

機械学習と データマイニングの基礎

大阪大学 産業科学研究所
原 聡

原担当パートの内容

■ 教師あり学習

- ・ 11/17(金) 回帰と分類
- ・ 11/24(金) スパース正則化

■ 教師なし学習

- ・ 12/ 1(金) 密度関数の推定
- ・ 12/ 8(金) 確率的生成モデル

■ 成績評価

- ・ レポート課題1回(12/1出題)にて評価

教師なし学習

確率密度関数の推定と異常検知

大阪大学 産業科学研究所

原 聡

教師なし学習とは

■ 教師あり学習

- ・ 入力データ x と対応する出力データ y の組が複数与えられた時に x から y を予測する関数 f を学習する: $y \approx f(x)$ 。

■ 教師なし学習

- ・ 入力データ x しか与えられていない場合の機械学習。
≈ 入力データの確率密度関数 $p(x)$ の推定問題。

■ 本日の内容

- ・ 確率密度関数の推定
- ・ 異常検知

講義内容

- 確率密度関数の推定
 - ・ 正規分布
 - 最尤推定
 - ・ 混合正規分布
 - EMアルゴリズム
 - 応用: クラスタリング
- 異常検知
 - ・ 確率密度関数を使った異常検知

確率密度関数 $p(x)$

- 確率密度関数 $p(x)$ はどれくらい入力点 x が「本来のデータに近しいか」を測る指標。
 - ・ 値が小さいと、そのような入力点 x は「本来のデータらしくない」と言える。
- 応用例1：外れ値の検知
 - ・ 密度関数 $p(x)$ の値が小さいデータ点 x は、通常のデータでは「ほぼありえない点」。そのような点は何かがおかしい(外れ値)と言える。
- 応用例2：データの生成・サンプリング
 - ・ 密度関数 $p(x)$ がわかれば、 $p(x)$ に従う「本来のデータらしい」新しいデータを生成・サンプリングできる。

講義内容

- 確率密度関数の推定
 - ・ 正規分布
 - 最尤推定
 - ・ 混合正規分布
 - EMアルゴリズム
 - 応用: クラスタリング
- 異常検知
 - ・ 確率密度関数を使った異常検知

正規分布

- 釣鐘型の確率密度関数。
 - ・ 確率変数 x が d 次元の場合： $x \in \mathbb{R}^d$

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} \sqrt{\det \Sigma}} \exp \left(-\frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu) \right)$$

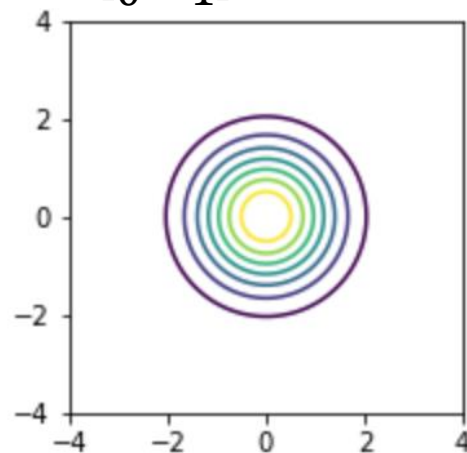
- ・ 分布の平均 $\mu \in \mathbb{R}^d$
- ・ 分布の分散共分散行列 $\Sigma \in \mathbb{R}^{d \times d}$

$$\mu = \mathbb{E}[x] = \int_{\mathbb{R}^d} x p(x) dx$$

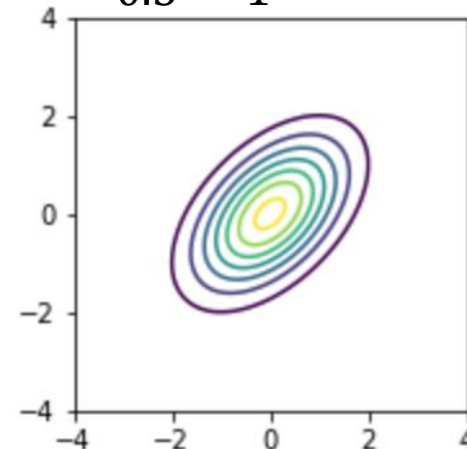
$$\Sigma = \mathbb{E}[(x - \mu)(x - \mu)^\top]$$

$$= \int_{\mathbb{R}^d} (x - \mu)(x - \mu)^\top p(x) dx$$

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I_2 \text{ の場合}$$



$$\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \text{ の場合}$$



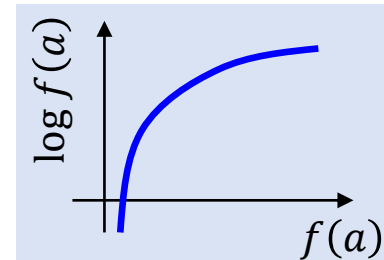
正規分布の推定: 最尤推定

- $p(x; \mu, \Sigma)$ の推定 = 分布の平均 μ と分散共分散行列 Σ の推定問題。
- 仮定: 観測データが独立同一分布に従う。
 - ・ 観測データ: $D = \{x^{(n)} \in \mathbb{R}^d\}_{n=1}^N$, $x^{(n)} \sim p(x; \mu, \Sigma)$ は i.i.d.
 - i.i.d.: independent and identically distributed
- 最尤推定: 観測データが“一番”尤もらしく見える”分布を推定する。
 - ・ 平均 μ と分散共分散行列 Σ のもとで、観測データ D が一番尤もらしいのは $\prod_{n=1}^N p(x^{(n)}; \mu, \Sigma)$ が最大するとき。

$$\hat{\mu}, \hat{\Sigma} = \operatorname{argmax}_{\mu, \Sigma} \prod_{n=1}^N \frac{1}{(2\pi)^{d/2} \sqrt{\det \Sigma}} \exp \left(-\frac{1}{2} (x^{(n)} - \mu)^\top \Sigma^{-1} (x^{(n)} - \mu) \right)$$

