1.

$$\underset{(\bar{w}, b)}{\operatorname{argmin}} \|\bar{w}\|^2 \text{ s.t. } \forall i \;\; y_i(\langle \bar{w}, \bar{x}_i \rangle + b) \geq 1$$

צ"ל: $Q, A, \bar{a}, \bar{d}$ כך שהבעיה שקולה ל:

$$\underset{v \in \mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2}\bar{v}^T Q \bar{v} + a^T \bar{v} \text{ s.t. } A\bar{v} \leq \bar{d}$$

$$\|w\|^2 = \langle w, w \rangle = w^T w = w^T I w$$

$$y_i(\langle \bar{w}, \bar{x}_i \rangle + b) \geq 1 \iff y_i \langle \bar{w}, \bar{x}_i \rangle \geq 1 - b y_i$$

$$\iff y_i \langle \bar{x}_i, \bar{w} \rangle \geq 1 - b y_i$$

$$\iff -y_i \langle \bar{x}_i, \bar{w} \rangle \leq b y_i - 1$$

$$A = \begin{pmatrix} — & -y_1 \bar{x}_1 & — \\ & \vdots & \\ — & -y_n \bar{x}_n & — \end{pmatrix}$$

$$\bar{d} = \begin{pmatrix} b y_1 - 1 \\ \vdots \\ b y_n - 1 \end{pmatrix} \qquad Q = I \quad \bar{a} = \bar{0}$$

ונקבל:

$$\frac{1}{2} v^T Q v + a^T \bar{v} = \frac{\|w^*\|^2}{2}$$

וגם

$$A\bar{v} = \begin{pmatrix} -y_1 \langle \bar{x}_1, \bar{w} \rangle \\ \vdots \\ -y_n \langle \bar{x}_n, \bar{w} \rangle \end{pmatrix}$$

1. המשך. ולכן

$$\underset{v\in\mathbb{R}^n}{\operatorname{argmin}} \ \frac{1}{2}v^TQv+\bar{d}^Tv \quad s.t. \quad Av\le d =$$

$$= \underset{v\in\mathbb{R}^n}{\operatorname{argmin}} \ \frac{\|v\|^2}{2} \quad s.t. \quad y_i(\langle\bar{w},\bar{x}_i\rangle+b)\geq 1 =$$

$$\overset{w=v}{=} \underset{w\in\mathbb{R}^n}{\operatorname{argmin}} \ \|w\|^2 \quad s.t. \quad y_i(\langle\bar{w},\bar{x}_i\rangle+b)\geq 1$$

$w=v$
והמינימיזציה
של $\frac{f}{2}$
היא זהה
למינימיזציה
של f

מ.ש.ל.

2

$$\underset{\bar{w},\{\varepsilon_i\}}{\operatorname{argmin}} \ \frac{\lambda}{2}\|\bar{w}\|^2 + \frac{1}{m}\sum_i \varepsilon_i \quad s.t. \ \forall i \ y_i\langle\bar{w},\bar{x}_i\rangle \geq 1-\varepsilon_i \wedge \varepsilon_i \geq \qquad \text{(i)}$$

$$\ell^{hinge}(a) = \max\{0, 1-a\}$$

נר' שקולות ל:

$$\underset{\bar{w},\{\varepsilon_i\}}{\operatorname{argmin}} \ \frac{\lambda}{2}\|\bar{w}\|^2 + \frac{1}{m}\sum_i \ell^{hinge}(y_i\langle\bar{w},\bar{x}_i\rangle) \qquad \text{(ii)}$$

יהי $\bar{w}$ שפותר מינ' של (ii).

עבור $\bar{w}' = \bar{w}$, $\varepsilon_i = 1-y_i\langle\bar{w},\bar{x}_i\rangle$ נקבל ש:

$$\frac{\lambda}{2}\|\bar{w}\|^2 + \frac{1}{m}\sum_i \varepsilon_i = \frac{\lambda}{2}\|\bar{w}\|^2 + \frac{1}{m}\sum_i \ell^{hinge}(1-y\langle\bar{w},\bar{x}_i\rangle)$$

