5.8 教師なし学習アルゴリズム

5.13 節で述べたように, 教師なしアルゴリズムは「特徴」だけを学習し, 教師信号を学習しないものである. 教師ありアルゴリズムと教師なしアルゴリズムの区別が厳密で形式的に定義されていないのは, 値が特徴であるか教師から提供されたターゲットであるかを区別する客観的なテストが存在しないからである. 非公式には, 教師なし学習とは, サンプルに注釈をつけるために人の労力を必要としない分布から情報を抽出する試みのほとんどを指す. この用語は通常, 密度推定, 分布からサンプルを抽出する方法の学習, ある分布からデータのノイズを除去する方法の学習, データが近くに存在する多様体を見つけること, あるいはデータを関連サンプルのグループにクラスタリングすることと関連づけられている.

古典的な教師なし学習のタスクは, データの「最適な」表現を見つけることである. 「最適」とはさまざまな意味を持ちますが, 一般的に言えば, $\boldsymbol{x}$ に関するできるだけ多くの情報を保存しながら, $\boldsymbol{x}$ 自体より簡潔で扱いやすい表現を保持することを目指したペナルティや制約に従うことである.

より単純な表現を定義する方法は複数ある. 最も一般的な 3 つは, 低次元表現, 疎な表現, そして独立した表現である. 低次元表現は, x に関するできるだけ多くの情報をより小さな表現で圧縮しようとする. 疎な表現 (1989 年 Barlow; 1996 年 Olshausen および Field; 1997 年 Hinton および Ghahramani) はほとんどの入力に対してエントリがほぼゼロである表現にデータセットを埋め込む. 疎な表現を使用するには, 通常, 表現の次元を増加させる必要があるため, 表現がほとんどゼロになっても情報があまり失われないようになる. これにより, 表現空間の軸に沿ってデータが分布する傾向がある表現の全体的な構造が得られる. 独立した表現は, 表現の次元が統計的に独立するように, データ分布の基礎となる変動の原因を解明しようとする.

もちろんこれら 3 つの基準は相互に排他的ではない. 低次元表現では, 元の高次元データよりも依存関係が少ないか弱い要素が生成されることがよくある. これは, 表現のサイズを削減する 1 つの方法が冗長性を見つけて削除することであるからである. より多くの冗長性を特定して削除することで, 次元削除アルゴリズムは, 破棄する情報を抑えながらより多くの圧縮を実現できる.

表現の概念は深層学習の中心的なテーマの 1 つであり, それゆえに本書の中心的なテーマの 1 つである. このセクションでは, 表現学習アルゴリズムの簡単な例をいくつか開発する. これらのアルゴリズムの例は, 上記の 3 つの基準を全て運用する方法を示している. 残りの章のほとんどで, これらの基準をさまざまな方法で開発したり, 他の基準を導入したりする追加の表現学習アルゴリズムを紹介する.

5.8.1 主成分分析

2.12 節では, 主成分分析アルゴリズムがデータを圧縮する手段を提供していることがわかった. PCA(主成分分析) をデータの表現を学習する教師なし学習アルゴリズムとみなすこともできる. この表現は, 上で説明した単純な表現の 2 つの基準に基づいている. PCA は元の入力よりも次元が低い表現を学習する. また, 要素間に線形相関がない表現も学習する. これは, 要素が統計的に独立している表現を学習するための基準に向けた最初のステップである. 完全な独立性を実現するためには, 表現学習アルゴリズムで変数間の非線形関係も除去する必要がある.

図 5.8: PCA は最大分散の方向を新たな空間の軸に合わせる線形射影を学習する. (左) 元のデータは $\boldsymbol{x}$ のサンプルで構成される. この空間では, 軸が揃っていない方向に沿って分散が生じる可能性がある. (右) 変換されたデータ $\boldsymbol{z} = \boldsymbol{x}^\top \boldsymbol{W}$ は, z_1 軸に沿って最も大きく変化する. 2 番目に大きい分散の方向は, z_2 に沿ったものになる.

PCA は, 図 5.8 に示すように入力 $\boldsymbol{x}$ を表現 $\boldsymbol{z}$ に射影するデータの交線形変換を学習する. 2.12 節では, (平均二乗誤差の意味で) 元のデータを最もよく再構成する 1 次表現を学習でき, この表現が実際にデータの第 1 主成分に対応していることがわかった. したがって, PCA をデータ内の情報をできるだけ多く保存する, 単純で効果的な次元削減手法として使用できる (これも, 最小二乗再構成誤差によって測定される). 以下では, PCA 表現が元のデータ表現 $\boldsymbol{X}$ とどのように無相関化されるかを検討する.