正規分布の推定: 最尤推定 = 対数尤度の最大化

- $f(a) > 0$ の最大化 = $\log f(a)$ の最大化
 - ・ $f(a)$ の最大化は一般に単調変換について不変



- 確率の最大化を「確率の対数(対数尤度)の最大化」に変換した方が解きやすいことが多い。

$$\hat{\mu}, \hat{\Sigma} = \operatorname{argmax}_{\mu, \Sigma} \prod_{n=1}^N \frac{1}{(2\pi)^{d/2} \sqrt{\det \Sigma}} \exp \left(-\frac{1}{2} (x^{(n)} - \mu)^\top \Sigma^{-1} (x^{(n)} - \mu) \right)$$



$$\hat{\mu}, \hat{\Sigma} = \operatorname{argmax}_{\mu, \Sigma} \log \prod_{n=1}^N \frac{1}{(2\pi)^{d/2} \sqrt{\det \Sigma}} \exp \left(-\frac{1}{2} (x^{(n)} - \mu)^\top \Sigma^{-1} (x^{(n)} - \mu) \right)$$

こっちを解く

正規分布の推定: 最尤推定 = 対数尤度の最大化

$$\hat{\mu}, \hat{\Sigma} = \operatorname{argmax}_{\mu, \Sigma} \log \prod_{n=1}^N \frac{1}{(2\pi)^{d/2} \sqrt{\det \Sigma}} \exp \left(-\frac{1}{2} (x^{(n)} - \mu)^\top \Sigma^{-1} (x^{(n)} - \mu) \right)$$

■ 式変形

- (板書)

正規分布の推定: 最尤推定 = 対数尤度の最大化

$$\hat{\mu}, \hat{\Sigma} = \operatorname{argmax}_{\mu, \Sigma} -\frac{1}{2} \sum_{n=1}^N (x^{(n)} - \mu)^{\top} \Sigma^{-1} (x^{(n)} - \mu) - \frac{N}{2} \log \det \Sigma - \frac{Nd}{2} \log 2\pi$$

- 平均 μ と分散共分散行列 Σ について微分して0とおいて解けば良い。

- (導出は板書)

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^N x^{(n)}, \quad \hat{\Sigma} = \frac{1}{N} \sum_{n=1}^N (x^{(n)} - \mu)(x^{(n)} - \mu)^{\top}$$

- 平均、分散共分散行列の最尤推定量はデータの平均、分散共分散行列に一致。
 - データの統計処理が、自然な形で正規分布のパラメータの推定に一致する。

数値実験例

■ 3次元の正規分布

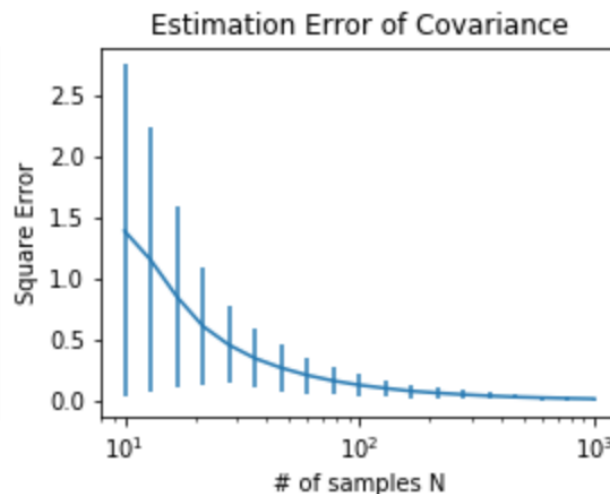
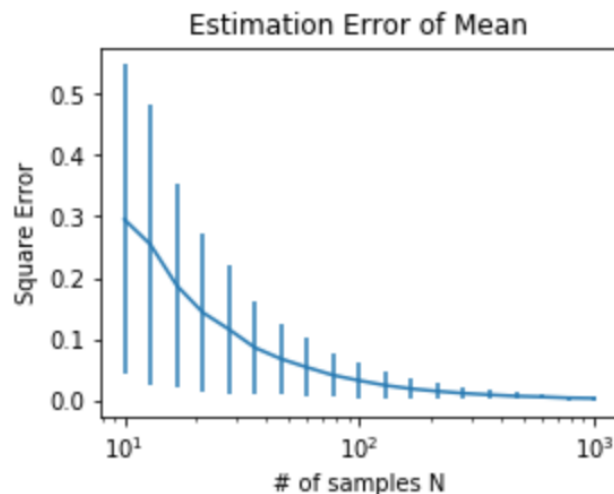
- 平均: $\mu = [2 \quad 1 \quad -1]^\top$

- 分散共分散行列: $\Sigma = \begin{bmatrix} 1 & -0.5 & 0 \\ -0.5 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

■ 二乗誤差で推定誤差を評価

- 推定誤差

- $\text{Err}_\mu = \|\mu - \hat{\mu}\|^2, \quad \text{Err}_\Sigma = \|\Sigma - \hat{\Sigma}\|_F^2 = \sum_{i,j} (\Sigma_{ij} - \hat{\Sigma}_{ij})^2$



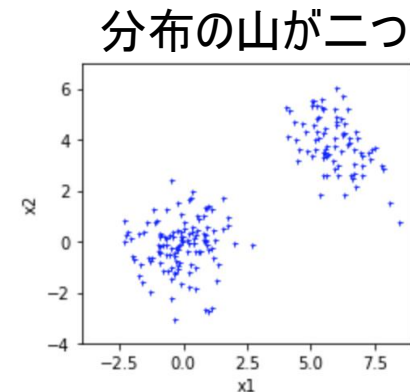
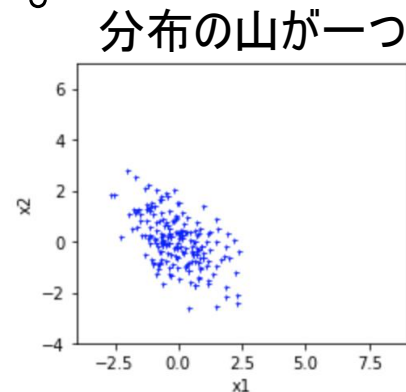
データ数Nが増え
ると推定誤差は
減少する

講義内容

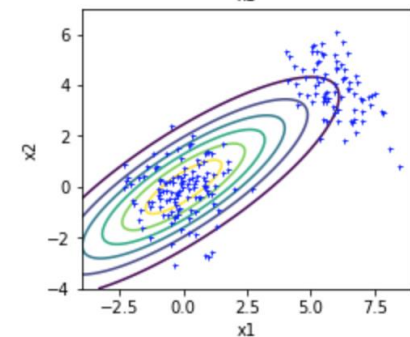
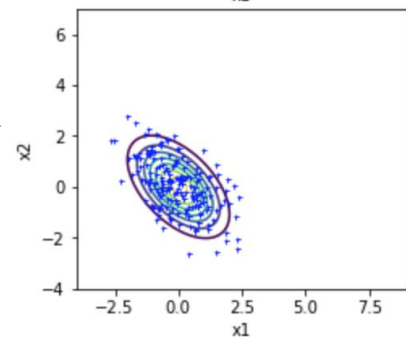
- 確率密度関数の推定
 - ・ 正規分布
 - 最尤推定
 - ・ 混合正規分布
 - EMアルゴリズム
 - 応用: クラスタリング
- 異常検知
 - ・ 確率密度関数を使った異常検知