נניח בשלילה שקיימים $\tilde{w}, \{\tilde{\varepsilon}_i\}$ כך שהם נותנים ערך נמוך יותר עבור (i). אזי:

$$\frac{\lambda}{2}|\tilde{w}|^2 + \frac{1}{m}\sum_i \tilde{\varepsilon}_i < \frac{\lambda}{2}\|\bar{w}\|^2 + \frac{1}{m}\sum_i \ell^{hinge}(1-y\langle\bar{w},\bar{x}_i\rangle$$

לכן

$$\frac{\lambda}{2}|\tilde{w}|^2 + \frac{1}{m}\sum_i \ell^{hinge}(1-y_i\langle\tilde{w},\bar{x}_i\rangle) \underset{(1-y_i\langle\tilde{w},\bar{x}_i\rangle)\leq\tilde{\varepsilon}_i}{\leq} \frac{\lambda}{2}|\tilde{w}|^2 + \frac{1}{m}\sum_i \tilde{\varepsilon}_i$$

ולכן

$$\frac{\lambda}{2}|\tilde{w}|^2 + \frac{1}{m}\sum_i \ell^{hinge}(1-y_i\langle\tilde{w},\bar{x}_i\rangle) \leq \frac{\lambda}{2}|\bar{w}|^2 + \frac{1}{m}\sum_i \ell^{hinge}(1-y_i\langle\bar{w},\bar{x}_i\rangle)$$

בסתירה למינימליות $\bar{w}$. מ.ש.ל. חלק ראשון

2. הוכחה: יהיו $\bar{w},\{\epsilon_i\}$ מינימליים עבור (i).

$$\forall i \quad \epsilon_i \geq 1 - y_i\langle \bar{w}, \bar{x}_i\rangle \ \wedge\ \epsilon_i \geq 0$$

נניח בשלילה שקיים $\tilde{w}$ כך שהוא נותן ערך נמוך יותר עבור (ii):

$$\underset{\bar{w},\{\epsilon_i\}}{\operatorname{argmin}} \ \frac{\lambda}{2}|\tilde{w}|^2 + \frac{1}{m}\sum_i l^{hinge}(y_i\langle \tilde{w}, \bar{x}_i\rangle) < \frac{\lambda}{2}|\bar{w}|^2 + \frac{1}{m}\sum_i l^{hinge}(y_i\langle \bar{w}, \bar{x}_i\rangle) \leq$$

$$\leq \frac{\lambda}{2}|\bar{w}|^2 + \frac{1}{m}\sum_i \epsilon_i$$

$\epsilon_i \geq 1 - y_i\langle \bar{w}, x_i\rangle$

$\epsilon_i \geq$

ולכן $\frac{\lambda}{2}|\tilde{w}|^2 + \frac{1}{m}\sum_i l^{hinge}(y_i\langle \tilde{w}, \bar{x}_i\rangle) < \frac{\lambda}{2}|\bar{w}|^2 + \frac{1}{m}\sum_i \epsilon_i$

כך ש- $\tilde{w},\{l^{hinge}(y_i\langle \tilde{w}, \bar{x}_i\rangle)\}$ מקבלים ערך נמוך יותר עבור (i)

ולכן $\bar{w},\{\epsilon_i\}$ אינם מינימליים, סתירה.

מ.ש.ל.

3. א.

$$y \sim Mult(\pi)$$

$$x_j | y=k \sim N(\mu_{kj}, \sigma^2_{kj})$$

$$L(\theta|\mathbf{x},\mathbf{y}) = f_{\mathbf{x},\mathbf{Y}|\theta}(\{(x_i, y_i)\}_{i=1}^m) =$$

$$\overset{ind}{=} \prod_{i=1}^m f_{X|Y=y_i}(x_i) \cdot f_{Y|\theta}(y_i) =$$

$$= \prod_{i=1}^m N(x_i | \mu_{k1}, \sigma^2_{k1}, k=y_i) \cdot Mult(y_i|\pi) =$$

$$= \prod_{i=1}^m \frac{1}{\sqrt{(2\pi)^1 \sigma^2_{y_i}}} \exp\left(-\frac{1}{2}(x-\mu_{y_i 1}) \frac{1}{\sigma^2_{y_i 1}} (x - \mu_{y_i 1})\right) \cdot \pi_{y_i}$$

