

情報管理

次の 2 問から, 1 問を選択し解答しなさい。

- 問 1. n 個の訓練データ $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ が観測されたときに, 特徴量 $x \in \mathbb{R}^d$ と目的変数 $y \in \mathbb{R}$ の関係をうまく説明する関数 $y = f(x, w) = \sum_{j=1}^d w_j x_j$ を推定することを考える。このように x と y の関係を一次関数として捉えるものを線形回帰モデルと呼ぶ。ただし訓練データは中心化されており, 回帰モデルに定数項は含まれていないものとする。
- 最も代表的な推定手法として, 訓練データに対する二乗損失和 $L(w) = \sum_{i=1}^n (y_i - \sum_{j=1}^d w_j x_{ij})^2$ を最小にするパラメータ $w = (w_1, w_2, \dots, w_d)$ を求める最小二乗推定がある。近年, 機械学習の分野においては, 推定精度の向上や過学習の抑制のために, 正則化項を導入した正則化付き最小二乗推定が用いられ, 正則化項として L_2 ノルムの二乗 $\|w\|_2^2 = \sum_{j=1}^d w_j^2$ を導入したものを Ridge 回帰と呼ぶ。通常最小二乗推定量と Ridge 推定量がどのように与えられるかを示し, それらの性質について説明しなさい。
- 問 2. 大量のデータから知識を取り出す技術であるデータマイニングの分析手法の 1 つに, 特定の事象が起きたときにそれに関連して別の事象が起きやすいという事象の関連性 (アソシエーション) を求めるアソシエーション分析がある。事象の関連性は, アソシエーションルールとして記述され, 評価指標によってその有用性が判断されるが, 確信度 (Confidence), 支持度 (Support), リフト値 (Lift) によってアソシエーションルールのどのような事柄を評価できるのか述べなさい。