**hq8314**

**Mostafa Sameer**

**Question 1.** *(10 points) Describe one problem in your study area that can be solved by machine learning techniques. Classify your problem in terms of supervised or unsupervised, classification or regression. Explain the unique challenges to the standard machine learning methods, such as logistic regression.*

Answer: Sentiment analysis is a challenge in the field of natural language processing (NLP) that can be resolved using machine learning methods. Sentiment analysis includes identifying the sentiment or emotional tone indicated in a text, such as a social networking post, a product review, or a news item. This issue falls within the categorization category and can be categorized as supervised learning.

Classification of the issue:

Supervised Learning: In sentiment analysis, supervised learning techniques are utilized since they call for labeled training data, where each text sample is assigned a sentiment label (such as positive, negative, or neutral).

Challenges and Particular Challenges to Traditional Machine Learning Methods:

Complex Contextual Understanding: Understanding the complex context in which words and phrases are employed is frequently necessary for sentiment analysis. Simple word-based features or bag-of-words representations are frequently used as the basis of common machine-learning techniques like logistic regression. However, they find it difficult to represent the nuanced interactions between words and the wide range of emotions conveyed in various circumstances. By considering the sequential structure of text and learning hierarchical representations of language, deep learning models, such as recurrent neural networks (RNNs) and transformers, have shown to be more successful at capturing these subtleties. Handling Sarcasm and Negation: Sentiment analysis must contend with difficulties like sarcasm and negation (e.g., "Oh, that's just great!" and "not good"). Due to the way that traditional machine learning approaches consider individual words, they are unable to handle these situations well.

Deep learning models are especially suited for dealing with these kinds of variables because they can capture long-range dependencies and comprehend how words interact with one another, particularly those based on transformers. Sentiment analysis datasets are frequently unbalanced, with an excessive quantity of positive or negative samples. The majority class may be favored by conventional categorization approaches. To handle the imbalance of classes, which is essential for ensuring the model's correctness, multiple strategies must be used, such as oversampling, under-sampling, or employing various loss functions. Deep learning models may also be improved to solve this problem successfully. Sentiment analysis across languages may be difficult using conventional methods for machine learning.

Multilingual transformers, a type of deep learning model, have demonstrated the capacity to perform effectively across a variety of languages without requiring considerable feature engineering or language-specific prepossessing. Interpretation Findings Although deep learning models can perform sentiment analysis with great accuracy, they can also be referred to as "black box" models, which makes it difficult to understand why a certain sentiment prediction was generated. This is particularly crucial in situations where readability and comprehension are vital, like the legal or medical fields. Using feature values, standard logistic regression methods provide simpler interoperability. In conclusion, while typical machine learning techniques like logistic regression may be employed for sentiment analysis, they have difficulties managing complex verbal occurrences and capturing nuanced context.

Modern state-of-the-art outcomes in sentiment analysis tasks have been attained by deep learning models, notably, those built on transformers, which have demonstrated considerable advances in resolving these issues. In contrast to conventional approaches, they could need more data and computer resources for training, and they might also be harder to comprehend. The needs and limitations of the application will determine which strategy is best.

**Question 2.** *(10 points) Regarding the squared loss function we covered in the lecture, we sum up the squares of differences between the actual value and the estimated value from the model output. This squared loss function is the one most frequently used for continuous output, but it can also be used for classification tasks. Since it sums up the squares of the differences, it is not robust to outliers. Propose a more robust loss function for classification tasks. Please define all mathematical notations clearly.*

Answer: The discrete character of class labels makes the squared loss function inappropriate for classification tasks, and it is also not resilient to outliers. Cross-entropy loss, usually referred to as log loss or logistic loss, is a new sort of loss function that is frequently used for classification tasks. This loss function, which is reliable for classification, is intended to quantify the discrepancy between actual class labels and expected class probabilities.

Following is a definition of the cross-entropy loss for binary classification:

Let,

* y be the true binary class label (0 or 1).
* p(y) is the predicted probability that the instance belongs to class y.

L(y,p(y))=−[y⋅log(p(y))+(1−y)⋅log(1−p(y))]

A dataset's overall loss is calculated as the average of all these instance losses.

*For classification problems, this cross-entropy loss function is appropriate since it:*

1. *0 or 1 discrete class labels are handled naturally.*
2. *assists the model in generating accurate probability estimates.*
3. *imposes severe penalties on the model when it firmly predicts the incorrect class, making it resilient to outliers in terms of prediction confidence*.

The cross-entropy loss is expanded to handle more than two classes for multi-class classification problems, often with a "softmax" activation function to calculate class probabilities. Considering the projected class probabilities, the loss is thus the negative log-likelihood of the true class.

In conclusion, it is recommended to employ cross-entropy loss rather than the squared loss when dealing with classification jobs since it is made expressly for classification issues and is more resilient to outliers in the context of classification.

**See the attached files for Ques 3, 4, 5.**