# 宿題 1

2 つの正規分布にそれぞれ従う 2 次元点群を 2 セット用意し，正規分布のパラメータを既知として識別関数を設計して，これらの点群を分類する。識別関数は，

- Case 1: $\boldsymbol{\Sigma}_i = \sigma^2 \boldsymbol{I}$
- Case 2: $\boldsymbol{\Sigma}_i = \boldsymbol{\sigma}$
- Case 3: $\boldsymbol{\Sigma}_i =$ arbitrary

の 3 ケースについて考える。

## 理論

ガウスモデル

$$p(\boldsymbol{x}|y) = \frac{1}{(2\pi)^{d/2}\det(\boldsymbol{\Sigma})^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{1}$$

を考えると，ベイズの定理より，

$$p(y|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y)p(y)}{p(\boldsymbol{x})} \tag{2}$$

$$\log(p(y|\boldsymbol{x})) = \log(p(x|y)) + \log(p(y)) - \log(p(x)) \tag{3}$$

$$= -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log(\det(\boldsymbol{\Sigma})) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) + \log(p(y)) - \log(p(\boldsymbol{x})) \tag{4}$$

$$= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) + \log(p(y)) - \frac{1}{2}\log(\det(\boldsymbol{\Sigma})) + \log(p(y)) + C \tag{5}$$

$$C = -\frac{d}{2}\log(2\pi) - \log(p(\boldsymbol{x})) \tag{6}$$

となる。いま，二値分類を考えると，識別関数 $g$ は，

$$g(\boldsymbol{x}) = \log(p(y=1|\boldsymbol{x})) - \log(p(y=2|\boldsymbol{x})) \tag{7}$$

$$= -\frac{1}{2}\left((\boldsymbol{x}-\boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) - (\boldsymbol{x}-\boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right) - \frac{1}{2}\log\left(\frac{\det(\boldsymbol{\Sigma}_1)}{\det(\boldsymbol{\Sigma}_2)}\right) + \log\left(\frac{p_1}{p_2}\right) \tag{8}$$

となる。この $g$ が正であれば予測は 1，負なら 2 である。

Case 1 のとき，$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \sigma^2\boldsymbol{I}$ から，

$$g(\boldsymbol{x}) = \frac{1}{\sigma^2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{x} + \left(-\frac{1}{2}(||\boldsymbol{\mu}_1||^2 - ||\boldsymbol{\mu}_2||^2) + \log\left(\frac{p_1}{p_2}\right)\right) \tag{9}$$

Case 2 のとき，$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \Sigma$ から，

$$g(\boldsymbol{x}) = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{x} + \left(-\frac{1}{2}(\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2) + \log\left(\frac{p_1}{p_2}\right)\right) \tag{10}$$

Case 3 のとき，

$$g(\boldsymbol{x}) = -\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\boldsymbol{x} + (\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}_2^{-1})\boldsymbol{x}$$
$$+ \left(\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_2 - \frac{1}{2}\log\left(\frac{\det(\boldsymbol{\Sigma}_1)}{\det(\boldsymbol{\Sigma}_2)}\right) + \log\left(\frac{p_1}{p_2}\right)\right) \tag{11}$$

## 条件設定

2 つの正規分布にそれぞれ従う 2 次元点群（以下では dataset と呼ぶ）は 3 種類用意した。それぞれの概要は表 1 にまとめた。

表 1: 3 つの dataset の概要

| dataset | 1 | 2 | 3 |
|---|---|---|---|
| 総点数 | 1000 | 1000 | 1000 |
| 点群 1 に属する点の数 | 307 | 415 | 517 |
| 点群 2 に属する点の数 | 693 | 585 | 483 |
| 点群 1 の平均 | $\begin{bmatrix}2\\0\end{bmatrix}$ | $\begin{bmatrix}2\\2\end{bmatrix}$ | $\begin{bmatrix}1\\1\end{bmatrix}$ |
| 点群 2 の平均 | $\begin{bmatrix}-2\\0\end{bmatrix}$ | $\begin{bmatrix}-2\\-2\end{bmatrix}$ | $\begin{bmatrix}1\\1\end{bmatrix}$ |
| 点群 1 の共分散行列 | $\begin{bmatrix}1&0\\0&9\end{bmatrix}$ | $\begin{bmatrix}5&0\\0&6\end{bmatrix}$ | $\begin{bmatrix}1&0\\0&2\end{bmatrix}$ |
| 点群 2 の共分散行列 | $\begin{bmatrix}1&0\\0&2\end{bmatrix}$ | $\begin{bmatrix}6&0\\0&4\end{bmatrix}$ | $\begin{bmatrix}8&0\\0&5\end{bmatrix}$ |

