adversely affected?
- Identify and describe an issue relating to data protection or data sharing raised in the lecture.
    - Which vendor(s) owns the data and/or determines how the data is shared or used?
    - To what extent is the privacy of users or persons represented in the data being protected? Is the data protection adequate?
- How does transparency and interpretability, or a lack thereof, affect users or other stakeholders? Are there black boxes?
- What incentives does the vendor (e.g., the data owner, company, or platform) have to ensure data protection, transparency, and fairness? How do these incentives shape the vendor's behavior?

Problem 3 (35 points): Generating Explanations with SHAP

For the programming portion of this assignment, we will use a subset of the text corpus from the 20 newsgroups dataset. This is the dataset used in the LIME paper to generate the Christianity/Atheism classifier, and to illustrate the concepts. However, rather than explaining predictions of a classifier with LIME, we will use this dataset to explain predictions with SHAP.

For guidance, you can use the HW3_Template here. Please make sure to duplicate this file rather than write your code directly here.

(a) **(5 points)** Use the provided Colab template notebook to import the 20 newsgroups dataset from sklearn.datasets, importing the same two-class subset as was used in the LIME paper: Atheism and Christianity. Use the provided code to fetch the data and split it into training and test sets. Then, fit a TF-IDF vectorizer to the data, and train a SGDClassifier classifier.

(b) **(10 points)** Generate a confusion matrix (hint: use sklearn.metrics.confusion_matrix) to evaluate the accuracy of the classifier. The confusion matrix should contain a count of correct Christian, correct Atheist, incorrect Christian, and incorrect Atheist predictions

from your SGDClassifier. Use SHAP's explainer to generate visual explanations for any 5 documents in the test set. The documents you select should include some correctly classified and some misclassified documents.

(c) **(20 points)** Use SHAP's explainer to study mis-classified documents, and the features (words) that contributed to their misclassification, by taking the following steps:

- Report the **accuracy** of the classifier, as well as the **number of misclassified documents**.

- For a document **doc_i** let us denote by **conf_i** the difference between the probabilities of the two predicted classes for that document. Generate a plot that shows **conf_i** for all misclassified documents (which, for misclassified documents, represents the magnitude of the error). Use any chart type you find appropriate to give a good sense of the distribution of errors.

- Identify all words that contributed to the misclassification of documents. (Naturally, some words will be implicated for multiple documents.) For each word (call it **word_j**), compute (a) the number of documents it helped misclassify (call is **count_j**) and (b) the total weight of that word in all documents it helped misclassify (**weight_j**) (sum of absolute values of **weight_j** for each misclassified document). The reason to use absolute values is that SHAP assigns a positive or a negative sign to **weight_j** depending on the class to which **word_j** is contributing. Plot the distribution of **count_j** and **weight_j**, and **discuss** your observations in the report.