Recent Developments in Document Clustering

7. Dimensionality Reduction

7.1 Principal component analysis

7.2 Nonegative matrix factorization

08t40721 全太俊

7 Dimensionality Reduction

- この章ではベクトル空間のサイズを大幅に削減し、精度を向上できる方法を２つ紹介する

- 次元削減技術は、管理ができるエラーを導入しながらAランクkに近似するAkを生成する定義を導入する（フロベニウスノルム）

フロベニウスノルム

$$|A-A_k|=\sqrt{\sum_{a\in A}\sum_{a_k\in A_k}(a-a_k)^2}$$

- フロベニウスノルムが小さいほど近似する

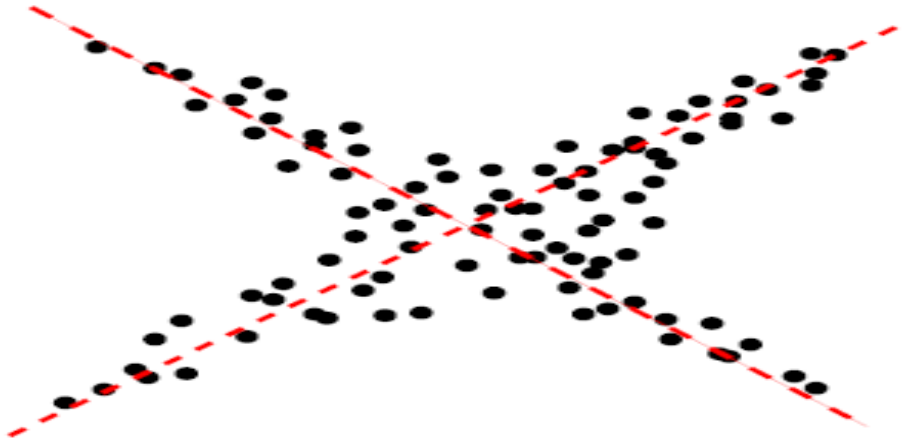
7.1 Principal component analysis

- 主成分の直交投影はデータセットの最大変動量で表示できる

- 実際に、主成分は特異値を計算することによって見つけることができる

- この方法はスペクトル投影と呼ばれる

7.1 Principal component analysis



- 二破線は、データセット内の変動をキャプチャする主成分を表す

特異値分解

$$A_n \approx U \sum V^T$$

↓

- Σは対角行列、UとVは直交行列である

特異値分解

$$A_k = U \sum_k V^T$$

- 上の式はAのランクkのスベクトル投影
- Σkのランクが高いほど元に近い

7.1 Principal component analysis

- 主成分分析法には近似と識別の属性がある
- **近似**

次元が４００から１２８に削減されるとき誤差が６０％になる

- **識別**

６章で紹介されて、もっと重要的である

7.1 Principal component analysis

- 主成分分析法の問題点

１．負数が含まれているので、直接クラスタできない

２．直交という制限があるので多くのトピックが飛ばされる

7.2 Nonnegative matrix factorization

- 非負行列因子分解（NMF）は 最初にコンピュータビジョンのアプリケーションで使われた
- 今はドキュメントクラスタリングに使われている。

7.2 Nonnegative matrix factorization

- NMFのいい点

近似の非負要因しか含まれていないので処理されてから後処理する必要がなくなった

7.2 Nonnegative matrix factorization

$$A \approx UV^T$$

- UとVはランダムに初期化される
- 期待値を最大化アルゴリズムで繰り返して評価する

7.2 Nonnegative matrix factorization

$$\Theta = \sum_i \sum_j (A_{ij} - \sum_l U_{il} V_{jl})$$

- 上式はUVの中で行列Aの列に対応する近似値のユークリッドの距離

- それはSVDよりも速く計算する。

7.2 Nonnegative matrix factorization

- NMFは問題

初期ランダムに依存して同じデータで違う結果がでる可能性がある