## プログラム

プログラムの本体は**??**ページの listing **??**に示す。以下に含まれる関数の簡単な説明を記載する。

- dataset1
  2 つの正規分布にそれぞれ従う 2 次元点群を生成する。点の総数，そのうちの点群 1 の割合，各正規分布の平均と共分散行列を設定し，必要な数だけサンプリングする。点群 1，点群 2 とそれぞれの平均・共分散行列を返す。dataset2，dataset3 についても同様。
- sampling_normal_dist
  正規分布の平均・共分散行列と生成したい数を入れると，そこから点群をサンプリングする関数。
- classifier_binary
  二値分類問題の識別関数。点群と係数行列を受け取り，識別関数の値を返す。
- classify_binary_1
  Case 1 について点群を二値分類する関数。識別結果の他，係数行列も返す。classify_binary_2，classify_binary_3 はそれぞれ Case 2，Case 3 用である。
- measure_accuracy
  識別結果と点群のラベル（1 か 2），および点の総数から正解数と精度を算出する。
- sampling_normal_dist_for_contour
  点群に対し正規分布の等高線を描くための値を計算する。
- plot_data
  点群・等高線・識別境界線をプロットする。
- main

実行用の関数。case 変数で各 Case を入れ替えられる。また，dataset1() 関数を他のものに入れ替えることで，dataset も入れ替えられる。

## 結果

以下に，識別の結果として精度と散布図を示す。散布図には点群 1 と 2 がともにプロットされており，求めた識別境界線も描かれている。プロット中では，群 1 を丸，群 2 をバツ印で表し，群 1 のうち正しく群 1 に分類されたものをオレンジ（darkorange），誤って群 2 に分類されたものを赤（red），群 2 のうち正しく群 2 に分類されたものを薄い青（royalblue），誤って群 1 に分類されたものを青（blue）で表す。また，その識別境界線を緑（darkcyan）で表す。

## 考察

紙面の都合上，各 dataset についての結果を示す前に，簡単な考察を示す。2 分布間に被りの少ない，dataset 1 のような設定では，Case 1 のような単純なモデルで十分分類できた。しかし，2 分布が共に含む領域を持つような dataset 2 では，当然分類の精度は低くなり，Case 3 を用いても，改善はしたがあまり大きくはなかった。2 つの正規分布の平均が一致するような設定である dataset 3 については，Case 1，Case 2 では全ての点が点群 1 に分類される結果となった。これは，$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \equiv \boldsymbol{\mu}$ かつ $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 \equiv \boldsymbol{\Sigma}$ のとき，式 (8) は

$$g(\boldsymbol{x}) = \log\left(\frac{p_1}{p_2}\right) \tag{12}$$

となり，識別関数がどんな $\boldsymbol{x}$ に対しても事前分布の大きい方のみに分類するようなものとなってしまうからである。なお，Case 3 ではこの問題は解消されることが結果からもわかる。

## dataset 1 の結果

表 2: dataset1 に対する識別の結果

| | Case 1 | Case 2 | Case 3 |
|---|---|---|---|
| 点群 1 の識別精度 | 302 | 302 | 298 |
| 点群 1 の識別精度 | 0.984 | 0.984 | 0.971 |
| 点群 2 の識別精度 | 688 | 688 | 692 |
| 点群 2 の識別精度 | 0.993 | 0.993 | 0.999 |
| 全体の正解率 | 0.990 | 0.990 | 0.990 |