正規分布の利用が不適切な例

- 正規分布は釣鐘型の分布。つまり山は一つ。
- 常にデータの密度関数の山が一つとは限らない。
 - ・ 山が複数あるデータのモデル化に正規分布を使うのは適切ではない。
 - ・ 山が複数あるデータを正規分布でモデル化すると、データの傾向を適切に捉えられない。



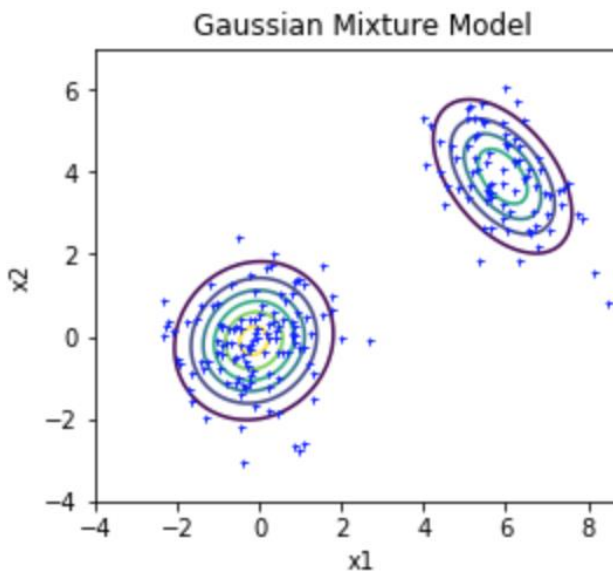
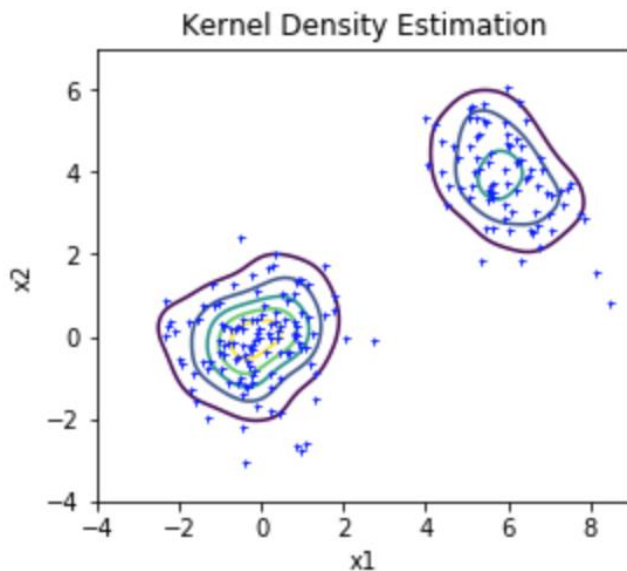
正規分布を使った
密度関数の推定



山が複数ある場合の密度関数の推定

■ 代表例

- ・ カーネル密度推定
- ・ 混合正規分布



山が複数あっても、
きちんと分布の傾向
を捉えられる。

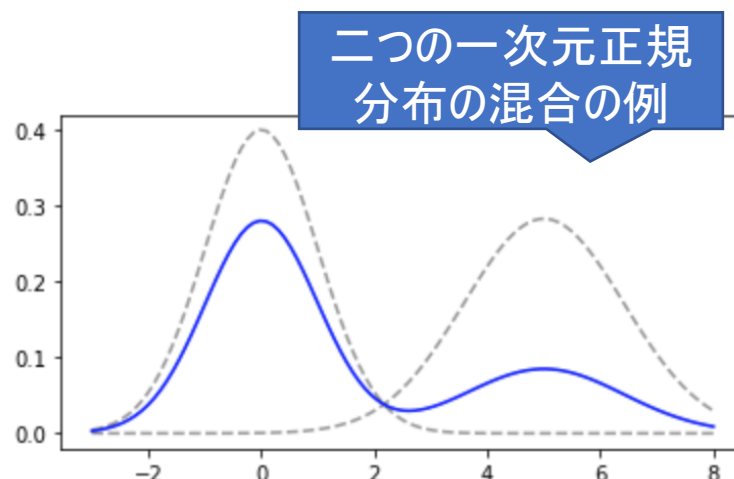
混合正規分布

- K 個の釣鐘型の密度関数でデータ全体の分布を表現。
 - ・ 混合正規分布は山が K 個ある分布が表現可能。

- 混合正規分布の定義

$$p(x; \theta) = \sum_{k=1}^K \pi_k p(x; \mu_k, \Sigma_k)$$

$$- \pi_k \geq 0, \sum_{k=1}^K \pi_k = 1$$



- ・ 混合正規分布は、 K 個の異なる平均 μ_k と分散共分散行列 Σ_k を持つ正規分布を重み π_k で足し合わせたもの。
- ・ $\theta = \{\mu_k, \Sigma_k, \pi_k\}_{k=1}^K$ が分布のパラメータ。

混合正規分布の推定

- 混合正規分布のパラメータ $\theta = \{\mu_k, \Sigma_k, \pi_k\}_{k=1}^K$ をデータから推定する。
- 仮定：観測データが独立同一分布に従う。
 - ・ 観測データ： $D = \{x^{(n)} \in \mathbb{R}\}_{n=1}^N$, $x^{(n)} \sim p(x; \theta)$ は i.i.d.
- 推定の方針：最尤推定 = 対数尤度の最大化

$$\hat{\theta} = \operatorname{argmax}_{\theta} \sum_{n=1}^N \log p(x^{(n)}; \theta)$$

混合正規分布の推定: 対数尤度の最大化

■ 対数尤度の最大化

$$\sum_{n=1}^N \log p(x^{(n)}; \theta) = \sum_{n=1}^N \log \left(\sum_{k=1}^K \pi_k p(x; \mu_k, \Sigma_k) \right)$$

