

Data compression using the Fourier transform

Keita Haruyama

2025 年 10 月 28 日

目次

1 はじめに

本書は、データアナリティクス高度研修の補足資料として作成された。

1.1 予測することの重要性と、期待値で管理するという考え方

予測することで、ビジネスを加速させる旨記載予定。

1.2 従来の仮説検定から脱却し、ビッグデータを使いこなす

複数のパラメータを用いて回帰分析を行うが、数値的な情報が多くなりすぎると、全ての定性情報を落として、数値のみでの解析が可能となる。

2 回帰分析

回帰分析とは、目的変数を特徴量の乗数多項式の形で表したものである。特徴量が x_1 つであれば、 x の 0 乗から n 乗までの線形結合で目的変数を表す。複数の特徴量がある場合は、 n 乗までの線形結合が特徴量の数だけ必要となる。基本形は以下である。

特徴量が 1 つの場合

$$f(x) = a_0 + a_1x + a_2x^2 + \dots a_nx^n = \sum_{j=0}^n a_jx^j \quad (2.1)$$

特徴量が 2 つの場合

$$\begin{aligned} f(x_1, x_2) &= a_0 + a_1x_1 + a_2x_1^2 + \dots + a_nx_1^n + \\ &\quad b_0 + b_1x_2 + b_2x_2^2 + \dots + b_mx_2^m \\ &= \sum_{i=0}^n a_ix_1^i + \sum_{j=0}^m b_jx_2^j \end{aligned} \quad (2.2)$$

となる。式 (??) は有名なエネルギーと質量の関係式です。

実験や、マーティングを行うシミュレーションにて、上記のような予測モデルを仮定した際、係数 a, b が分かれば目的変数 f を数値化することが可能となる。

これらの係数を決定する方法について、数学的な証明を用いて説明していく。

まず前提とする条件は、観測データを取得している状態である。

2.1 線型回帰

2.2 最小二乗法

2.3 最尤推定法

3 決定木

3.1 サブタイトル

3.2 サブサブタイトル

4 数値計算法

4.1 Newton-Rapson 法

5 補足

5.1 相関と偏相関の違い

相関は、ピアソンの積率相関係数偏相関は、2 つの変数の関係を他の 1 つ、または複数の変数の影響を取り除いた上で測定する方法

5.2 変数選択法

変数選択は、Python ならば `mlex` で行う。

変数増減は、ステップノイズとも呼ばれ、F 値と関係している。優位水準=F 統計量の p 値

F 統計量=帰無仮説 (2 つ以上のグループの平均値が同じである) を検証するグループ間の変動、グループ内の変動の比率

p 値が小さい程、「帰無仮説のもとで」得られる確率が低い だから、その仮説が違う ということ

最大 R² 乗値改良法 R²(決定係数) は、回帰分析におけるモデルの適合度を表す。