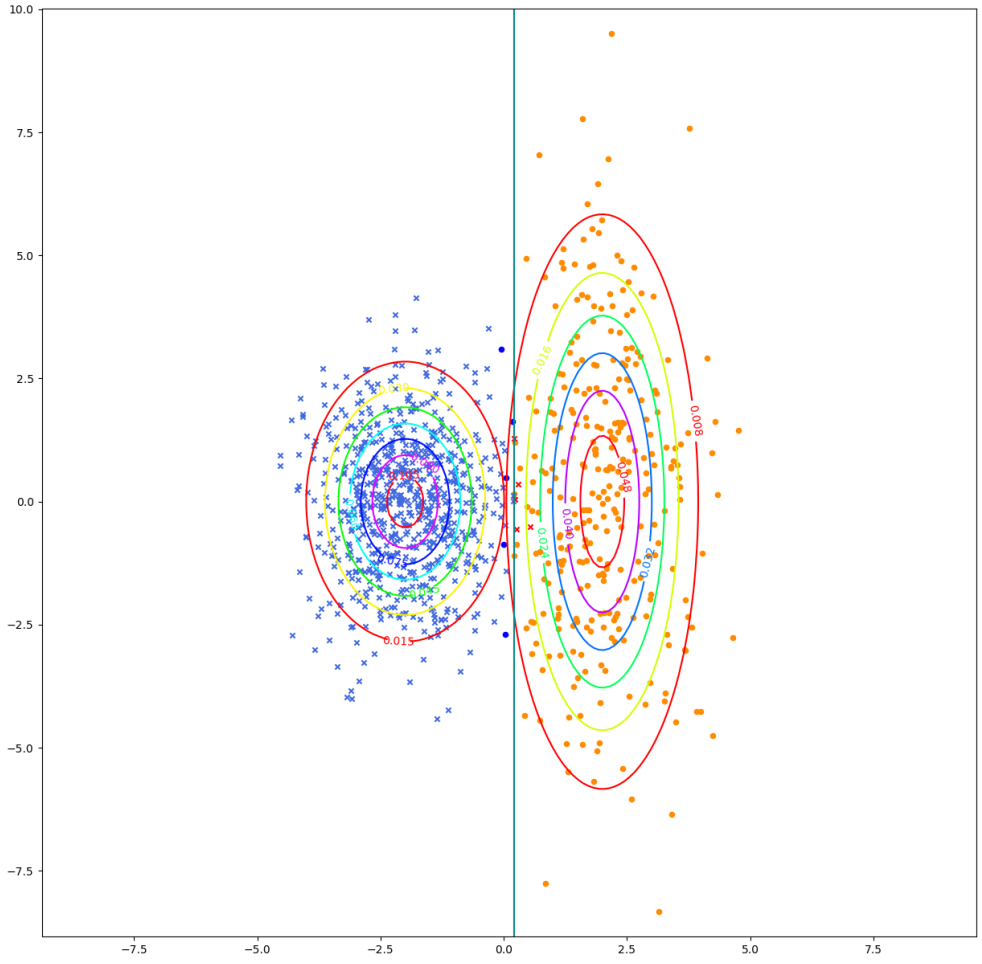


図 1: dataset 1 に対して Case 1 の識別関数を適用したときの散布図と識別境界線

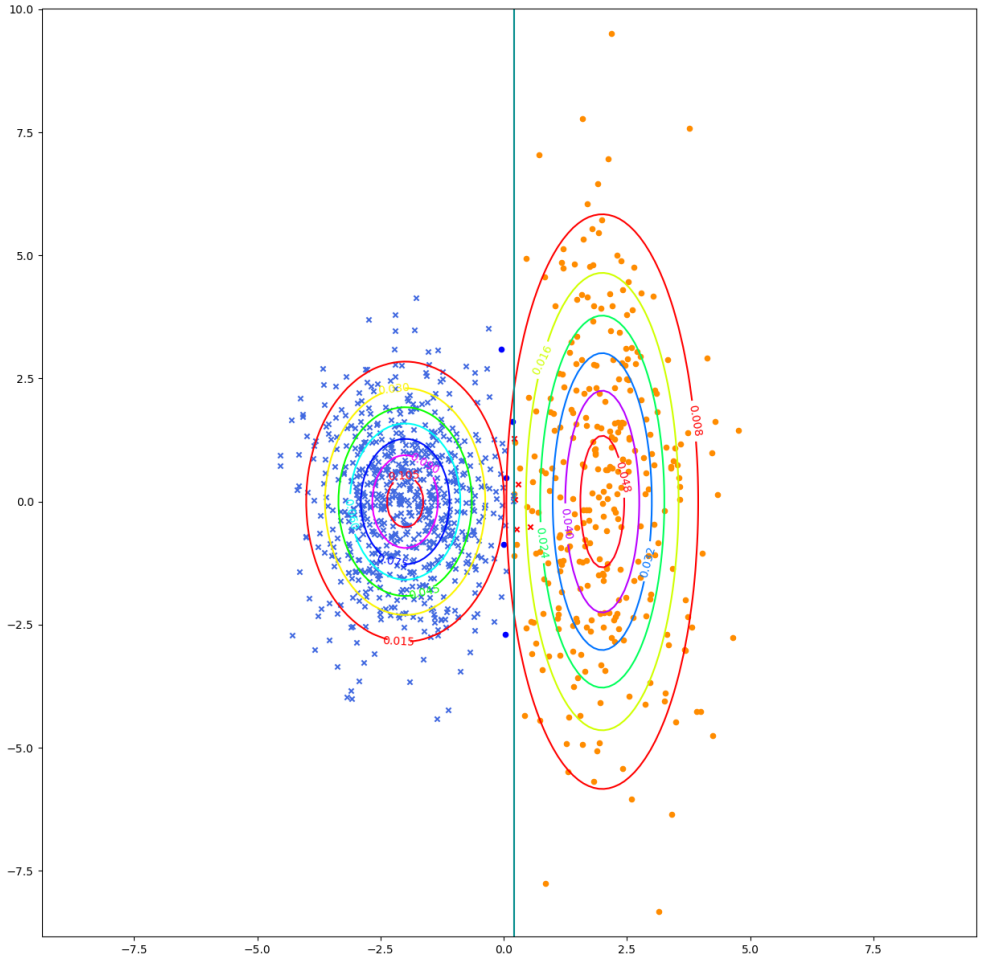


図 2: dataset 1 に対して Case 2 の識別関数を適用したときの散布図と識別境界線

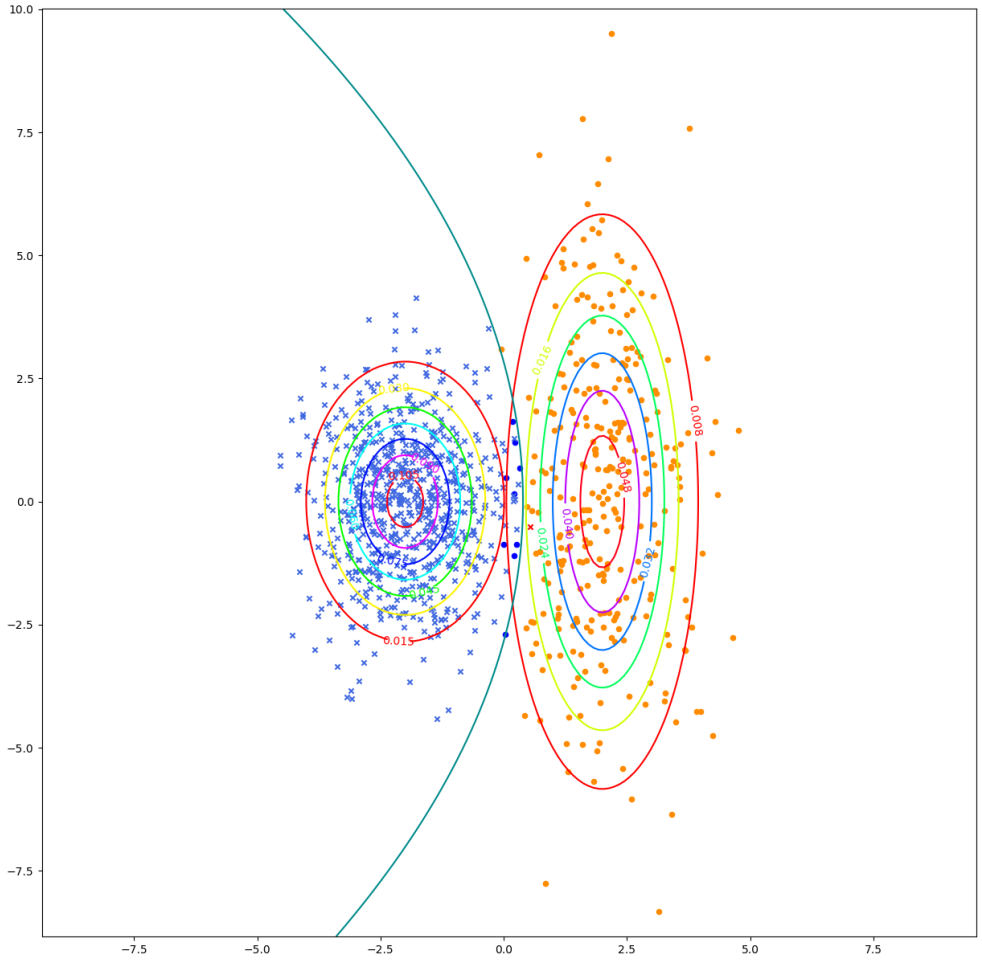


図 3: dataset 1 に対して Case 3 の識別関数を適用したときの散布図と識別境界線

## dataset 2 の結果

表 3: dataset2 に対する識別の結果

| | Case 1 | Case 2 | Case 3 |
|---|---|---|---|
| 点群 1 の識別精度 | 370 | 357 | 361 |
| 点群 1 の識別精度 | 0.892 | 0.860 | 0.870 |
| 点群 2 の識別精度 | 540 | 549 | 550 |
| 点群 2 の識別精度 | 0.923 | 0.938 | 0.940 |
| 全体の正解率 | 0.910 | 0.906 | 0.911 |

