特徴選択と特徴空間の変換

- パターン認識の手順
  - 特徴抽出: 特徴ベクトル(パターン)の定義
  - 学習パターンを用いて特徴空間を各クラスに分割

- 問題点
  - 特徴ベクトルの各成分間のスケーリング，正規化
  - 特徴空間の次元数，次元の呪い
    特徴の数を増やせば情報量が増え，認識率も向上するという誤解
    * 特徴の数を増やせば増やすほど相関の高い特徴の組が混入する可能性が高まる
    * 統計計算に要する計算量は少なくとも次元のべき乗のオーダになる．計算量の爆発.
    * 有限個のパターンから識別器を設計する際，次元を高くしていくと誤識別率がかえって上昇する．ヒューズの現象

- 情報量正規化
- 次元削減
  - KL 展開: パターン全体の分布をもっともよく近似する部分空間
  - 線形判別法: 各クラスのパターン分布の分離度を最大にする部分空間

特徴量正規化

- 元特徴ベクトル: $\boldsymbol{x}$，変換後のベクトル: $\boldsymbol{y}$，変換行列: A

$$\boldsymbol{y} = A\boldsymbol{x}$$

- パターン: $\boldsymbol{x}_p = (x_{p1}, x_{p2}, \cdots, x_{pd})^T$
- 変換行列: $A = \mathrm{diag}(a_1, a_2, \cdots, a_d)$
- 変換後のベクトル: $\boldsymbol{y}(y_{pi} = a_i x_{pi})$
- p番目のパターンとその他のパターンの平均距離

$$r_p^2 = \frac{1}{n-1} \sum_{q=1}^{n} \sum_{j=1}^{d} (y_{pj} - y_{qj})^2$$

- 全パターンの平均距離

$$\begin{aligned}
R^2 &= \frac{1}{n} \sum_{p=1}^{n} r_p^2 = \frac{1}{n(n-1)} \sum_{p=1}^{n} \sum_{q=1}^{n} \sum_{j=1}^{d} a_j^2 (x_{pj} - x_{qj})^2 \\
&= \frac{n}{n-1} \sum_{j=1}^{d} a_j^2 \frac{1}{n} \sum_{p=1}^{n} \sum_{q=1}^{n} \left(\frac{1}{n} x_{pj}^2 - \frac{2}{n} x_{pj} x_{qj} + \frac{1}{n} x_{qj}^2 \right) \\
&= \frac{n}{n-1} \sum_{j=1}^{d} a_j^2 \left(\frac{1}{n} \sum_{q=1}^{n} \frac{1}{n} \sum_{p=1}^{n} x_{pj}^2 \right. \\
&\qquad \left. - 2 \frac{1}{n} \sum_{p=1}^{n} x_{pj} \frac{1}{n} \sum_{q=1}^{n} x_{qj} + \frac{1}{n} \sum_{p=1}^{n} \frac{1}{n} \sum_{p=1}^{n} x_{qj}^2 \right) \\
&= \frac{2n}{n-1} \sum_{j=1}^{d} a_j^2 (\bar{x_j^2} - \bar{x}_j^2) \\
&= 2 \sum_{j=1}^{d} a_j^2 \sigma_j^2
\end{aligned}$$

- 特徴の分散

$$\sigma_j^2 = \frac{1}{n-1} \sum_{p=1}^{n} (x_{pj} - \bar{x}_j)^2 = \frac{n}{n-1} \sum_{p=1}^{n} (\bar{x_j^2} - \bar{x}_j^2)$$

パターン相互の距離の最大化

- 制約条件 (特徴空間の体積不変)

$$\prod_{j=1}^{d} a_j = 1$$

- Lagrange 未定乗数法

$$L = 2\sum_{j=1}^{d} a_j^2 \sigma_j^2 - \lambda \left(\prod_{j=1}^{d} a_j - 1 \right)$$

$$\frac{\partial L}{\partial a_j} = 4a_j\sigma_j^2 - \lambda \prod_{k \neq j}^{d} a_k = 0$$

$$a_j = \frac{\sqrt{\lambda}}{2\sigma_j}, \quad \lambda = 4\left(\prod_{j=1}^{d} \sigma_j \right)^{2/d}$$

- 最適係数

$$a_j = \frac{1}{\sigma_j}\left(\prod_{k=1}^{d} \sigma_k \right)^{1/d}$$

各特徴軸を標準偏差で正規化し，平均の回りの分散 (すなわちパターンの広がり) を均一化

KL展開

- Karhunen-Loève展開: 線形空間における特徴ベクトルの分布をもっともよく近似する部分空間を求める方法
  - 分散最大基準
  - 平均二乗誤差最小基準
- 手法は多変量解析 (主成分分析)
- d 次元の特徴量ベクトル $\boldsymbol{x}$ から $\bar{d}$ $(< d)$ 次元の特徴 $\boldsymbol{y}$ への変換 A

$$\boldsymbol{y} = A\boldsymbol{x}, \quad A = (\boldsymbol{u}_1, \cdots, \boldsymbol{u}_{\bar{d}}), \quad A^T A = I$$