נעבור ל log-likelihood כמו בתרגול:

$$l(\theta|x,y) = \sum_i \log(\pi_{y_i}) - \frac{1}{2}\sum_i \log(2\pi) - \frac{1}{2}\sum_i \frac{(x-\mu_{y_i 1})^2}{\sigma_{y_i 1}^2} - \frac{1}{2}\sum_i \log(\sigma_{y_i})$$

נסמן ונקבל:

$$= \sum_k \left[n_k \log(\pi_k) - \frac{1}{2} \sum_{i: y_i = k} \frac{(x-\mu_k)^2}{\sigma_k^2} \right] - \frac{m}{2}\log(2\pi) - \frac{1}{2}\log(\sigma_k) \cdot n_k$$

נגדיר $g(\pi) = \sum_k \pi_k - 1$

ואז $\mathcal{L} = l(\theta|x,y) - \lambda g(\pi)$

נגזור ונשווה: (משתמש בג"מ).

3.א. המשך.

$$\frac{\partial L}{\partial \pi_k} = \frac{n_k}{\pi_k} - \lambda = 0 \Rightarrow \pi_k = \frac{n_k}{\lambda} = \boxed{\frac{n_k}{m}}$$

$$1 = \sum_k \pi_k = \sum_k \frac{n_k}{\lambda} \Rightarrow \boxed{\lambda = m}$$

$$\hat{\pi}_k^{MLE} = \frac{n_k}{m}$$ לק

מינוס מגזירה לפי μ_k:

$$\sum_{i:y_i=n_k} x_i - \mu_k = 0 \Rightarrow \frac{1}{n_k}\sum_i \mathbb{1}[y_i=k]x_i = \hat{\mu}_k^{MLE}$$

מינוס:

$$\frac{\partial L}{\partial \sigma_k} = 0 \Rightarrow +\frac{1}{2}\sum_{i:y_i=k}\frac{(x_i-\mu_k)^2}{(\sigma_k^3)} - \frac{n_k}{2\sigma_k} = 0$$

$$\cancel{\frac{1}{2}\sum_k\frac{(x-\mu_k)^2}{\sigma_k^3} - \frac{n_k}{2\sigma_k} = 0}$$

$$\sum_{i:y_i=k}\frac{(x_i-\mu_k)^2}{1} = \sigma_k^2 n_k$$

$$\cancel{(x-\mu_k)^2 = \sigma_k^2}$$

$$\cancel{\boxed{\hat{\sigma}_k^{MLE} = |x-\mu_k|}}$$

$$\boxed{\hat{\sigma}^2_{k_{MLE}} = \frac{1}{n_k}\sum_{i:y_i=k}(x_i-\mu_k)^2}$$

מ.3

$y \sim Mult(\pi) \qquad \bar{x} \in \mathbb{R}^d$

$x_j | y=k \sim N(\mu_{kj}, \sigma_{kj}^2)$

סעיף קודם

$$L(\theta, X, y) = f_{X,Y|\theta}\left(\{(\bar{x}_i, y_i)\}_{i=1}^{m}\right) = \cdots \overset{\downarrow}{=}$$

$$= \prod_{i=1}^{m} f_{X|Y=y_i}(x_i) \cdot f_{Y|\theta}(y_i) =$$

$$= \prod_{i=1}^{m} \pi_{y_i} \cdot N\left(\bar{x}_i \middle| \bar{\mu}_k, \operatorname{diag}(\sigma_{k1}^2, \dots, \sigma_{kd}^2), k=y_i\right)$$

מבונן מ

log-likelihood :

$$\ell(\theta | X, y) = \sum_i \left[\log(\pi_{y_i}) - \frac{d}{2}\log(2\pi) - \frac{1}{2}\log\left(\prod_{j=1}^{d} \sigma_{y_i j}\right) - \frac{1}{2}(\bar{x}_i - \bar{\mu}_{y_i})^T \operatorname{diag}\left(\frac{1}{\sigma_{y_i 1}^2}, \dots, \frac{1}{\sigma_{y_i d}^2}\right)(\bar{x}_i - \bar{\mu}_{y_i}) \right]$$

לגרנג׳:

$$\mathcal{L} = \ell(\theta | X, y) - \lambda g(\pi)$$

נגזור:

$$\frac{\partial \mathcal{L}}{\partial \pi_k} = 0 \Rightarrow \frac{n_k}{\pi_k} - \lambda = 0 \Rightarrow \hat{\pi}_k^{MLE} = \frac{n_k}{\lambda} = \boxed{\frac{n_k}{m}}$$

