Partial Dependence Plots (PDP) & Individual Conditional Expectation (ICE) Research Report

Xin Wei a1912958
RAIL-PG-2

1. Introduction

With the rapid development of machine learning and artificial intelligence across critical fields like healthcare, finance, transportation, and industrial safety, predictive models are achieving remarkable accuracy. However, many of these models—particularly ensemble methods and deep neural networks—remain opaque, making them difficult to interpret. In high-risk applications, such as medical diagnoses, financial loan approvals, and railway failure prediction, users and stakeholders require not only accurate predictions but also an understanding of why a particular prediction was made. This necessity for transparency drives the growing importance of explainable AI (XAI).

Among the various XAI techniques, Partial Dependence Plots (PDP) and Individual Conditional Expectation (ICE) are two popular and effective methods for visualizing the relationship between input features and model predictions. PDP provides insights into the average effect of features on predictions, while ICE highlights how individual instances are affected by feature changes. Together, they offer complementary views—global and local—of model behavior.

This report introduces the theoretical foundations of PDP and ICE, explores their applicability to different model types, evaluates their advantages and limitations, and demonstrates their complementary role in enhancing model transparency and trust.

2. Method Overview: PDP & ICE

2.1 Partial Dependence Plots (PDP)

PDP is a method used to understand the average marginal effect of one or two features on the model's predictions. It works by varying the value of the target feature(s) while keeping all other features fixed and averaging the model's predicted outcomes across the entire dataset. The resulting plot typically displays the relationship between the feature(s) and the predicted output as a curve (for one feature) or a heatmap (for two features).

For example, in a credit risk prediction model, PDP can be used to visualize how changes in loan amount impact the probability of default, averaged over all samples. If the PDP curve shows a steady increase, it suggests that, on average, higher loan amounts lead to higher default probabilities.

PDP is especially useful for identifying global patterns such as monotonicity or non-linear effects. However, because it averages across all instances, it can obscure individual differences.

2.2 Individual Conditional Expectation (ICE)

Individual Conditional Expectation (ICE) extends the idea of PDP by focusing on individual-level behavior. Instead of providing a global average, ICE calculates the predicted output for each individual instance as the target feature(s) vary, holding all other features constant. This produces one curve for each sample, illustrating how the model's prediction changes for that specific instance as the feature of interest changes.

By overlaying multiple ICE curves, we can see how different samples respond to the same feature change, revealing heterogeneity in model behavior that PDP might obscure. For example, while PDP may show a monotonic increase in default probability with loan amount, ICE might uncover that for certain individuals, the probability decreases, reflecting complex interactions with other features.

The key strength of ICE lies in its ability to capture hidden variation and provide more granular insights into model behavior.

3. Applicability

3.1 Applicable Models

Both PDP and ICE are model-agnostic techniques, meaning they can be applied to any machine learning model as long as the model can generate predictions. This makes them highly flexible and suitable for a variety of models, including:

- Ensemble models such as Random Forests, Gradient Boosting Machines, and XGBoost;
- Kernel-based models such as Support Vector Machines (SVMs);
- Deep learning models such as fully connected neural networks or Transformer-based models.

However, their usability can be limited in high-dimensional settings. When the number of features increases significantly, visualizing the interactions between them becomes more difficult, and PDP and ICE can become less informative. Therefore, these methods are most effective for analyzing one or two features at a time.

3.2 Applicable Scenarios

PDP and ICE are particularly useful in the following scenarios:

- Feature influence analysis: Assessing the overall effect of one or two features on model predictions;
- Trend exploration: Identifying monotonic or non-linear relationships between features and predictions;
- Interaction effect analysis: Investigating whether interactions between features influence predictions, using two-dimensional PDPs or comparing ICE curves.

These methods are particularly valuable when models are used in decision-critical applications. For instance, in medical risk prediction, PDP might show the overall trend of increasing disease risk with age, while ICE could reveal how this risk varies significantly across different patient profiles.

4. Advantages and Limitations

4.1 Advantages

- Interpretability: PDP and ICE are visual tools that make complex model behavior more accessible and understandable for non-experts, facilitating communication between data scientists and domain experts.
- Model-agnostic: These methods can be applied to virtually any predictive model, making them highly versatile across various machine learning tasks.
- Complementarity: While PDP reveals global patterns (e.g., the average effect of a feature), ICE uncovers individual-level variation, making these methods complementary in understanding both general trends and sample-specific effects.

4.2 Limitations

- Independence assumption: PDP assumes features are independent. In real-world datasets, strong correlations among features may distort results and lead to misleading interpretations.
- High-dimensional challenges: PDP and ICE are most effective for visualizing relationships between one or two features. When applied to high-dimensional data, these methods struggle to provide meaningful insights without dimensionality reduction.

- Computational cost: Generating ICE curves for large datasets requires significant computational resources, especially when visualizing multiple individual curves.
- Ambiguity in interpretation: Different feature distributions or ranges can lead to varying patterns in PDP and ICE plots, necessitating cautious interpretation.

5. Conclusion

PDP and ICE provide two complementary methods for interpreting machine learning models. PDP is ideal for understanding global trends by visualizing the average effect of features on model predictions, while ICE offers more detailed, sample-specific insights, revealing how individual instances respond to changes in feature values. Together, these methods help improve model transparency and build trust with stakeholders by providing clear explanations of model behavior.

Despite their advantages, PDP and ICE have limitations, especially when dealing with high-dimensional data or highly correlated features. These challenges highlight the need for careful application of these methods, often in conjunction with domain knowledge.

Future work may involve integrating PDP and ICE with other advanced interpretability techniques, such as Accumulated Local Effects (ALE) or counterfactual explanations, to create more comprehensive frameworks for understanding model behavior. These advances will continue to drive the deployment of machine learning models in high-stakes fields, ensuring that they are not only accurate but also interpretable and trustworthy.

6. References

Friedman, JH 2001, ‘Greedy function approximation: A gradient boosting machine’, The Annals of Statistics, vol. 29, no. 5, pp. 1189–1232.

Goldstein, A, Kapelner, A, Bleich, J & Pitkin, E 2015, ‘Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation’, *Journal of Computational and Graphical Statistics*, vol. 24, no. 1, pp. 44–65.

Muschalik, M, Fumagalli, F, Jagtani, R, Hammer, B & Hüllermeier, E 2023, ‘iPDP: On Partial Dependence Plots in Dynamic Modeling Scenarios’, in L Longo (ed.), *Explainable Artificial Intelligence*, vol. 1901, Springer, Switzerland, pp. 177–194.

Wood, D, Papamarkou, T, Benatan, M & Allmendinger, R 2024, ‘Model-agnostic variable importance for predictive uncertainty: an entropy-based approach’, *Data Mining and Knowledge Discovery*, vol. 38, no. 6, pp. 4184–4216.

Mehdiyev, N, Majlatow, M & Fettke, P 2024, ‘Communicating Uncertainty in Machine Learning Explanations: A Visualization Analytics Approach for Predictive Process Monitoring’, in S Lapuschkin, C Seifert & L Longo (eds), *Explainable Artificial Intelligence*, vol. 2155, Springer, Switzerland, pp. 420–438.