$m \times n$ 次元の計画行列 $\boldsymbol{X}$ を考えてみる. データの平均はゼロ, $\mathbb{E}[\boldsymbol{x}] = \boldsymbol{0}$ であると仮定する. そうでない場合は, 前処理ステップで全てのサンプルから平均を減算することで, データを簡単に中心に合わせることができる.

$\boldsymbol{X}$ に関連づけられた不偏標本の共分散行列は次の式で与えられる:

$$\mathrm{Var}[\boldsymbol{x}] = \frac{1}{m-1}\boldsymbol{X}^\top \boldsymbol{X}. \tag{5.85}$$

PCA は (線形変換を通じて), $\mathrm{Var}[\boldsymbol{z}]$ が対角である表現 $\boldsymbol{z} = \boldsymbol{x}^\top \boldsymbol{W}$ を見つける.

2.12 節では, 計画行列 $\boldsymbol{X}$ の主成分が $\boldsymbol{X}^\top \boldsymbol{X}$ の固有ベクトルによって与えられることがわかった. この視点から見ると,

$$\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{W}\boldsymbol{\Lambda}\boldsymbol{W}. \tag{5.86}$$

この節では, 主成分の代替導出を利用する. 主成分は, 特異値分解によって取得することもできる. 具体的には, これらは $\boldsymbol{X}$ の右特異ベクトルである. これを確認するには, 分解 $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{W}^\top$ の右特異ベクトルを $\boldsymbol{W}$ とする. 次に, 固有ベクトルの基底として, $\boldsymbol{W}$ を使用して元の固有ベクトル方程式を復元する:

$$\boldsymbol{X}^\top \boldsymbol{X} = (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{W}^\top)^\top \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{W}^\top = \boldsymbol{W}\boldsymbol{\Sigma}^2\boldsymbol{W}^\top. \tag{5.87}$$

SVD(特異値分解) は, PCA の結果が対角 $\mathrm{Var}[z]$ になることを示すのに役立つ. $\boldsymbol{X}$ の SVD を使用すると, $\boldsymbol{X}$ の分散は次のように表現できる:

$$\begin{aligned}
\mathrm{Var}[\boldsymbol{x}] &= \frac{1}{m-1}\boldsymbol{X}^\top \boldsymbol{X} && (5.88)\\
&= \frac{1}{m-1}(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{W}^\top)^\top \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{W}^\top && (5.89)\\
&= \frac{1}{m-1}\boldsymbol{W}\boldsymbol{\Sigma}^\top \boldsymbol{U}^\top \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{W}^\top && (5.90)\\
&= \frac{1}{m-1}\boldsymbol{W}\boldsymbol{\Sigma}^2\boldsymbol{W}^\top, && (5.91)
\end{aligned}$$

ここで, 特異値行列の $\boldsymbol{Y}$ 行列は正規直行であると定義されているため, $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}$ という事実を使用する. これは, $\boldsymbol{z} = \boldsymbol{x}^\top \boldsymbol{W}$ とすれば, $\boldsymbol{z}$ の共分散が必要に応じて対角であることを保証できることを示す.

$$\begin{aligned}
\mathrm{Var}[\boldsymbol{z}] &= \frac{1}{m-1}\boldsymbol{Z}^\top \boldsymbol{Z} && (5.92)\\
&= \frac{1}{m-1}\boldsymbol{W}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{W} && (5.93)\\
&= \frac{1}{m-1}\boldsymbol{W}^\top \boldsymbol{W}\boldsymbol{\Sigma}^2\boldsymbol{W}^\top \boldsymbol{W} && (5.94)\\
&= \frac{1}{m-1}\boldsymbol{\Sigma}^2, && (5.95)
\end{aligned}$$

ここで, 今回はやはり SVD の定義から $\boldsymbol{W}^\top \boldsymbol{W} = \boldsymbol{I}$ という事実を使用する.

上記の分析は, 線形変換 $\boldsymbol{W}$ を介してデータ $\boldsymbol{x}$ をデータ $\boldsymbol{z}$ に射影すると, 結果として得られる表現には ($\boldsymbol{\Sigma}^2$ で与えられる) 対角共分散行列が含まれることを示す. これは, $\boldsymbol{z}$ の個々の要素が相互に相関していないことを直ちに意味する.

要素が互いに相関しない表現にデータを変換する PCA のこの機能は, PCA の非常に重要な特性である. これは, データの基盤となる変動の未知の要因を解明することを試みた表現の簡単な例である. PCA の場合, この解き方は, 分散の主軸を $\boldsymbol{z}$ に関連づけられた新たな表現空間の基底に揃えるための ($\boldsymbol{W}$ で記述される) 入力空間の回転を見つけるという形式を取る.

相関関係は, データの要素間の依存関係の重要なカテゴリであるが, より複雑な形式の特徴の依存関係を解明する表現の学習にも興味がある. このためには, 単純な線形変換で実行できるもの以上のものが必要になる.