- 正規分布の場合のように「微分=0」と置いても解析的には解けない。
 - ・ 正規分布の場合に「微分=0」で解けたのは正規分布が”キレイ”な分布だったから。
 - ・ 混合正規分布は正規分布より複雑なので、「微分=ゼロ」では駄目。
- 方法: EMアルゴリズム
 - ・ 繰り返しアルゴリズムによる対数尤度の最大化。

混合正規分布の推定: 対数尤度の最大化

- 解きやすい形に問題を変換する。
 - ・ 正規分布の最尤推定が簡単だったのは、 $\log \exp$ の形のおかげで、対数尤度がキレイな形で書けたから。
- 混合正規分布では $\log \text{sum exp}$ の形のため、同様の式変形ができない。

$$\sum_{n=1}^N \log p(x^{(n)}; \theta) = \sum_{n=1}^N \log \left(\sum_{k=1}^K \pi_k p(x^{(n)}; \mu_k, \sigma_k^2) \right)$$

- EMアルゴリズムでは、 $\log \text{sum exp}$ を sum log exp の形に書き換えるテクニックを使って、計算を簡単化する。

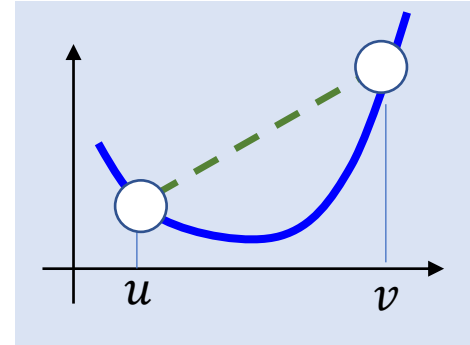
→ Jensenの不等式

【準備】 Jensenの不等式

- 仮定1: 関数 $f(u)$ は凸関数。

- ・ 凸関数の定義

$$f(au + (1 - a)v) \leq af(u) + (1 - a)f(v) \\ \forall a \in [0, 1], \forall u, v \in \mathbb{R}$$



- 仮定2: $\alpha_k \geq 0, \sum_{k=1}^K \alpha_k = 1$

- Jensenの不等式

$$\sum_{k=1}^K \alpha_k f(u_k) \geq f\left(\sum_{k=1}^K \alpha_k u_k\right)$$

凸関数の出力の重み付き和は、入力の重み付き和に対する関数の出力より大きい

【準備】 Jensenの不等式

- 証明
 - (板書)

【準備】 Jensenの不等式

- 負の対数 $f(u) = -\log u$ は凸関数。
- Jensenの不等式: $f(u) = -\log u$ の場合

$$\sum_{k=1}^K \alpha_k f(u_k) \geq f\left(\sum_{k=1}^K \alpha_k u_k\right)$$


$$\Rightarrow -\sum_{k=1}^K \alpha_k \log u_k \geq -\log\left(\sum_{k=1}^K \alpha_k u_k\right)$$

$$\Rightarrow \sum_{k=1}^K \alpha_k \log u_k \leq \log\left(\sum_{k=1}^K \alpha_k u_k\right)$$

混合正規分布の推定

■ 対数尤度の下限を導出する。

- Jensenの不等式を使って、 $\log \text{sum exp}$ を sum log exp の形に書き換える。

$$\sum_{n=1}^N \log p(x^{(n)}; \theta) = \sum_{n=1}^N \log \left(\sum_{k=1}^K \pi_k p(x^{(n)}; \mu_k, \Sigma_k) \right)$$


(板書)

$$\geq \sum_{n=1}^N \sum_{k=1}^K \alpha_k^{(n)} \log \frac{\pi_k p(x^{(n)}; \mu_k, \Sigma_k)}{\alpha_k^{(n)}}$$

- $p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} \sqrt{\det \Sigma}} \exp \left(-\frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu) \right)$ なので、これは sum log exp 。

混合正規分布の推定

■ ここまでのまとめ

- ・ 対数尤度を最大にする $\theta = \{\mu_k, \Sigma_k, \pi_k\}_{k=1}^K$ を見つけたい。

$$\max_{\theta} L(\theta) =: \sum_{n=1}^N \log \left(\sum_{k=1}^K \pi_k p(x^{(n)}; \mu_k, \Sigma_k) \right)$$

- ・ Jensenの不等式から、 $L(\theta)$ の下限 $L_{\text{LB}}(\theta, \alpha)$ が導出できる。

$$\begin{aligned} \max_{\theta, \alpha} L_{\text{LB}}(\theta, \alpha) &= \sum_{n=1}^N \sum_{k=1}^K \alpha_k^{(n)} \log \frac{\pi_k p(x^{(n)}; \mu_k, \Sigma_k)}{\alpha_k^{(n)}} \\ \text{s.t. } \alpha_k^{(n)} &\geq 0, \sum_{k=1}^K \alpha_k^{(n)} = 1 \end{aligned}$$

- $L_{\text{LB}}(\theta, \alpha)$ はsum log expの形なので最適化が簡単。
- 任意の α について $L(\theta) \geq L_{\text{LB}}(\theta, \alpha)$ なので、 $L_{\text{LB}}(\theta, \alpha)$ を θ と α について最大化すれば、間接的に $L(\theta)$ が最大化できる。

混合正規分布の推定

■ EMアルゴリズム

- $L_{LB}(\theta, \alpha)$ を θ と α について交互に最大化する。

$$\begin{aligned} \max_{\theta, \alpha} L_{LB}(\theta, \alpha) &= \sum_{n=1}^N \sum_{k=1}^K \alpha_k^{(n)} \log \frac{\pi_k p(x^{(n)}; \mu_k, \Sigma_k)}{\alpha_k^{(n)}} \\ \text{s. t. } \alpha_k^{(n)} &\geq 0, \sum_{k=1}^K \alpha_k^{(n)} = 1 \end{aligned}$$

- ただし

$$p(x; \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{d/2} \sqrt{\det \Sigma_k}} \exp \left(-\frac{1}{2} (x - \mu_k)^\top \Sigma_k^{-1} (x - \mu_k) \right)$$