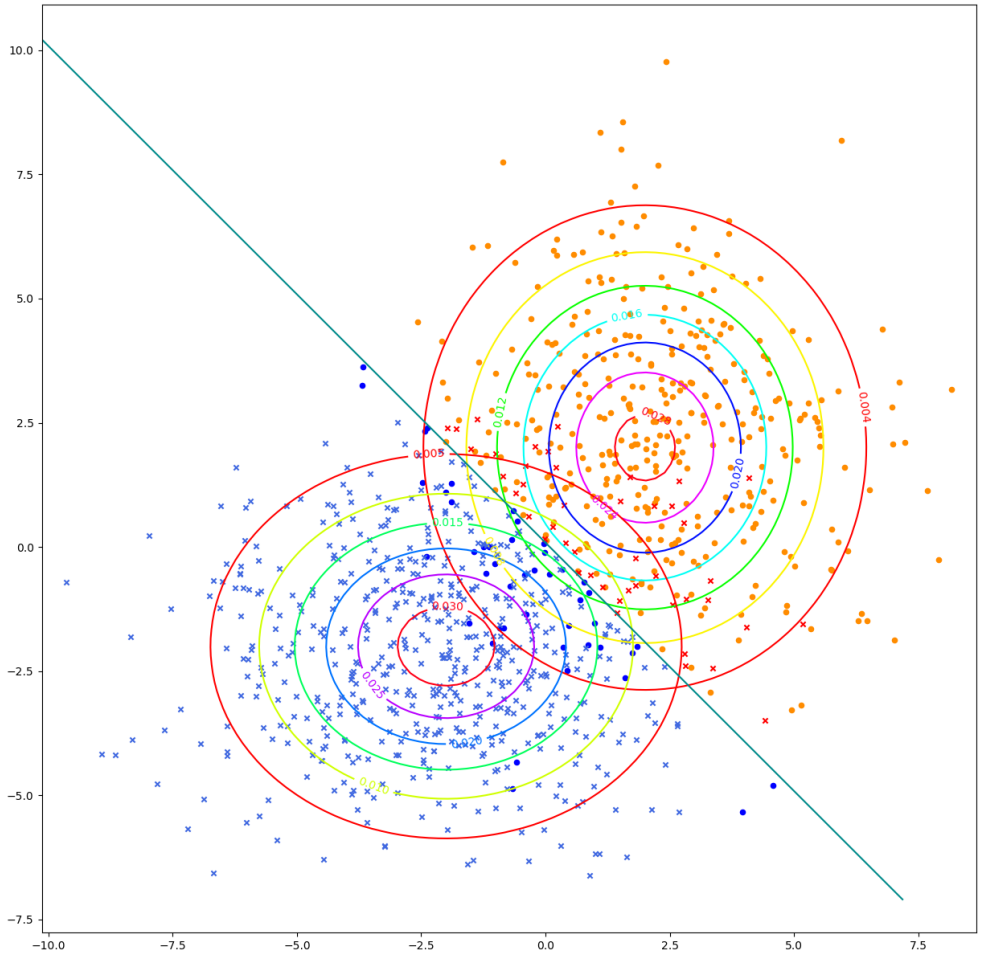


図 4: dataset 2 に対して Case 1 の識別関数を適用したときの散布図と識別境界線

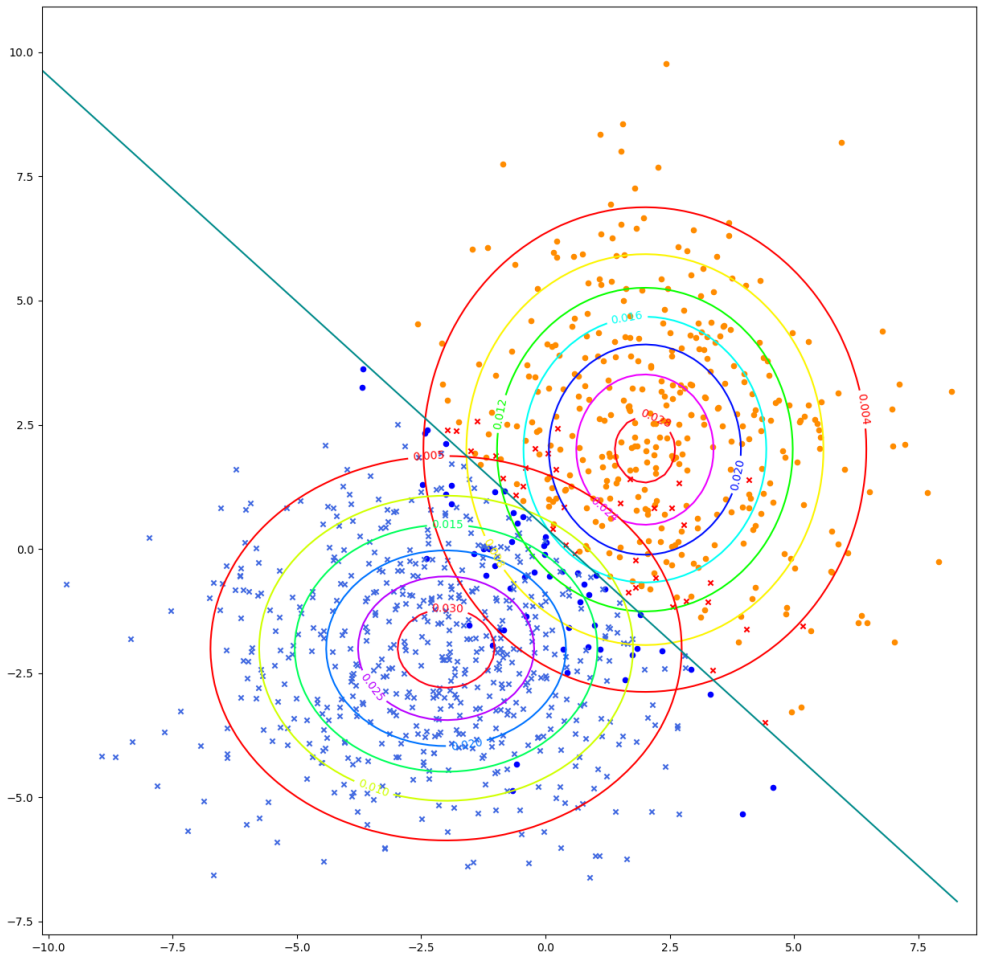


図 5: dataset 2 に対して Case 2 の識別関数を適用したときの散布図と識別境界線

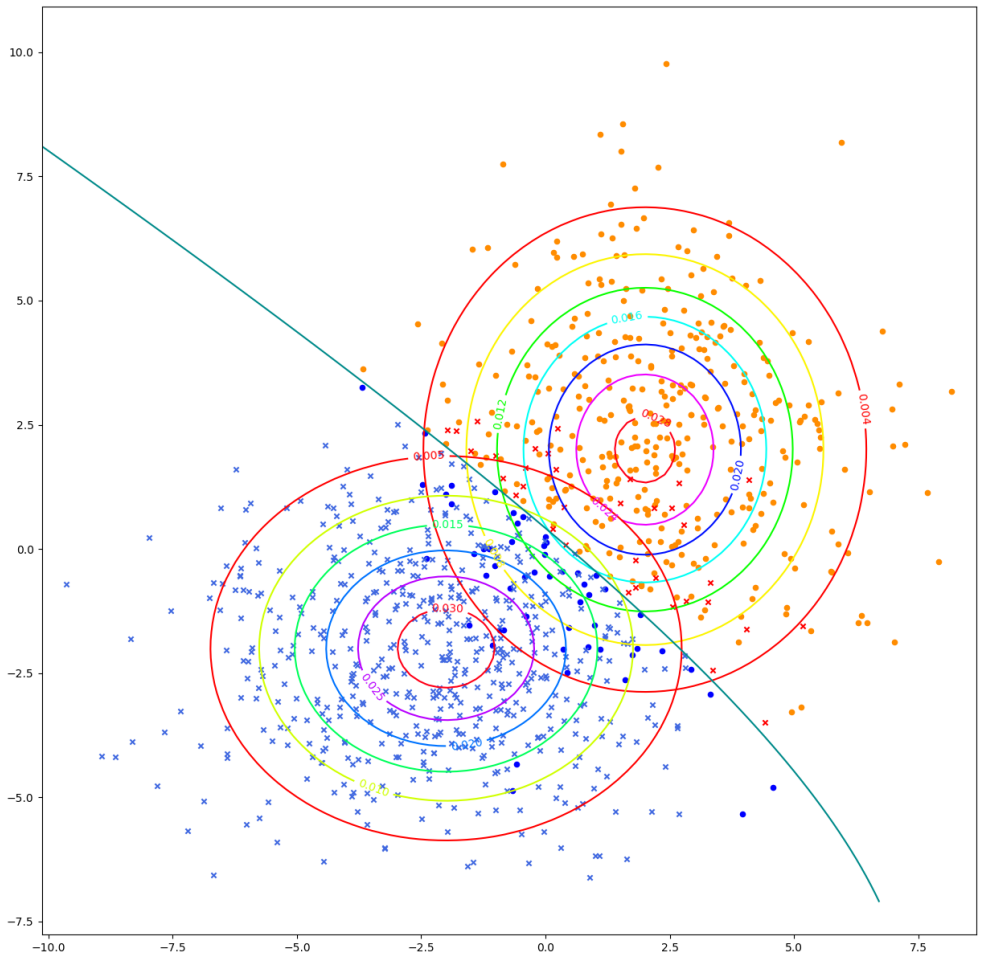


図 6: dataset 2 に対して Case 3 の識別関数を適用したときの散布図と識別境界線

## dataset 3 の結果

表 4: dataset3 に対する識別の結果

| | Case 1 | Case 2 | Case 3 |
|---|---|---|---|
| 点群 1 の識別精度 | 517 | 517 | 465 |
| 点群 1 の識別精度 | 1.000 | 1.000 | 0.899 |
| 点群 2 の識別精度 | 0 | 0 | 295 |
| 点群 2 の識別精度 | 0.000 | 0.000 | 0.611 |
| 全体の正解率 | 0.517 | 0.517 | 0.760 |