$\boxed{\lambda = m}$ מוסף

[illegible]:

$$\hat{\mu}_k^{MLE} = \frac{1}{n_k}\sum_i \mathbb{1}_{[y_i = k]} \bar{x}_i \Rightarrow \boxed{\hat{\mu}_{kj}^{MLE} = \frac{1}{n_k}\sum_i \mathbb{1}_{[y_i = k]} (\bar{x}_i)_j}$$

3.ב. המשך.

$$0=\frac{\partial \mathcal{L}}{\partial \sigma_{kj}}=\sum_{i:y_i=k}\left[-\frac{1}{2\sigma_{kj}}+\frac{1}{2}\left((x_i)_j-\mu_{kj}\right)\cdot\frac{1}{\sigma_{kj}^3}\cdot\left((x_i)_j-\mu_{kj}\right)\right]$$

$$\frac{n_k}{\cancel{2}\sigma_{kj}}=\frac{1}{\sigma_{kj}^3}\sum_{i:y_i=k}(x_{ij}-\mu_{kj})^2$$

$$\boxed{\hat{\sigma}^2_{kj_{MLE}}=\frac{1}{n_k}\sum_{i:y_i=k}(x_{ij}-\mu_{kj})^2}$$

תשובות לחלק הפרקטי

3.1) 1. ניתן להסיק שהמסווג מתכנס ל-0.
ז"א שהperceptron מצליח לסווג את
המדגם בין שתי סוגי הdata לאחר
מספר סופי (וקטן) של איטרציות.

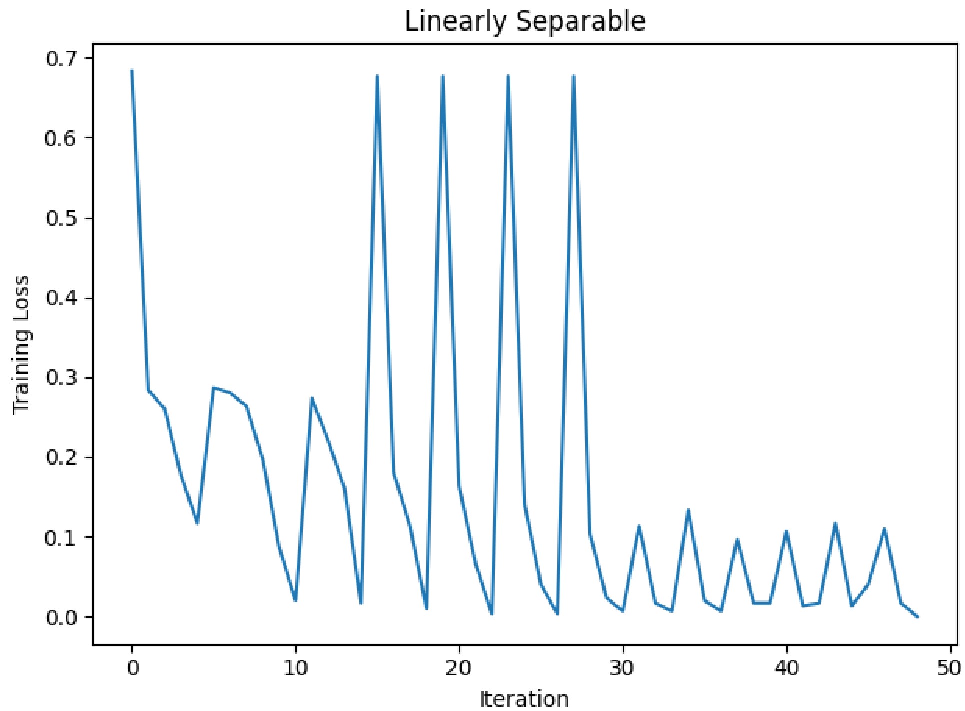
2. במקרה הלא ניתן להפרדה
ליניארית ניתן לראות שהמודל
לעולם אינו מתכנס, ~~[illegible]~~
הוא ממשיך לצעוד קדימה
ואחורה ולא מאפס את
ה-Loss.