混合正規分布の推定

■ E-ステップ: $L_{LB}(\theta, \alpha)$ の α についての最大化

- 適当な $\theta = \{\mu_k, \Sigma_k, \pi_k\}_{k=1}^K$ の値が与えられ固定されているとする。
- α についての最大化問題は以下の通り。

$$\hat{\alpha} = \operatorname{argmax}_{\alpha} \sum_{n=1}^N \sum_{k=1}^K \alpha_k^{(n)} \log \frac{\pi_k p(x^{(n)}; \mu_k, \Sigma_k)}{\alpha_k^{(n)}}, \text{ s. t. } \alpha_k^{(n)} \geq 0, \sum_{k=1}^K \alpha_k^{(n)} = 1$$

• 最適解

$$\hat{\alpha}_k^{(n)} = \frac{\pi_k p(x^{(n)}; \mu_k, \Sigma_k)}{\sum_{k=1}^K \pi_k p(x^{(n)}; \mu_k, \Sigma_k)} \quad (\text{板書})$$

混合正規分布の推定

- M-ステップ: $L_{LB}(\theta, \alpha)$ の θ についての最大化
 - ・ 適切な α の値が与えられ固定されているとする。

$$\hat{\theta} = \operatorname{argmax}_{\theta} \sum_{n=1}^N \sum_{k=1}^K \alpha_k^{(n)} \left(\log \pi_k + \log p(x^{(n)}; \mu_k, \Sigma_k) \right), \text{ s. t. } \pi_k \geq 0, \sum_{k=1}^K \pi_k = 1$$

- π_k, μ_k, Σ_k についての最適化

(課題)

$$\hat{\pi}_k = \frac{\sum_{n=1}^N \alpha_k^{(n)}}{\sum_{n=1}^N \sum_{k=1}^K \alpha_k^{(n)}}$$

$$\hat{\mu}_k = \frac{\sum_{n=1}^N \alpha_k^{(n)} x^{(n)}}{\sum_{n=1}^N \alpha_k^{(n)}}$$

$$\hat{\Sigma}_k = \frac{\sum_{n=1}^N \alpha_k^{(n)} (x^{(n)} - \mu)(x^{(n)} - \mu)^{\top}}{\sum_{n=1}^N \alpha_k^{(n)}}$$

混合正規分布の推定

■ EMアルゴリズムのまとめ

- ・ パラメータの初期化

- $\theta = \{\mu_k, \Sigma_k, \pi_k\}_{k=1}^K$ を適当な方法(乱数など)で初期化する。

- ・ E-ステップ

- θ を固定して α を最適化する。

- $\hat{\alpha}_k^{(n)} = \frac{\pi_k p(x^{(n)}; \mu_k, \Sigma_k)}{\sum_{k=1}^K \pi_k p(x^{(n)}; \mu_k, \Sigma_k)}$

- ・ M-ステップ

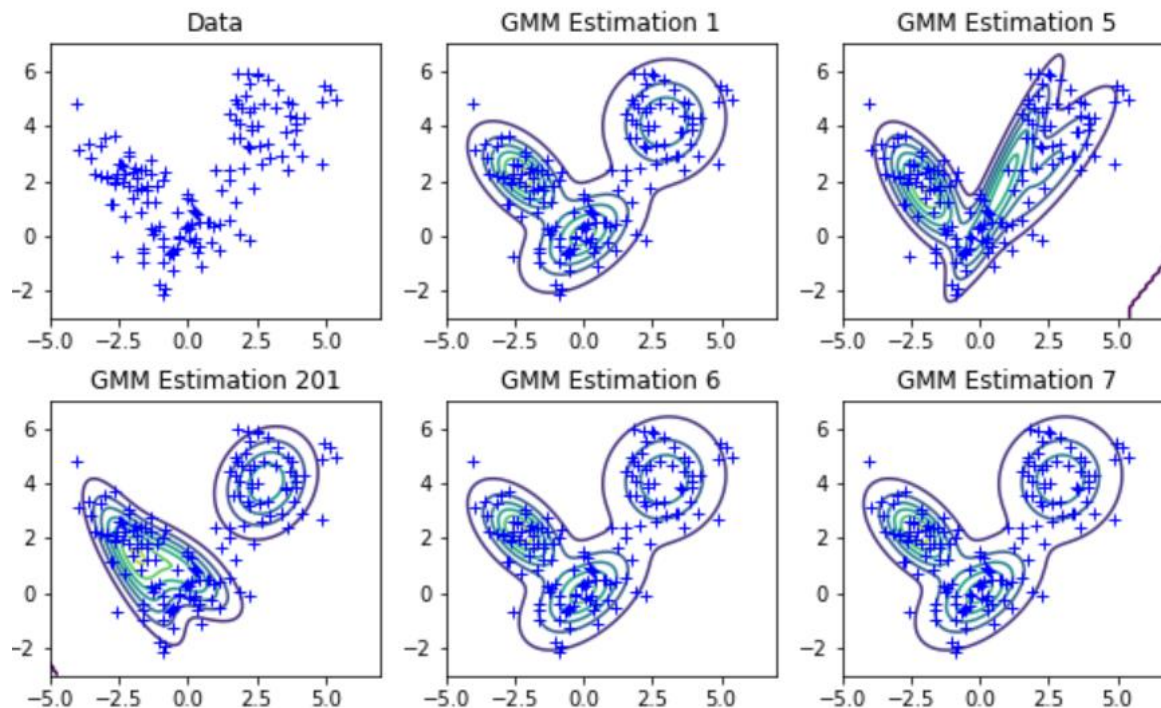
- α を固定して θ を最適化する。

- $\pi_k = \frac{\sum_{n=1}^N \alpha_k^{(n)}}{\sum_{n=1}^N \sum_{k=1}^K \alpha_k^{(n)}}, \hat{\mu}_k = \frac{\sum_{n=1}^N \alpha_k^{(n)} x^{(n)}}{\sum_{n=1}^N \alpha_k^{(n)}}, \Sigma_k = \frac{\sum_{n=1}^N \alpha_k^{(n)} (x^{(n)} - \mu)(x^{(n)} - \mu)^T}{\sum_{n=1}^N \alpha_k^{(n)}}$

- ・ E, Mステップをパラメータが収束するまで繰り返す。

【参考】EMアルゴリズムの解は局所最適解

- EMアルゴリズムで得られるパラメータ θ は局所最適解
 - ・ 一般に大域的な最適性は保証されない。
 - ・ EMアルゴリズムでは、パラメータの初期値によって推定される密度関数の形状が異なる。
 - ・ 実用上は、色々な初期値からEMアルゴリズムを回して得られた複数の局所解の中から一番良いものを選ぶ。