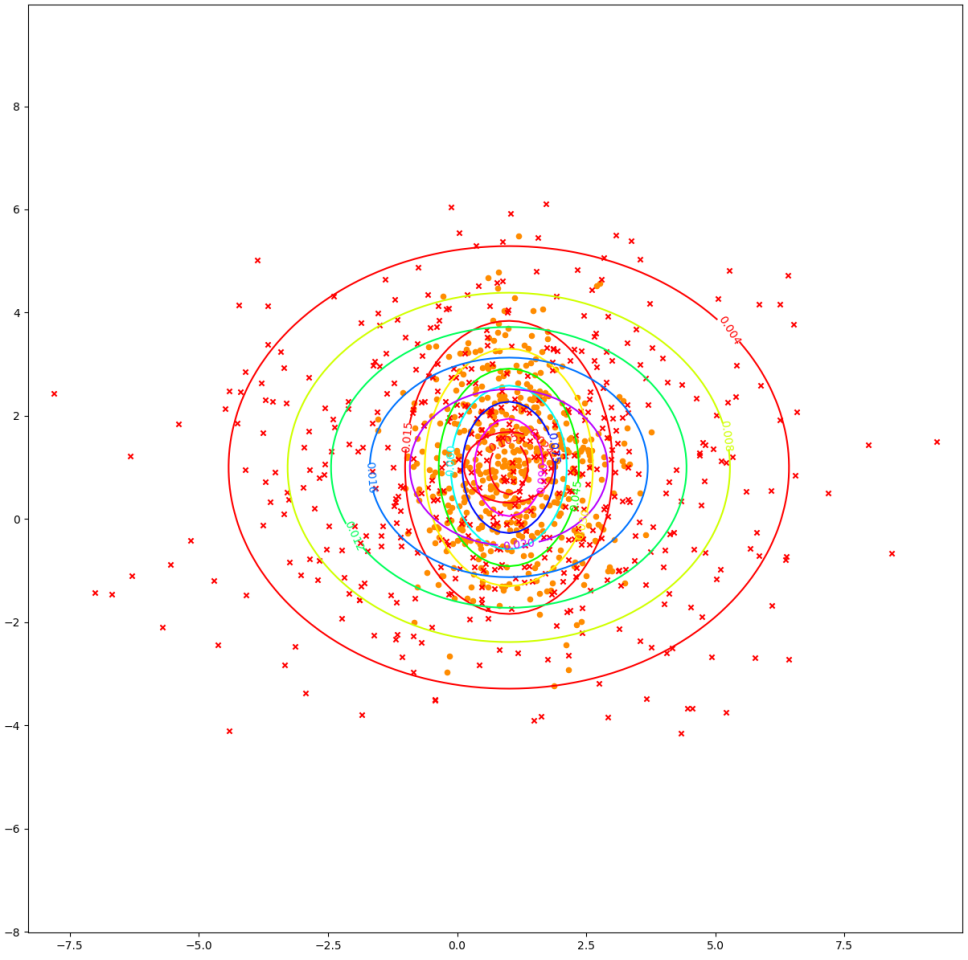


図 7: dataset 3 に対して Case 1 の識別関数を適用したときの散布図と識別境界線

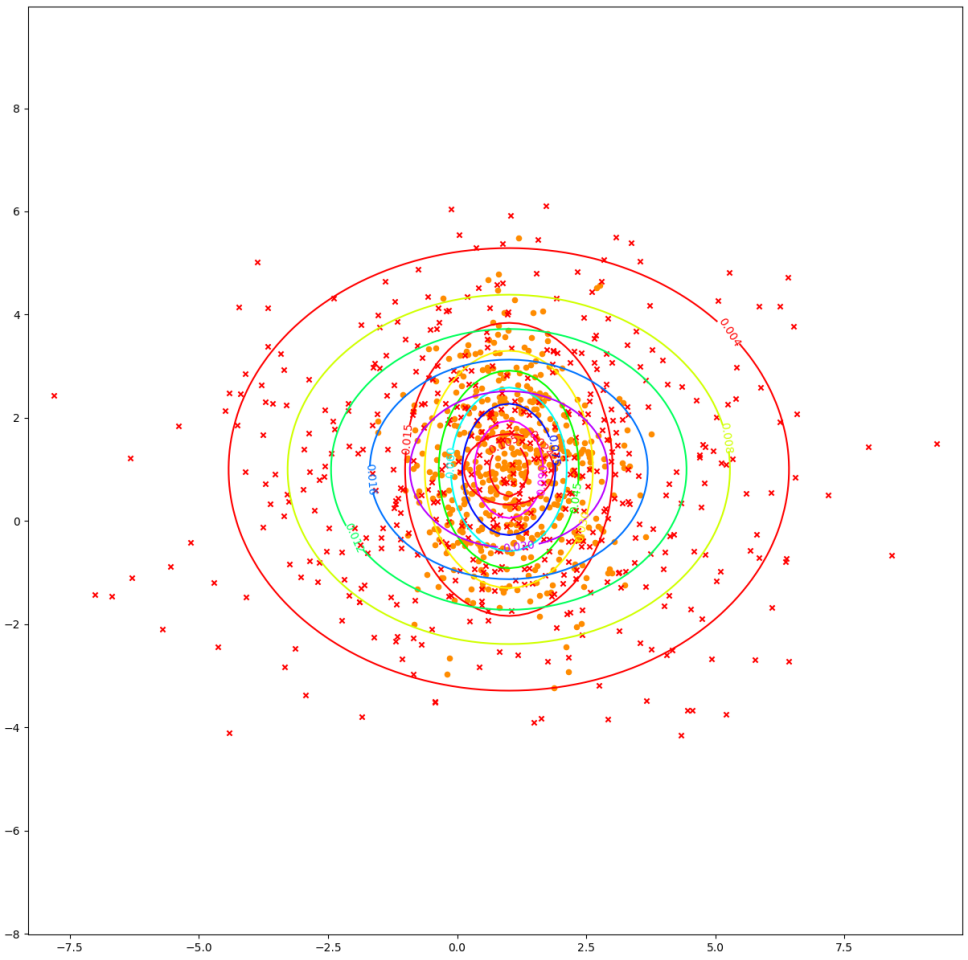