- 分散最大基準

$$\begin{aligned} \tilde{\sigma}^2(A) &= \frac{1}{n}\sum_{\boldsymbol{y}}(\boldsymbol{y}-\tilde{\boldsymbol{m}})^T(\boldsymbol{y}-\tilde{\boldsymbol{m}}) \qquad (\tilde{\boldsymbol{m}} = A^T\boldsymbol{m}) \\ &= \mathrm{tr}(A^T\Sigma A) \qquad (\Sigma = \frac{1}{n}\sum_{\boldsymbol{x}}(\boldsymbol{x}-\boldsymbol{m})(\boldsymbol{x}-\boldsymbol{m})^T) \end{aligned}$$

$$\max\{\tilde{\sigma}^2(A)\} = \max\{\mathrm{tr}(A^T\Sigma A)\} = \max\{\Lambda\} = \sum_{i=1}^{\bar{d}} \lambda_i$$

- 平均二乗誤差最小

$$\begin{aligned} \varepsilon^2(A) &= \frac{1}{n}\sum(A\boldsymbol{y}-\boldsymbol{x})^T(A\boldsymbol{y}-\boldsymbol{x}) \\ &= \mathrm{tr}R - \mathrm{tr}(A^T R A)) \quad (R = \frac{1}{n}\sum \boldsymbol{x}\boldsymbol{x}^T) \end{aligned}$$

$$\min\{\varepsilon^2(A)\} = \mathrm{tr}R - \sum_{i=1}^{\bar{d}} \lambda_i$$

線形判別法 (Fisher's discriminant method)

- 特徴空間上の 2 クラスのパターンの分布からこの 2 クラスを識別するのに最適な 1 次元軸を求める手法 (クラス内変動・クラス内分布比最大)
- 変動行列: S_i

$$S_i = \sum_{\boldsymbol{x} \in \mathcal{X}_i} (\boldsymbol{x} - \boldsymbol{m}_i)(\boldsymbol{x} - \boldsymbol{m}_i)^T$$

- クラス内変動行列: S_W

$$S_W = S_1 + S_2 = \sum_{i=1,2} \sum_{\boldsymbol{x} \in \mathcal{X}_i} (\boldsymbol{x} - \boldsymbol{m}_i)(\boldsymbol{x} - \boldsymbol{m}_i)^T$$

- クラス間変動行列: S_B

$$S_B = \sum_{i=1,2} n_i (\boldsymbol{m}_i - \boldsymbol{m})(\boldsymbol{m}_i - \boldsymbol{m})^T = \frac{n_1 n_2}{n} (\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^T$$

- d 次元特徴空間から 1 次元空間への変換: $A \quad y = A\boldsymbol{x}$
- 変換後のクラス内変動行列:$\tilde{S}_W$，クラス間変動行列: $\tilde{S}_B$

$$\tilde{S}_W = \tilde{S}_1 + \tilde{S}_2 = \sum_{i=1,2} \sum_{y \in \mathcal{Y}_i} (y - \tilde{m}_i)^2 = n_1 \tilde{\sigma}_1^2 + n_2 \tilde{\sigma}_2^2 = A^T S_W A$$

$$\tilde{S}_B = \sum_{i=1,2} n_i (\tilde{m}_i - \tilde{m})^2 = \frac{n_1 n_2}{n} (\tilde{m}_1 - \tilde{m}_2)^2 = A^T S_B A$$

- フィッシャーの評価基準 (クラス間変動のクラス内変動に対する比)

$$J_S(A) = \frac{\tilde{S}_B}{\tilde{S}_W} = \frac{n_1 n_2}{n} \frac{(\tilde{m}_1 - \tilde{m}_2)^2}{n_1 \tilde{\sigma}_1^2 + n_2 \tilde{\sigma}_2^2} = \frac{A^T S_B A}{A^T S_W A}$$

評価基準最大化

- $J_S(A)$ の最大化: $\tilde{S}_W = I$ の条件下で $\tilde{S}_B$ の最大化
- Lagrange 乗数

$$J(A) = A^T S_B A - \lambda(A^T S_W A - I)$$

$$\frac{\partial J(A)}{\partial A} = S_B A - \lambda S_W A = 0$$

- S_W が正則なら

$$(S_W^{-1} S_B - \lambda I)A = 0$$

- $S_W^{-1} S_B$ の最大固有値を λ_1 とすると

$$\max\{J_S(A)\} = \lambda_1$$

J_S を最大にする A は最大固有値に対応する固有ベクトル

KL 展開と線形判別法

- KL 展開
  - 特徴ベクトルの分布全体の持つ情報をなるべく最大限に反映できるように特徴空間の次元を削減する方法
  - クラスの識別に有効かどうかはわからない
  - 目的は表現のための (圧縮のための) 次元削減
- 線形判別法
  - クラスの分布の分離度を考慮した空間の変換方法
  - 判別のための次元削減
- それにも関わらず KL 展開が用いられる理由
  - 文字認識や音声認識などの高度な認識には，高次元の特徴ベクトルが必要であり，次元の呪いから逃れるためには次元削減が不可欠である.
  - はじめに選ばれた (人間が選んだ) 特徴には相関を持つ特徴の組が含まれている可能性がある.