P-15

住宅価格予測ニューラルネットワークにおける特徴量とハイパーパラメータの最適化

Optimization of Features and Hyperparameters for a Neural Network in House Price Prediction

○小林晴 [1], 青柳美輝 [2]
Haru Kobayashi[1], Miki Aoyagi[2]

Abstract: In this paper, we built a house price prediction model using a neural network. Through feature engineering and hyperparameter optimization, the enhancement of prediction accuracy together with the reduction of computational cost demonstrated its effectiveness

1. はじめに

住宅価格の予測は、立地や築年数といった多様な変数が複雑かつ非線形に影響するため、長年の課題とされてきた。近年、この課題を解決するアプローチとして、変数間の複雑な関係性を自動で学習するニューラルネットワークの活用が期待されている。本稿は、ニューラルネットワークを用いた住宅価格予測モデルを構築し、その精度と有効性を検証することを目的とする。

2. 実験手法

本研究では、ニューラルネットワークを用いて住宅価格を予測するモデルを構築し、その精度向上を目的とした実験を行った。実験は、データの前処理、モデル構築、そして独自の精度向上プロセスという手順で進めた。

2.1　使用データと前処理

本実験では、[アットホームデータセット] から取得した香川県の不動産賃貸データを用いた。このデータは価格が 4 万円から 5 万円台の物件が中心である。目的変数を「月額賃料」、説明変数を物件スペックに関する各種データとした。以下のデータを入力特徴量として使用した。
物件情報: 「築年」「専有・建物面積（平米）」「設備数」
立地情報: 「徒歩 (分)」「バス (分)」「停歩 (分)」「所在階」
地理座標: 「緯度」「経度」
費用情報: 「管理費（円）」「共益費（円）」「駐車場料金（円）」
カテゴリカル特徴量において処理をしたデータには、以下の 2 手法を適用した。

One-Hot エンコーディング: カテゴリ間に順序のない「所在地名 1」「建物構造」に One-Hot エンコーディングを適用し、独立した二値特徴量へ変換した。

ターゲットエンコーディング: 「間取りタイプ」には、各カテゴリの価格の中央値で数値を置き換えるターゲットエンコーディングを適用した。

その他の前処理：上記の過程で生じた欠損値は全て 0 で補完した。また、学習の安定化を目的として、全特徴量に Z スコア標準化を適用した（節 2.3 にて説明する）訓練データとテストデータの分割時には、元のデータとの対応を維持するためインデックスも同時に分割した。

2.2　モデルアーキテクチャ

本研究で用いた予測モデルには、PyTorch を用いて多層パーセプトロン（MLP）を構築した。このモデルのアーキテクチャは、実験を通じて一貫して固定されており、変更は加えていない。モデルは、入力層、1 つの中間層、出力層から成るシンプルな三層構造である。データの流れは以下の通りである。
入力層: 複数個の特徴量を受け取る。
中間層: 10,000 個のニューロンで構成され、活性化関数として ReLU を使用。
出力層: 最終的に 1 つの数値（予測価格）を出力する。

本研究における一連の試行錯誤は、このモデル構造を変更することではなく、入力する特徴量の選択や、学習率・エポック数といったハイパーパラメータをいかに最適化するか、という点に主眼を置いて行われた。

2.3　Z スコア標準化

本研究で用いる特徴量は単位や尺度が異なるため、学習の安定化を目的として前処理に Z スコア標準化を適用した。これは、全特徴量と目的変数の平均を 0、標準偏差を 1 に変換する処理であり、モデルが各特徴を公平に評価し、効率的に学習する基盤を構築する。実装には scikit-learn の StandardScaler を用いた。

$$z = \frac{x - \mu}{\sigma}$$

ここで、μ は元のデータセットの平均値、σ は標準偏差を指す。

2.4　平均二乗誤差

本研究では、モデルの予測精度を測る損失関数（Loss）として、回帰問題で標準的に用いられる**平均二乗誤差（Mean Squared Error, MSE）**を採用した。MSE は、予測値と実

1: 日大理工・院（前）・数学　2: 日大理工・教員・数学

測値の差（誤差）を二乗し、その平均を取る指標であり、数式では以下のように表される。

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

・n: データの総数
・y_i: i 番目のデータの実際の値
・$\hat{y}_i$: i 番目のデータのモデルの予測値

誤差を二乗する特性上、外れ値の影響を強く受ける一方で、数学的には滑らかで微分可能な凸関数であるため、勾配降下法による最適解の探索が容易であるという利点を持つ。

3. 実験結果

本章では、前章で述べた手法を用いて構築したモデルを使用し、住宅価格予測の最適解を探索するために行った、複数のニューラルネットワークモデル構造および学習プロセスの評価結果について報告する。