図 8: dataset 3 に対して Case 2 の識別関数を適用したときの散布図と識別境界線

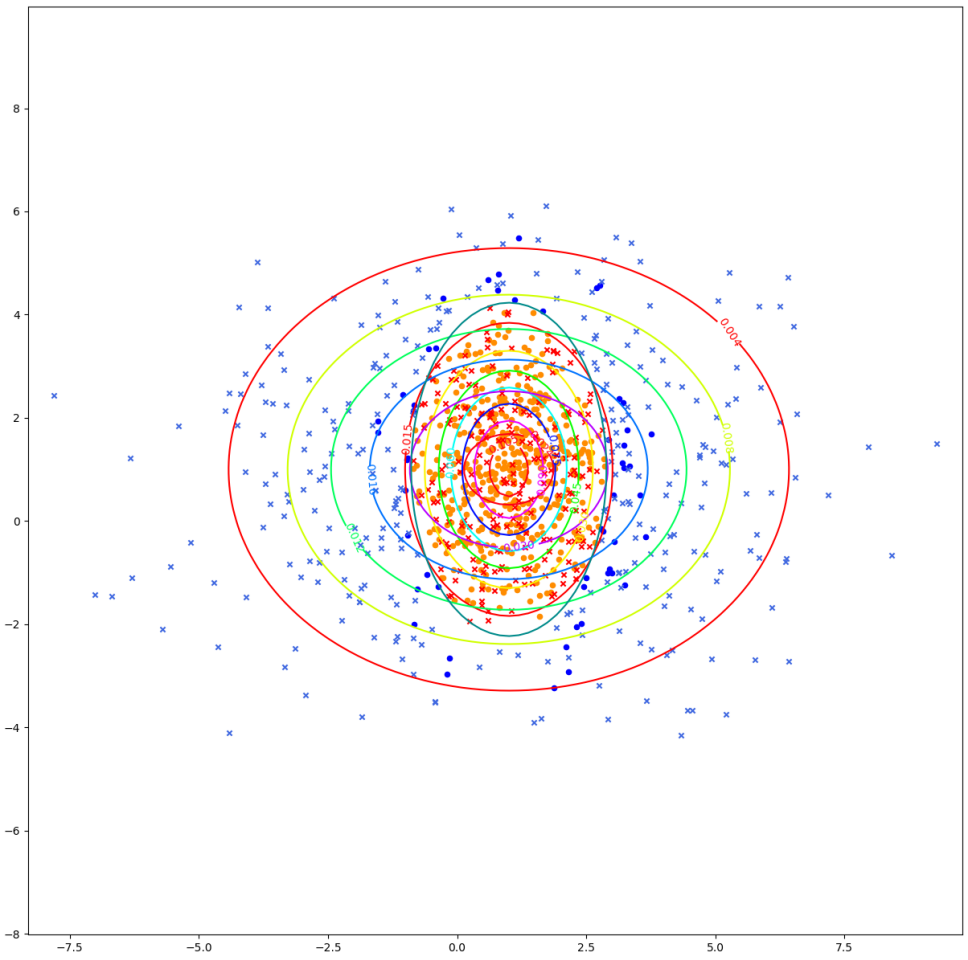


図 9: dataset 3 に対して Case 3 の識別関数を適用したときの散布図と識別境界線