講義内容

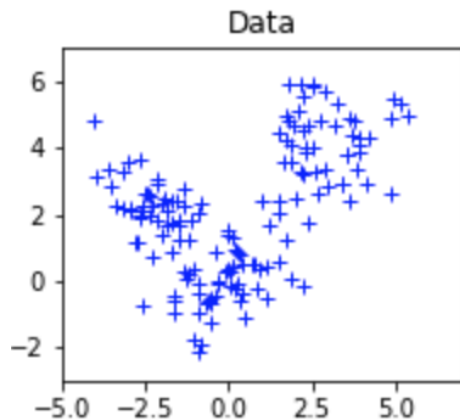
- 確率密度関数の推定
 - ・ 正規分布
 - 最尤推定
 - ・ 混合正規分布
 - EMアルゴリズム
 - 応用: クラスタリング
- 異常検知
 - ・ 確率密度関数を使った異常検知

混合正規分布の応用: クラスタリング

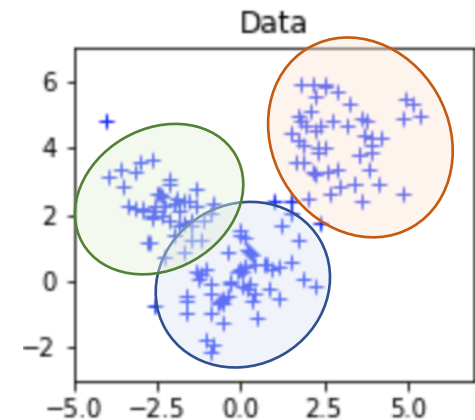
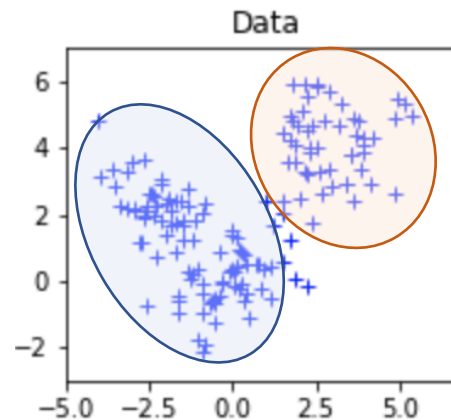
- 通常のカテゴリ分け問題 (教師あり学習) では、データ点 x が所属するカテゴリ y が教師信号として与えられている。
- 教師なし学習では教師信号はない。
 - ・ しかし、データの背後には何らかのカテゴリ分けが暗黙的に存在する、と想定される場合がある。
 - 例: ニュース記事の背後には、「政治」や「経済」、「スポーツ」といったカテゴリ分けがある。
- クラスタリング: データの背後の暗黙のカテゴリ分けの推定
 - ・ データ $D = \{x^{(n)} \in \mathbb{R}^d\}_{n=1}^N$ を K 個のデータの部分集合 (クラスタ) に分割する。
 - 多くの場合、各クラスタ内のデータは”似ている”と期待される。
 - そのため、”似ている”データ点同士を一つのクラスタにまとめる処理をする。

混合正規分布の応用: クラスタリング

- クラスタリング: データの背後の暗黙のカテゴリ分けの推定
 - データ $D = \{x^{(n)} \in \mathbb{R}^d\}_{n=1}^N$ を K 個のデータの部分集合 (クラスタ) に分割する。
 - 多くの場合、各クラスタ内のデータは”似ている”と期待される。
 - そのため、”似ている”データ点同士を一つのクラスタにまとめる処理をする。



クラスタリング



混合正規分布の応用: クラスタリング

■ 潜在変数モデル

- 実際に観測される変数の他に、観測されない変数(潜在変数)の存在を仮定したモデル。

■ 潜在変数モデルとしての混合正規分布

- 潜在変数 $z \in \{1, 2, \dots, K\}$: データ点 x が何番目の正規分布から生成されたか。
 - $z = k$ のとき、データ点 x は k 番目の正規分布から生成。
 - $p(x|z = k) = p(x; \mu_k, \Sigma_k)$
 - k 番目の正規分布が選ばれる確率を $p(z = k) = \pi_k$ とする。
- 混合正規分布
 - 実際には z の値は観測できないので、周辺化することで消去する。

$$p(x; \theta) = \sum_{k=1}^K p(z = k) p(x|z = k) = \sum_{k=1}^K \pi_k p(x; \mu_k, \Sigma_k)$$

