４章 確率モデルと識別関数

12T4069L
佐鳥 恭太郎

特徴を減らす

特徴間に相関がある

=>どちらか一方あればよい

相関を無くすためには学習データの統計量が必要

（例:平均ベクトル,共分散行列）

=>確率分布のパラメータ と呼ばれる

パラメータが分かれば、

特徴量の線形変換で相関がなくせる

平均ベクトル

観測データ（d次元の特徴ベクトル）

$$x = (x_1, \ldots, x_d)^T \in R^d$$

平均ベクトルμ

$$\mu = (\mu_1, \ldots, \mu_d)^T = (E\{x_1\}, \ldots, E\{x_d\})^T$$

$E\{x_i\}$はi番目の特徴の期待値

$$\mu_i = E\{x_i\} = \int_{R^d} x_i p(x) dx = \int_{-\infty}^{\infty} x_i p(x_i) dx_i$$

$p(x_i)$はi番目の特徴を表す$p(x)$の周辺確率

$$p(x_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(x_1, \ldots, x_d) dx_1 \ldots dx_{i-1} dx_{i+1} \cdots dx_d$$

平均ベクトルの算術平均

平均ベクトルの算術平均

観測データがN個のとき以下のように表せる

$$\mu = \bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

共分散行列

共分散行列Σ

$$\Sigma = Var\{x\} = E\{(x - \mu)(x - \mu)^T\}$$

$$= E\left\{\begin{pmatrix} x_1 - \mu_1 \\ \vdots \\ x_d - \mu_d \end{pmatrix}(x_1 - \mu_1, \ldots, x_d - \mu_d)\right\}$$

$$= \begin{pmatrix} E\{(x_1 - \mu_1)(x_1 - \mu_1)\} & \cdots & E\{(x_1 - \mu_1)(x_d - \mu_d)\} \\ \vdots & \ddots & \vdots \\ E\{(x_d - \mu_d)(x_1 - \mu_1)\} & \cdots & E\{(x_d - \mu_d)(x_d - \mu_d)\} \end{pmatrix}$$

$$= (\sigma_{ij}) = \begin{cases} i = j & \text{分散} \\ i \neq j & \text{共分散} \end{cases}$$

標準偏差と共分散

$\sigma_{ii} = \sigma_i^2$と表し、σ_iを標準偏差という

観測データがN個与えられているとき、

n番目のデータの

i番目の特徴をx_{ni}

j番目の特徴をx_{nj}

とすると、共分散は

$$\sigma_{ij} = \mathrm{E}\{(x_i - \mu_i)(x_j - \mu_j)\}$$

$$= \frac{1}{N}\sum_{n=1}^{N}(x_{ni} - \mu_i)(x_{nj} - \mu_j)$$

相関係数

i番目とj番目の特徴間の相関係数ρ_{ij}は

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

となるので、$-1 < \rho_{ij} < 1$の範囲の値になる

観測データの標準化

学習データの特徴は、測定単位による影響で分布の形状が大きく変化してしまう

=>平均0,分散1に標準化することで影響を取り除く

特徴xの線形変換$y = ax + b$とする

平均$E\{y\} = E\{ax + b\} = aE\{x\} + b = a\mu + b$

分散$\mathrm{V}ar\{y\} = E\{(y - E\{y\})^2\}$

$$= E\{(ax + b - a\mu - b)^2\} = E\{(ax - a\mu)^2\}$$

$$= a^2E\{(x - \mu)^2\} = a^2\ Var\{x\} = a^2\sigma^2$$

標準化

xの平均μと標準偏差σを用いた線形変換

$$z = \frac{x-\mu}{\sigma}$$

より、$a = \frac{1}{\sigma}, b = -\frac{\mu}{\sigma}$ なので、

$$E\{z\} = \frac{\mu}{\sigma} - \frac{\mu}{\sigma} = 0, Var\{z\} = \left(\frac{1}{\sigma}\right)^2 \sigma^2 = 1$$

となり、zの平均は0,分散は1になる

=>元のデータの平均と標準偏差を用いてzのように線形変換することを、データの標準化という

観測データの無相関化

共分散行列Σの固有値問題

$$\Sigma s = \lambda s$$

を解いて得られたd個の固有値を

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$$

とし、対応する固有ベクトルをならべて行列Sを定義する

$$S = (s_1, s_2, \cdots, s_d)$$

行列Sは直交行列なのでSによる線形変換は元の座標系を固有ベクトル方向に回転する

=> Sは回転行列と呼ばれる

無相関化

観測データxを,S^Tで線形変換$(y = S^T x)$する

平均$E\{y\} = E\{S^T x\} = S^T \mu$

$$\text{分散}\mathrm{Var}\{\mathrm{y}\} = \mathrm{E}\{(\mathrm{y} - \mathrm{E}\{\mathrm{y}\})(y - E\{y\})^T\}$$

$$= S^{-1}\mathrm{E}\{(\mathrm{y} - \mathrm{E}\{\mathrm{y}\})(y - E\{y\})^T\}\mathrm{S}$$

$$= S^{-1}\Sigma\mathrm{S} = \Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda_d \end{pmatrix}$$

=>対角化

白色化

共分散行列Σの対角化をする
=>特徴間の相関がなくなる
=>特徴量の標準偏差に違いがある
=>標準偏差を1に正規化し、中心化する
白色化
白色化後の座標系は

$$u = (u_1, \ldots, u_d)^T = \Lambda^{-1/2} S^T (x - \mu)$$

となる
※$\Lambda^{-1/2}$:Λの各対角要素の平方根を取った逆行列

確率モデル

パラメトリックモデル

=>学習データから推定したパラメータを用いて構成した確率モデルで学習データの分布を表現する

・確率変数が離散的な値を取る

(二項分布,多項分布,ポアソン分布など)