3.1　初期モデルの性能評価

実験の出発点として、まず単一の中間層を持つ比較的シンプルな多層パーセプトロン（MLP）を構築し、その性能を評価した。最初期のモデルは 70,100 エポックで Loss: 0.0868 であり、予測値と実際の値の散布図を図 1 に示す。この時点での予測値と実測値の散布図（図 1）では、一定の相関は見られるものの、予測値のばらつきが大きく、特に低価格帯の物件において予測精度に課題が残ることが確認された 。また 3 万円から 4 万円の間にもばらつきが見受けられた

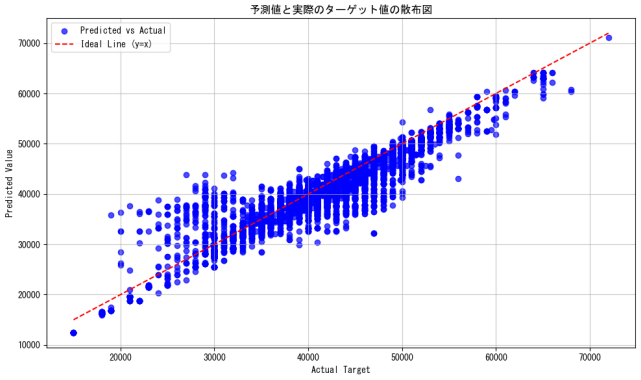

図 1: 予測値と実際の値の散布図（初期モデル）

3.2　モデル構造と学習プロセスの最適化、最終モデルの評価

初期モデルの性能を向上させるため、ニューラルネットワークのアーキテクチャは固定したまま、学習プロセスの最適化を試みた。具体的な改善アプローチとして、主に特徴量エンジニアリングとハイパーパラメータのチューニングに取り組んだ。特徴量エンジニアリングでは、新たな特徴量の追加や、先に述べたカテゴリカル特徴量に対するエンコーディング手法の追加を試みた。この試行錯誤の過程で、特に学習率を下げて学習を長時間継続することにより、損失が緩やかではあるが着実に低下していくことが観測された。

また初期の試行では、特定のハイパーパラメータ設定のもとで 70,100 エポックという長大な学習を行い、損失（Loss）を 0.0868 まで低下させることができた。しかし、これは非常に多くの計算時間を要するプロセスであった。その後、特徴量の追加や学習率の最適化をさらに進めた結果、最終的なモデルでは約 20,000 エポックの学習で、損失を 0.0734 まで改善させることに成功した

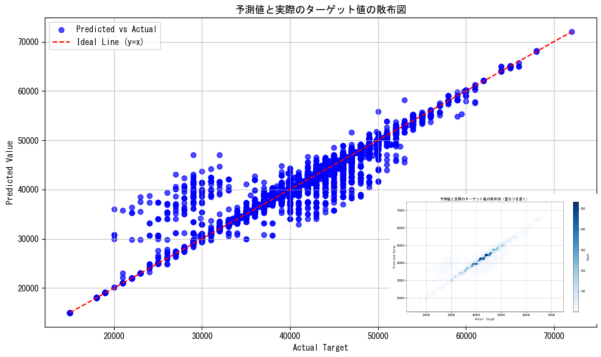

図 2: 予測値と実際の値の散布図（最終モデル）

4. 考察

本稿で構築したニューラルネットワークモデルは、特徴量とハイパーパラメータの最適化を通じて、その有効性が実証された。具体的には、当初 70,100 エポックを要して達成した損失（Loss）0.0868 に対し、最終モデルでは 20,000 エポックという 3 分の 1 以下の計算コストで、それを上回る損失 0.0734 を達成した。この結果は、特徴量の質と学習プロセスを改善することが、精度と効率の両面で極めて重要であることを示している。したがって、本研究で採用した最適化アプローチは、複雑な住宅価格予測問題に対して有効であると結論付けられる。

5. 参考文献

[1] アットホーム株式会社 (2024): アットホームデータセット. 国立情報学研究所情報学研究データリポジトリ. (データセット).

[2] 斎藤 康毅 (2016): 『ゼロから作る Deep Learning ― Python で学ぶディープラーニングの理論と実装』, オライリー・ジャパン.