これは混合正規分布そのもの

混合正規分布の応用: クラスタリング

■ ベイズの定理による潜在変数の推定

- ・ データ点 x が何番目の正規分布から生成されたかを推定する。

$$p(z = k|x) = \frac{p(z = k)p(x|z = k)}{\sum_{k=1}^K p(z = k)p(x|z = k)}$$

- ・ $p(z = k|x)$ が最大の正規分布からデータ点 x が生成されたと考えられる。

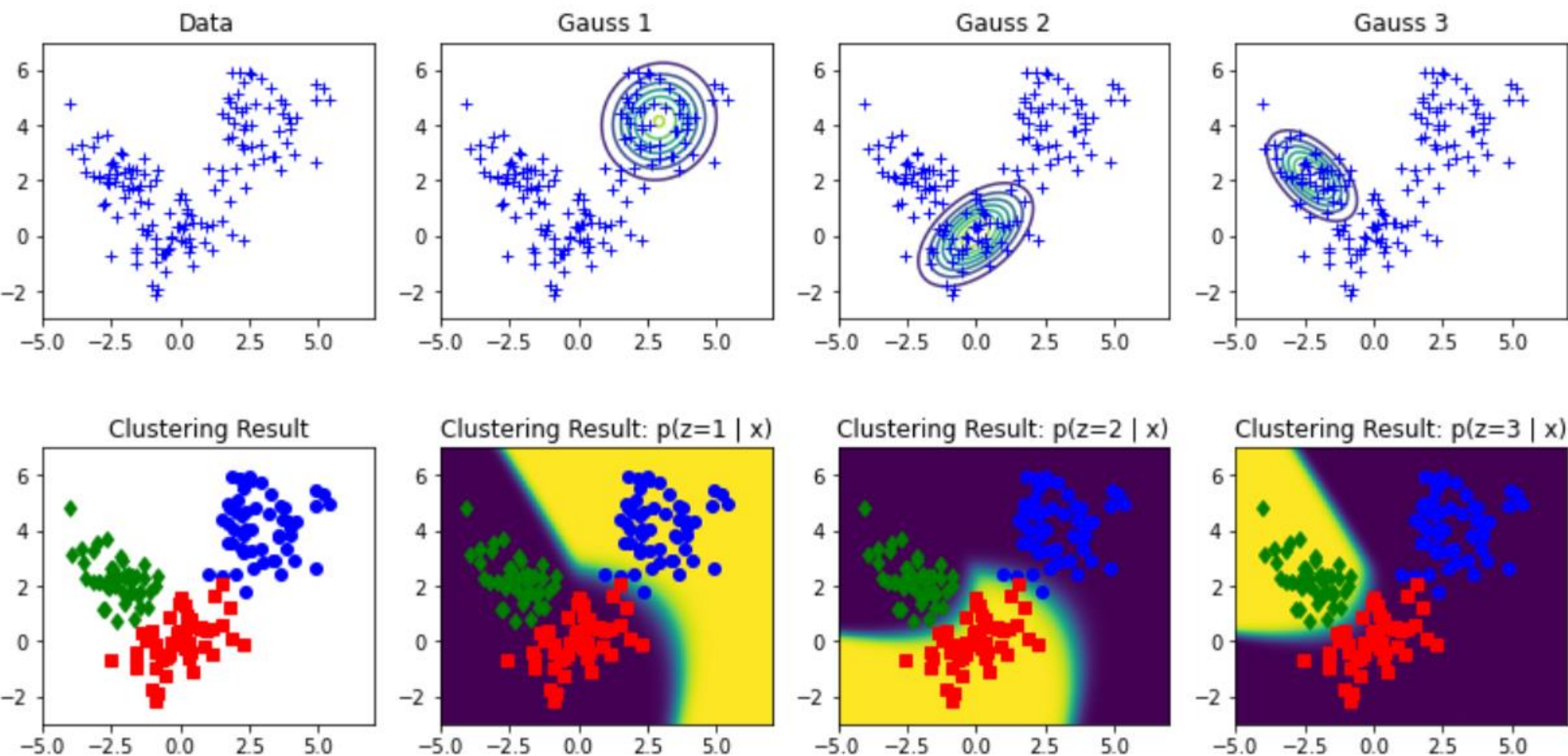
■ 混合正規分布を使ったクラスタリング

- ・ 同じ正規分布から生成されたと推定されたデータ点同士を一つのクラスタにまとめる。

混合正規分布の応用: クラスタリング

■ 混合正規分布を使ったクラスタリング

- 混合正規分布の k 番目の正規分布について、 $p(z = k|x)$ が最大のデータ点 x を一つのクラスタとみなす。



【参考】混合正規分布に基づくクラスタリングの拡張

- 混合正規分布は正規分布を組み合わせたもの
- 対象のデータの性質に応じて、正規分布以外の分布に置き換えることで、色々なデータのクラスタリングができる。
 - ・ 混合ディリクレ分布などは、文書のクラスタリングでよく使われる。
- 混合xx分布は、特にベイズ的な機械学習との相性が良い。
 - ・ ベイズ的な機械学習では、ベイズの定理を元に潜在変数を多く含むモデルを扱える。

講義内容

■ 確率密度関数の推定

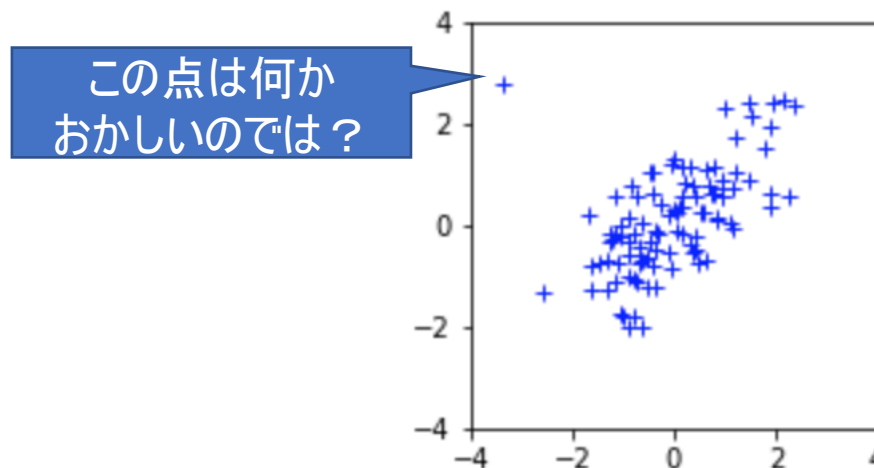
- ・ 正規分布
 - 最尤推定
- ・ 混合正規分布
 - EMアルゴリズム
 - 応用: クラスタリング

■ 異常検知

- ・ 確率密度関数を使った異常検知

異常検知(外れ値検知)

- 外れ値検知: データに変な値が含まれていないかを見つける問題。
- 外れ値検知は実データ分析において欠かせない技術である。
 - ・ 理由1: 通常の統計処理・データ分析は外れ値に弱い。
 - ・ 理由2: データに含まれる外れ値自体が有用な情報源である。