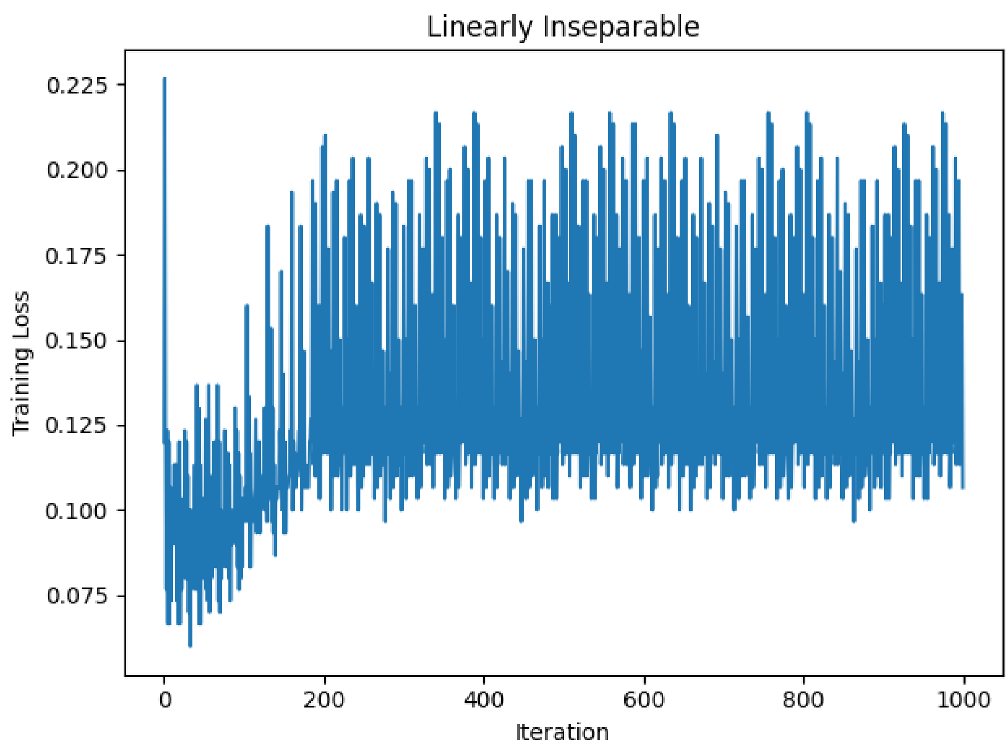
~~תשובות~~ לחלק הפרקטי

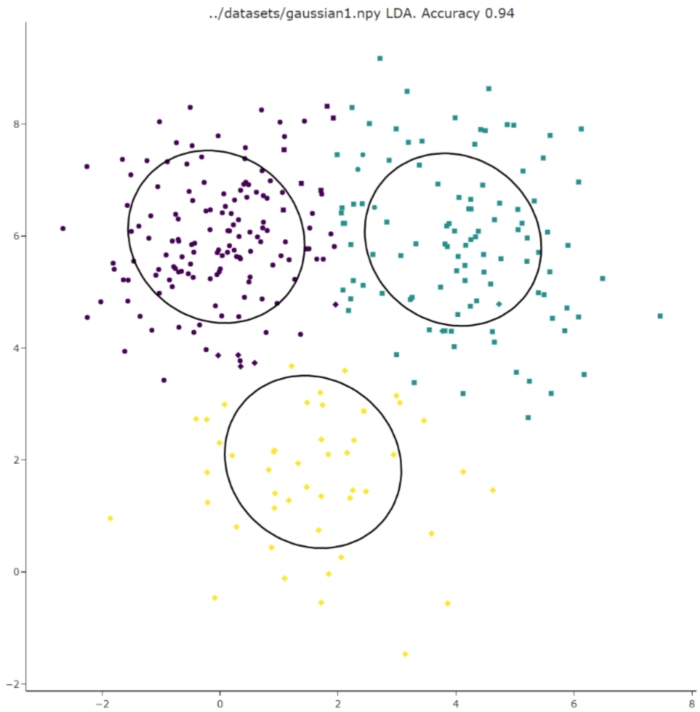
3.2) ~~[illegible]~~ 1. ניתן ללמוד מהגרפים שהמידע נוצר ע"י פיזור גאוסיאני סביב שלוש נקודות שונות עם cov-mats שונות ~~[illegible]~~ בעלות Variance גבוה ביחס לש ויחסים אחיד בצירי X וY.

