第8章　次元縮約
pLSIについて

7月1日

鈴木直也

次元縮約

- 現実のデータ解析では、動もするとデータを表すベクトルが非常に高次元になります。
  → ベクトル間の距離が離れている。
  →妥当な結果を得られない

次元が多くて困っているなら減らせばよい。

→ただ減らせばよいというものではなくできるだけ元の形を残して減らしていく。

射影

n次元空間の点をk次元空間に変換する処理を射影といいます。ここでn次元のデータがm個あるとき、このデータセットは行をデータに対応させるｍ×ｎの行列Xで表現できます。

データをk次元に射影する1つの方法が、ｎ×ｋの行列Aを掛けることです。このときXAの行ベクトルがk次に縮約されたデータのベクトルになります。次元縮約を行うには多数あるAの中から、何らかの条件を満たしたAを求める必要があります。

特異値分解

次元縮約を行う標準的な手法です。特異値分解では
m×nの行列Xを3つの行列 U, Σ, V^t の積に分解します。

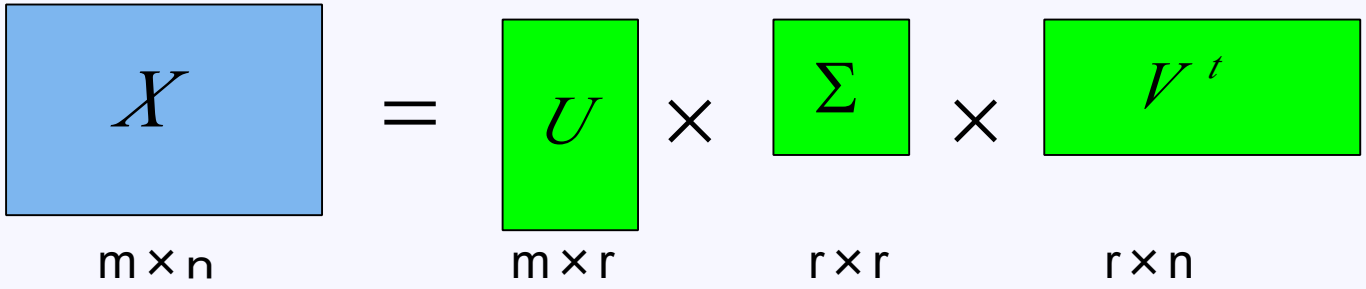
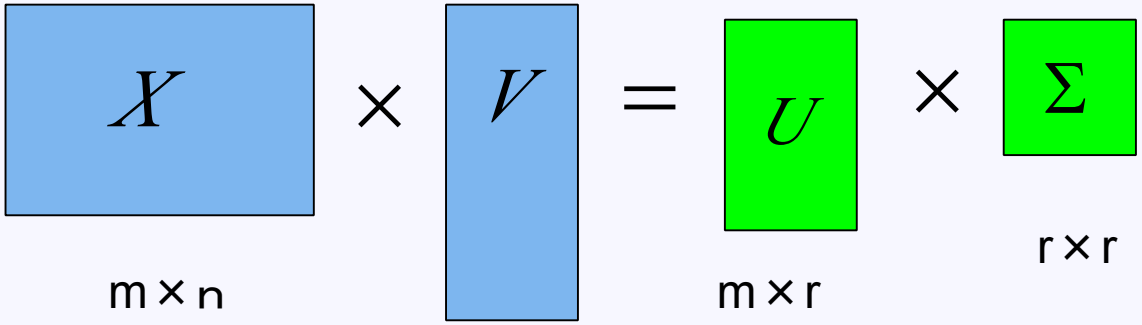


図8.1：特異値分解

Σは対角行列で、i行i列の対角要素を λ_i と置くと
$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \lambda_r \geq 0$ の関係があります。

ここで行列Vを右から掛けると



となります。

ここでXVはm×rの行列であり、元のm×nの行列が縮約されたことになります。

次元縮約の結論

- Uの列ベクトルはXの列ベクトルの張る空間の正規直行基底である。
- V^tの行ベクトルはXの行ベクトルの張る空間の正規直行基底である。
- λ_iはUのi番目の列ベクトル（あるいはV^tのi番目の行ベクトル）の基底としての重要度を表す。

LSI

これは次元縮約の手法の1つですが、特異値分解そのものです。ベクトル空間モデルで表現された文書ベクトルを特異値分解によって次元縮約する場合にLSIと呼びます。

文書ベクトルの各要素に重みをつける必要がある

→理論的な弱点

pLSI

文書ベクトルを想定した次元縮約の手法。
クラスタリングに直接利用することもできます。

文書ベクトルに重みも必要なく、頻度を要素とするベクトルでかまわない。

Aspectモデル

文書と単語を結ぶつける潜在的なクラスを想定したモデル。

文書dと単語wの出現をクラスzを用いて以下のようにモデル化をする。

$$p(w|d) = \sum_z p(w|z)p(z|d)$$

ベイズの定理から

$$p(z|d) = \frac{p(z)p(d|z)}{p(d)}$$

また、$p(d,w) = p(d)p(w|d)$ なので

続き

$$p(d,w)=\sum_{z}p(z)p(w|z)p(d|z) \qquad (8.1)$$

となります。

K次に縮約する場合クラスをK個に設定します。

$$z_1, z_2, z_3, \cdots z_K$$

文書dに対して $(p(z_1,d), p(z_2,d), \cdots p(z_K,d))$

が縮約されたベクトルとなります。

潜在的なクラスをそのままクラスタリングにおけるクラスと捉える。

→次元縮約の結果がクラスタリングを表す。

つまり、以下の式でデータdのクラスタ番号が得られます。

$$\arg\max_{K} p(d, z_k) = \arg\max_{K} p(z_k)\, p(d \mid z_k)$$

$p(z), p(w|z), p(d|z)$を求めるにおいて一般的な式を求める必要はない。

与えられた文書集合が $D=\{d_1, d_2, \cdots, d_N\}$ で

それに関した単語の集合が $W=\{w_1, w_2, \cdots, w_m\}$

である場合、各kに対する $p(z_k)$、各kとmに対する $p(w_m|z_k)$、各kとnに対する $p(d_n|z_k)$

が求まればよい。

→全部でK(1+M+N)個のパラメータを求める。

未知数（パラメータ）

パラメータは最尤法で求めることができる。

文書dに含まれる単語wの数を$n(d,w)$　で表すと対数尤度関数Lは以下になります。

$$L=\sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i,w_j)\log p(d_i,w_j)$$