外れ値検知の重要性

- 通常の統計処理・データ分析は外れ値に弱い。
 - ・ 例：身体測定データの入力誤り

身長	体重
172cm	68kg
179cm	71kg
161cm	58kg
165cm	60kg
170cm	70kg

平均身長
169.4cm
平均体重
65.4kg

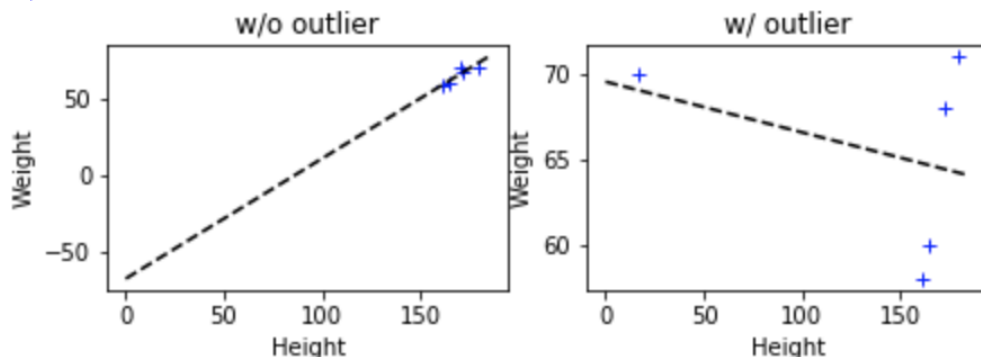
入力ミスでゼロ
が足りなかった

身長	体重
172cm	68kg
179cm	71kg
161cm	58kg
165cm	60kg
17cm	70kg

平均身長
138.8cm
平均体重
65.4kg

平均身長の見
積もりが大きく
ずれてしまう

- ・ 外れ値があると、回帰直線も想定外の推定がされてしまう。
- ・ 外れ値を見つけて除去することで適切なデータ分析ができる。

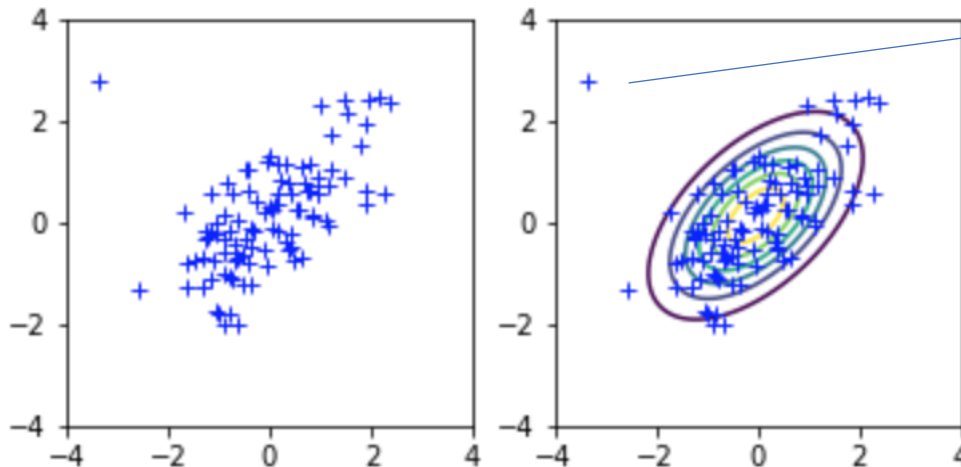


外れ値検知の重要性

- データに含まれる外れ値自体が有用な情報源である。
 - ・ 例：工場のセンサーデータ
 - 機械が通常に動作している時と、故障した時とでデータの振る舞いは異なる。
 - 外れ値の発生は故障の兆候である可能性が高い。
 - ・ 例：クレジットカードの利用履歴
 - 普段、大阪で本やテレビゲームの購入に使われているクレジットカードが、ある日ブラジルの電器店での高額家電の購入に使われた。
 - 外れ値の発生(=利用傾向の変化)は、カードの盗難が原因かもしれない。
 - ・ 外れ値が検知できれば、これら危険の兆候に適切に対処できる。

確率密度関数を使った異常検知

- 密度関数 $p(x)$ をデータセット D から推定する。
- データ点 x での密度関数の値 $p(x)$ がある値 δ を下回ったら、その点は外れ値であると判定する。
 - ・ 点 x の密度関数 $p(x)$ の値が小さい場合、これは x がデータセット D と大きく異なる「ほぼありえない点」であることを示している。そのような点は何かがおかしい(外れ値)と言える。



この点の密度は極めて小さい。
→ 外れ値である。

確率密度関数を使った異常検知

例. 正規分布を用いる場合

- 1. 正常なデータの集合 $D = \{x^{(n)}\}_{n=1}^N$ から、正規分布のパラメータを推定する。

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^N x^{(n)}, \quad \hat{\Sigma} = \frac{1}{N} \sum_{n=1}^N (x^{(n)} - \hat{\mu})(x^{(n)} - \hat{\mu})^\top$$

確率密度関数を使った異常検知

例. 正規分布を用いる場合

- 2. 新しい点 x について確率密度 $p(x; \hat{\mu}, \hat{\Sigma})$ を評価する。
 - ・ 負の対数尤度が扱いやすいので、 $-\log p(x; \hat{\mu}, \hat{\Sigma})$ を外れ値の度合いの指標とする。

$$-\log p(x; \hat{\mu}, \hat{\Sigma}) = \frac{1}{2} (x - \hat{\mu})^\top \hat{\Sigma}^{-1} (x - \hat{\mu}) + \frac{1}{2} \log (2\pi)^d \det \hat{\Sigma}$$

点 x が外れ値 \Leftrightarrow 密度 $p(x; \hat{\mu}, \hat{\Sigma})$ が小 $\Leftrightarrow -\log p(x; \hat{\mu}, \hat{\Sigma})$ が大

- ・ 点 x に依存しない項は定数とみなせるので除外。
以下の $a(x)$ を指標とする。

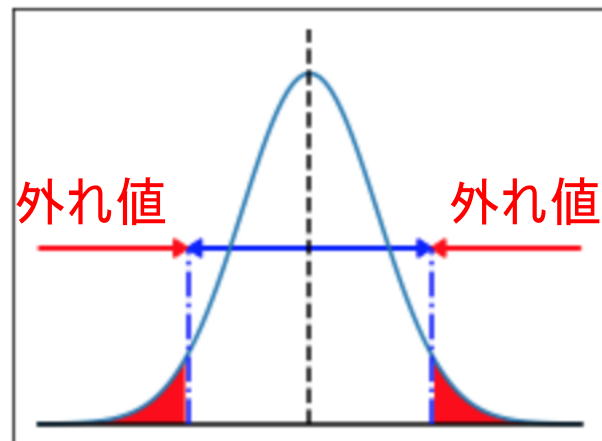
$$a(x) = (x - \hat{\mu})^\top \hat{\Sigma}^{-1} (x - \hat{\mu})$$

確率密度関数を使った異常検知

例. 正規分布を用いる場合

- 3. 指標 $a(x)$ が閾値 δ を超えたら点 x を外れ値と判定する。

$$a(x) = (x - \hat{\mu})^\top \hat{\Sigma}^{-1} (x - \hat{\mu}) \geq \delta$$



正常範囲

まとめ

- 確率密度関数の推定
 - ・ 教師なし学習の基盤技術
- 正規分布
 - ・ 最尤推定：対数尤度の最大化
- 混合正規分布
 - ・ EMアルゴリズム
 - 逐次的アルゴリズムによるパラメータ推定
 - ・ 応用：クラスタリング
 - 混合分布の事後分布としてのクラスタリング
- 確率密度関数を使った異常検知