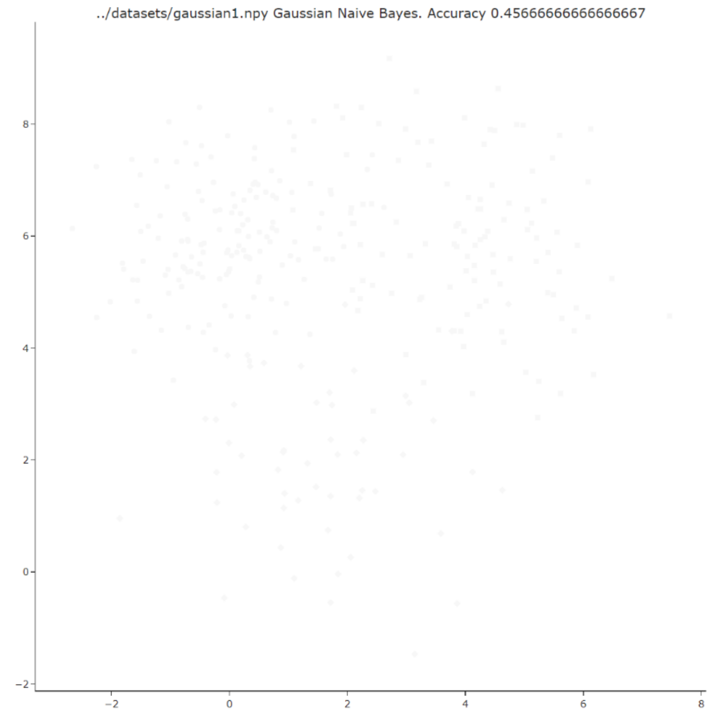
2. ה dataset הזה הותאם יותר טוב ע"י ה LDA אבל בעיקר בגלל שלא הספקתי לממש את GNB עד הסוף.

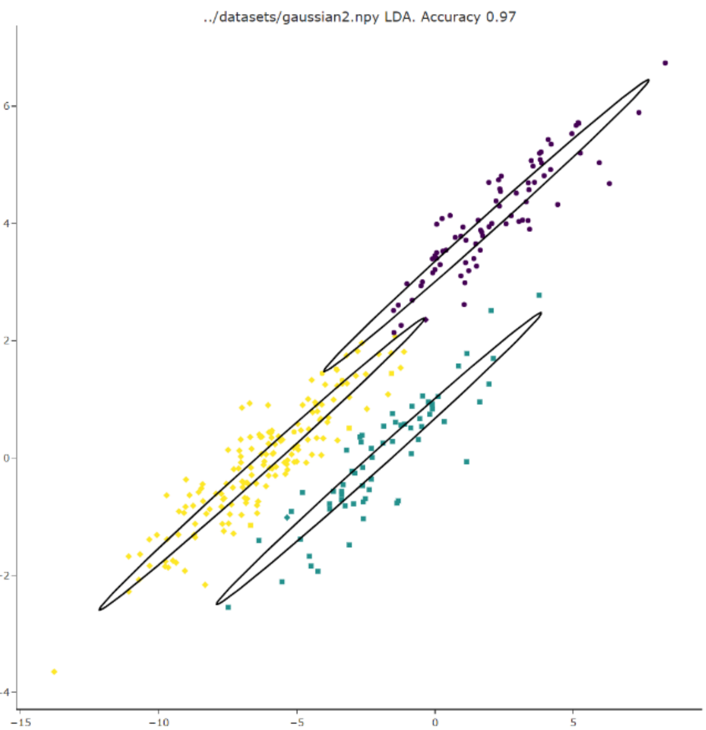
במידע זה יש Var גדול בציר אחד ונמוך במאונך לו, והנקודות מופרדות יחסית טוב.


Linearly Separable
Training Loss
Iteration
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
50


Linearly Inseparable
Training Loss
Iteration
0.225
0.200
0.175
0.150
0.125
0.100
0.075
0
200
400
600
800
1000


../datasets/gaussian1.npy LDA. Accuracy 0.94


../datasets/gaussian1.npy Gaussian Naive Bayes. Accuracy 0.45666666666666667


../datasets/gaussian2.npy LDA. Accuracy 0.97


../datasets/gaussian2.npy Gaussian Naive Bayes. Accuracy 0.27