Summary of ”Feature selection, L1 vs. L2 regularization” by Andrew Y. Ng

Isaac Akintaro

11/11/2023

1 Introduction

This summary presents key insights from Andrew Y. Ng's paper on the comparison of L1 and L2 regularization methods in the context of logistic regression, particularly in scenarios involving a high number of irrelevant features. I think about it's application to my Aylesbury Estate project.

2 Key Terms

- Sample Complexity - the number of training examples (data points) that an algorithm needs to generalise well on unseen data.
- Loss or Cost function - measures how well a specific model performs. It does this by quantifying the difference between the predicted values by the model and the actual values of the data it's trying to predict or classify.
- Regularization - this technique involves adding a penalty to the loss function to prevent the model from becoming too complex, which can lead to overfitting. Overfitting happens when a model memorizes the training data too well, impairing its performance on unseen data.
- Coefficient - indicates the direction and strength of the relationship between a feature and the prediction.
- L1 Regularization (Lasso Regularization) - this type of regularization adds a penalty to the model's loss function based on the absolute value of the coefficients. The L1 regularization penalty is given by $\lambda \sum_{i=1}^{n} |\beta_i|$, where λ is the regularization parameter, n is the number of features, and β_i are the coefficients of the model for feature i. This approach can force some coefficients to become zero, meaning it effectively eliminates some features from influencing the model's outcome. Therefore, with L1 Regularization, only the most useful features are retained, enhancing the model's ability to generalize and perform well on unseen data.

- L2 Regularization (Ridge Regularization) - this regularization technique adds a penalty to the model's loss function based on the square of the coefficients. The L2 regularization penalty is formulated as $\lambda \sum_{i=1}^{n} \beta_i^2$, where λ is the regularization parameter, n is the number of features, and β_i are the coefficients of the model for feature i. Unlike L1 regularization, L2 does not reduce the coefficients to zero but rather shrinks them closer to zero. This diminishes the role of less influential features in the model's predictions, helping to reduce the likelihood of overfitting and improving the model's generalizability without completely discarding any features.
- Rotationally Invariant Algorithm - an algorithmn where if you transform the input features of the dataset by rotating them (imagine spinning the axes of a graph around the origin), a rotationally invariant algorithm would still make the same predictions as it would on the unrotated data. This characteristic is due to the mathematical properties of the algorithm. L1 is not rotationally invariant.

3 L1 Regularization

- L1 regularized logistic regression exhibits a sample complexity that grows logarithmically with the number of irrelevant features.
- This method is effective in high-dimensional input spaces, handling exponentially many irrelevant features compared to the number of training examples.
- The approach involves solving a convex optimization problem and demonstrates efficiency even in large input spaces.

4 L2 Regularization and Rotational Invariance

- L2 regularization and its rotational invariance in algorithms like logistic regression, SVMs, and neural networks are discussed.
- The sample complexity for rotationally invariant algorithms grows linearly with the number of irrelevant features, which is less efficient in high-dimensional settings.

5 Empirical Comparisons and Experiments

- Empirical results highlight that logistic regression with L1 regularization outperforms L2 regularization in the presence of many irrelevant features.
- L1 regularization exhibits less sensitivity to irrelevant features, thereby enhancing performance in high-dimensional spaces.
- Next steps:

6 Theoretical Analysis

- The paper provides a detailed theoretical analysis, using concepts like covering numbers and properties of logistic regression models.
- These analyses support the superiority of L1 regularization over L2 in high-dimensional data scenarios.

7 Applications to Aylesbury Estate Project and Future Projects

- Feature Selection and Regularization - selecting the most relevant features and avoiding overfitting are crucial. If I have a model with many irrelevant features it is good to know L1 regularization would be more effective for feature selection.
- Handling High-Dimensional Data - L1 would enhance my model's performance by reducing the impact of less informative features.
- Model Complexity and Sample Size - using L1 regularization might help in building a more generalisable model without the need for an extensively large dataset
- Next steps - experiment with no regularization, L1, L2 and see what performs best for my data.

8 Conclusion

The paper underscores the advantages of L1 regularization in logistic regression, especially in contexts with a large number of irrelevant features. It combines theoretical and empirical evidence to establish the effectiveness of L1 regularization over L2 in such scenarios.