Intuition Report 1

Ramtin Mojtahedi Saffari - 20307293

Assessed link 1: A visual introduction to machine learning

Self-reflection

- **Article Short Summary**

The provided article discusses main machine learning (ML) concepts and rationales in an easy-to-understand and straightforward way. These concepts are being introduced in a home classification problem between the cities of San Francisco and New York. While the issue is being explored, the article expounds on concepts of model training, features (predictors or variables), false positive, false negative, dataset split, recursion, prediction, accuracy, and precision. By providing explnations in training and testing an ML model with decision tree as the classifier, the article sheds light on the mentioned ML major concepts.

- **Hypothesis and Expectation**

Considering the accuracy results achieved for the training dataset, I hypothesize that growing and making a deeper decision tree model results in better outcomes for training and test data sets. Therefore, it is expected is to exist a direct relationship between results in training and test data set.

- **What I Achieved**

Although the accuracy increased and reached out to 100% by adding more layers to the decision tree model (making a deeper model), the accuracy on the unseen(test) data set couldn’t reach out to the same accuracy in the training data set, and it reached to only 89.7%. This contradicts the proposed hypothesis as the test data’s accuracy is less than the training data set.

- **What I Learned**

I learned that overfitting is an important factor that can reduce the performance of the proposed ML model on the test data set and caused the trained decision tree model to not perform well on the unseen data. Overfitting happens when the model tightly fits all samples in the given training data to a great extent that it would be inaccurate in predicting the outcomes of the unseen and test data. This also prevents building a generalized ML model. Thus it ends up with branches with strict rules of sparse data, and the model learns the noisy data rather than focusing on the important patterns of data.

To address the overfitting issue, there are many methods, which I found very interesting as written in [1]. This should be added to the article to fill some gaps in understanding the concept of overfitting. As I explored, the effective techniques to prevent overfitting in decision tree models are pre-pruning, post-pruning, and using the Ensemble-Random Forest model. In the selected article, the decision tree model is trained to its full length and branches, tempting the model for overfitting. One of the common techniques is pre-pruning. In this way, the model is not trained to its full length to stop generating non-significant tree branches. This can be done by controlling the model’s hyperparameters of max_depth, min_samples_leaf, min_samples_split.

On the other hand, in the post-pruning technique full tree is generated, and then the non-significant branches are removed. This is done by controlling the Cost Complexity Pruning (CCP) considered in the ccp_apha values. Alongside these two techniques, the other method to prevent overfitting is using Random forest(can be implemented by Scikit-Learn library). In this bagging (bootstrap aggregation technique), multiple decision tree models are being built and merged their predictions to get a more accurate and stable prediction. This helps in the reduction of variance, which helps prevent overfitting. In this essence, we can control the number of used trees (learners) by modifying the value of n_estimator.

As one of the article’s gaps, it is recommended to provide a comprehensive explanation of effective techniques in overfitting prevention. Here is a recommended link for useful strategies to prevent overfitting, which can be considered for different models, not specifically the proposed decision tree model in the assessed article [2]. The information can be added to the article to make it comprehensive and fill some related gaps. Overall, I can enumerate the strength and weak points of the assessed article as follows:

Strengths

- Visualization graphs
- Easy-understanding context
- Providing a straightforward and real-world example

Weakness

- Shallow explanation of the concepts
- Inadequate explanation for addressing the overfitting issue

- Inadequate conclusion

Assessed link 2: This person doesn’t exist

Self-reflection

- **Short Summary**

The website generates fake face images using style-based generative adversarial networks (StyleGAN2) to generate the fake images. GAN is an algorithmic architecture consisting of two neural networks (generator and discriminator) that pits against each other to generate fake and synthetic data/images to pass on the real data/images. The provided link uses StyleGAN2, an improved version of StyleGAN that uses baseline progressive GAN architecture. Progressive GAN is an extension to GAN, which gradually increases the size of generated images from a very low resolution (e.g., 4×4) to high resolution (e.g., 1024 x1024). On the other hand, StyleGAN2 was redesigned in its progressive growth with a new regularization technique to improve conditional generation without changing the network’s architecture and topology. This allows network design with great depth and good training stability and secures that the outcome of the proposed network will be in a higher resolution and realistic images compared to the StyleGAN [3].

- **Hypothesis**

Considering the generated images, it is hypothesized that these fake images can also be recognizable with eyes without using a discriminator network by considering some defects in the visual features of images

- **Testing the Hypothesize and Explain What I Have Learned**

Although StyleGAN2 has improved the generated images' quality, there are a number of visual defects in the generated images. Through assessment of more than 150 images, it was explored that those visible defects can be detectable as follows:

- Defects in the background or texts in the background (most frequent deficiency)
- Flaws in the symmetry of head components (ears, eyes, eyebrows, etc.), upper body, or clothes shape and color (painterly rendering)
- Defects in the form of teeth
- Defects in the hair (messy hair)
- Defects in the earrings and jewelry

One of the problems with the background is that if there is a complex and noisy background, the background is being generated with some levels of defects. An example of this is when the person has long hairs or hairs covering some parts. For the other kinds of defects, there is a need for a high level of punctiliousness, and the proposed model can generate close to flawless and realistic images.

The StyleGAN2 requires lots of computational resources, and the later studies focused on how to decrease complexity in computationally and network parameters. An example of an improved model can be found in [4], which provides x3.5 fewer parameters and is x9.5 less computationally complex than StyleGAN2. Also, It is recommended that the website add more explanations or easy-understandable information to improve viewers' interaction with the website, such as using some general information that is available in [5].

Self-evaluation:

In this intuition report, I have gone through an in-depth analysis of two of the provided links. In my assessment, I have completely considered the required expectation for deep exploration, including proposing a hypothesis and what I expected, reporting on what I achieved and explored, in-depth discussion of what I have learned, and providing gaps and some recommendations to fill them. Considering the quality and assessment level, I deserve to get the full mark (4 points) for this intuition report.

In advance, thank you very much for your time and consideration of this report.

Best regards,

Ramtin

References

[1] Fumo, D. (2018, June 21). Types of Machine Learning Algorithms You Should Know. Medium. https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861

[2] Chuan-En Lin, D. (2020, July 2). *8 Simple Techniques to Prevent Overfitting - Towards Data Science*. Medium. https://towardsdatascience.com/8-simple-techniques-to-prevent-overfitting-4d443da2ef7d

[3] Singh, P. C. (2021, February 16). *Understanding the StyleGAN and StyleGAN2 Architecture*. Medium. https://medium.com/analytics-vidhya/understanding-the-stylegan-and-stylegan2-architecture-add9e992747d

[4] *MobileStyleGAN: A Lightweight Convolutional Neural Network for High-Fidelity Image Synthesis*. (2021, April 10). MobileStyleGAN: A Lightweight Convolutional Neural Network for High-Fidelity Image Synthesis. https://paperswithcode.com/paper/mobilestylegan-a-lightweight-convolutional

[5] *This person does not exist: AI generates fake faces on the website*. (2019, February 15). CTVNews. https://www.ctvnews.ca/sci-tech/this-person-does-not-exist-ai-generates-fake-faces-on-website-1.4299515