・連続な値を取る

(一様分布,指数分布,正規分布など)

ノンパラメトリックモデル

=>学習データそのものを用いてデータ分布を表現

・ヒストグラム法,K最近傍法,パルツェン密度推定法などがある

正規分布

なぜ正規分布が確率モデルの代表なのか

・多くの観測データが正規分布に従う

・正規分布と仮定することで解析的な解が得られやすい

・データが正規分布していなくても、データの平均の分布は正規分布になる（中心極限定理）

・確率分布が、平均値と共分散で決まる

・正規分布しているデータの線形変換で得られる分布も正規分布

・正規分布をする複数の確率変数の線形和は、正規分布(再生性)

・正規分布の周辺分布も正規分布

・正規分布に限り、無相関であることと統計的に独立であることが等価

=>共分散行列を対角化でき、統計的に独立な要素に分解できる

多次元正規分布

確率変数がd個の要素をもつベクトルのとき、
d次元の多次元正規分布関数となる

$$N(x|\mu,\Sigma)=\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mathrm{\mu})\right)$$

正規分布関数の指数部は、任意点xと平均ベクトルμとの距離（マハラノビス距離）を表している

$$d(x,\mu)=\sqrt{(x-\mu)^T\Sigma^{-1}(x-\mathrm{\mu})}$$

識別関数

正規分布から識別関数を導く

i番目のクラスのクラス条件付き確率が次の正規分布

$$p(x|C_i) = \frac{1}{(2\pi)^{d/2}|\Sigma_{\mathrm{i}}|^{1/2}} exp\left(-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mathrm{\mu}_i)\right)$$

をしていると仮定して、ベイズの誤り率最小識別規則を満たす識別関数を求める。

事後確率

クラスの事前確率を$P(C_i)$とすると、事後確率は

$$P(C_i|x) = \frac{p(x|C_i)P(C_i)}{p(x)}$$

クラスiに関係ないので無視

$$\propto \frac{P(C_i)}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)^T \Sigma_\mathrm{i}^{-1}(x-\mu_i)\right)$$

対数を取ると

$$\ln P(C_i) - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(x-\mu_i)^T \Sigma_\mathrm{i}^{-1}(x-\mu_i)$$

となる。

評価値

共通項を除き、符号を反転させてi番目のクラスの事後確率から求められる評価値は

$$g_i(x) = (x - \mu_i)^T \Sigma_\mathrm{i}^{-1} (x - \mu_i) + \ln|\Sigma_i| - 2\ln P(C_i)$$

と書ける。よってもっとも小さな値をとるクラスを選択すれば誤り最小基準のベイズの識別規則が得られる。

$$\text{識別クラス} = arg\min_i |g_i(x)|$$

識別境界

クラス間の識別境界は、2クラスの事後確率が等しくなる点の軌跡。

$$f_{ij}(x) = g_i(x) - g_j(x) = \cdots = x^T S x + 2c^T x + F = 0$$

※$S = \Sigma_{\mathrm{i}}^{-1} - \Sigma_j^{-1}$

$c^T = \mu_j^T \Sigma_j^{-1} - \mu_i^T \Sigma_{\mathrm{i}}^{-1}$

$$F = \mu_i^T \Sigma_{\mathrm{i}}^{-1} \mu_i - \mu_j^T \Sigma_{\mathrm{j}}^{-1} \mu_j + \ln \frac{|\Sigma_i|}{|\Sigma_j|} - 2 \ln \frac{P(C_i)}{P(C_j)}$$

$f_{ij}(x)$は正負の値をとるので関数値によってはクラスを識別することができる

=>２次識別関数

線形識別関数

$\Sigma_i = \Sigma_j = \Sigma$のとき、行列$\mathrm{S} = 0$なので、

$$f_{ij}(x) = g_i(x) - g_j(x) = 2c^T x + F = 0$$

と表せる。

=>線形識別関数

距離による識別

$\Sigma = \sigma I$のように、2つのクラスの共分散行列が同じ等方性分散をもち、クラスの事前確立が等しい($P(C_i) = P(C_j)$)とき

$$(x - \mu_i)^T (x - \mu_i) = (x - \mu_j)^T (x - \mu_j)$$

が成り立つ。入力ベクトルと、2つのクラスの平均ベクトルとの間の距離（の2乗）が小さな方のクラスに識別される。

=>最近傍法と同じ

確率モデル

学習データ$x_i (i = 1, \dots, N)$
=>真の分布から独立にサンプル（i.i.d.標本）
確率モデル$f(x|\theta)$
=>真の分布を,パラメータθにもつ
確率モデル$f(x|\theta)$に従うN個の学習データの同時分布から、

$$f(x_1, \dots, x_N|\theta) = \prod_{i=1}^{N} f(x_i|\theta)$$

と表せる。

尤度関数

関数fを確率モデルと考えると

=>パラメータθは定数,xは確率変数

ここではパラメータθを変数と考える尤度関数

$$L(\theta) = f(x_1, \ldots, x_N | \theta)$$

この尤度を最大にするパラメータθを見つける

=>最尤推定法

最尤推定法

尤度関数$L(\theta)$または$\ln L(\theta)$を微分し、0と置いて解くことで最適なパラメータθを得る。

$$\theta = (\theta_1, \ldots, \theta_M)^T$$

$$\frac{\partial L(\theta)}{\partial \theta_i} = 0 \text{ または } \frac{\partial \ln L(\theta)}{\partial \theta_i} = 0 \ (i = 1, \ldots, M)$$

普通、微分しやすい対数を取ったほうを用いる。