式8.1より

$$L=\sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i,w_j)\log p\left(\sum_{k=1}^{K} p(z_k)p(w_i\mid z_k)p(d_i\mid z_k)\right) \qquad (8.2)$$

Lを最大化するパラメータを求める

→EMアルゴリズム[*1] を用いる。

ここで尤度関数に隠れ変数zが埋め込まれているのでQ関数は直接求められる。

---

*1：混合分布モデル参照

zの分布

z以外のパラメータを固定したときのzの分布を求める。
　→つまり　$p(z|d,w)$

Aspectモデルでは $p(d,w,z)=p(z)p(w|z)p(d|z)$

が定義されているので

$$p(z_k|d,w)=\frac{p(d,w,z_k)}{p(d,w)}=\frac{p(z_k)p(w|z_k)p(d|z_k)}{\sum_{k=1}^{K}p(z_k)p(w|z_k)p(d|z_k)}$$

続き

簡略化のために$p(z_k \mid d_i, w_j) = Q_{ijk}$　と置くと

$$Q_{ijk}^{(t+1)} = \frac{p(z_k)^{(t)}\, p(w \mid z_k)^{(t)}\, p(d \mid z_k)^{(t)}}{\sum_{k=1}^{K} p(z_k)^{(t)}\, p(w \mid z_k)^{(t)}\, p(d \mid z_k)^{(t)}}$$

となります。

zを固定した場合

zを固定つまり Q_{ijk} を固定したときのその他のパラメータを求める。

$$L = \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) \log\left(\sum_{k=1}^{K} p(z_k) p(w_j \mid z_k) p(d_i \mid z_k) \right) \qquad (8.3)$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) \log\left(\sum_{k=1}^{K} Q_{ijk} \frac{p(z_k) p(w_j \mid z_k) p(d_i \mid z_k)}{Q_{ijk}} \right) \qquad (8.4)$$

$$\leq \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) \sum_{k=1}^{K} Q_{ijk} \log \frac{p(z_k) p(w_j \mid z_k) p(d_i \mid z_k)}{Q_{ijk}} \qquad (8.5)$$ ＊1

＊1：Jensenの不等式より

続き

さらに変形して

$$L = \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) \sum_{k=1}^{K} Q_{ijk} \log\left(p(z_k) p(w_j \mid z_k) p(d_i \mid z_k) \right)$$

$$-\sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) \sum_{k=1}^{K} Q_{ijk} \log Q_{ijk}$$

第2項は定数になる。

→第1項を再びLと置く

ラグランジュの未定乗数法の利用

前式よりLの最大化は

$$\begin{cases} \sum_{i=1}^{N} p(d_i, z_k) = 1 \\ \sum_{j=1}^{M} p(w_i, z_k) = 1 \\ \sum_{k=1}^{K} p(z_k) = 1 \end{cases}$$

の関係があるのでラグランジュの未定乗数法を利用してとける。

$$L+\sum_{k=1}^{K}\alpha_k\left(1-\sum_{i=1}^{N}p(d_i\mid z_k)\right)+\sum_{k=1}^{K}\beta_k\left(1-\sum_{j=1}^{M}p(w_j\mid z_k)\right)+\gamma\left(1-\sum_{j=1}^{M}p(z_k)\right)$$

$$p(d_i\mid z_k)=u_{ik},\ p(w_j\mid z_k)=v_{jk},\ p(z_k)=w_k$$

とおいて上記の式を u_{ik}, v_{jk}, w_k
で偏微分して極値問題を解く

まず、u_{ik} を解きます。

$$\frac{\sum_{j=1}^{M} n(d_i, w_j) Q_{ijk}}{u_{ik}} - \alpha_k = 0$$

よって

$$u_{ik} = \frac{\sum_{j=1}^{M} n(d_i, w_j) Q_{ijk}}{\alpha_k}$$

両辺をiに関して和をとると

$$\sum_{i=1}^{N} u_{ik} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) Q_{ijk}}{\alpha_k}$$

1

よって前式は

$$\alpha_k = \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) Q_{ijk}$$

以上より

$$u_{ik} = p(d_i, z_k) = \frac{\sum_{j=1}^{M} n(d_i, w_j) Q_{ijk}}{\sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) Q_{ijk}}$$

更新式の形で書くと

$$p(d_i, z_k)^{(t)} = \frac{\sum_{j=1}^{M} n(d_i, w_j) Q_{ijk}^{(t)}}{\sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) Q_{ijk}^{(t)}}$$

同様に ν_{jk}, w_k でそれぞれ偏微分を行うと以下の式を得られます。

$$p(w_j, z_k)^{(t)} = \frac{\sum_{i=1}^{N} n(d_i, w_j) Q_{ijk}^{(t)}}{\sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) Q_{ijk}^{(t)}}$$

$$p(z_k)^{(t)} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) Q_{ijk}^{(t)}}{\sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) Q_{ijk}^{(t)}}$$