Machine Learning HW2

Chutian Chen cc4515

1

(a)

$$p(w) = (2\pi)^{-\frac{k}{2}}\,\det(\boldsymbol{\Sigma})^{-\frac{1}{2}}\,e^{-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\mathbf{w}-\boldsymbol{\mu})}$$

$$\Sigma = \tau^2 I$$

$$Let\ L(w) = \left(\sum_{i=1}^{N} \ln p\left(y^{(i)}|\mathbf{x}^{(i)};\mathbf{w}\right)\right) + \ln p(\mathbf{w})$$

$$\frac{\partial L}{\partial w} = -\sum_{i=1}^{N} \frac{w^T x^{(i)} - y^{(i)}}{\sigma^2} x^{(i)} - \frac{1}{\tau^2} w$$

$$\text{Let } A = (x^{(1)}, x^{(2)}, .., x^{(N)}),\ y = (y^{(1)}, y^{(2)}, .., y^{(N)})^T$$

$$\frac{\partial L}{\partial w} = -A(\frac{A^T w - y}{\sigma^2}) - \frac{w}{\tau^2} = 0$$

So

$$w_{MAP} = (AA^T + \frac{\sigma^2}{\tau^2} I)^{-1} Ay$$

It's same as the form of the solution of ridge regression.

(b)

$$p(w_i) = \frac{1}{2b}\exp\left(-\frac{|w_i|}{b}\right)$$

So

$$lnp(w_i) = -\frac{|w_i|}{b} - ln(2b)$$

$$\begin{aligned}
\text{So }\ w_{MAP} &= \arg\max -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y^{(i)} - \hat{\mathbf{w}}^T\mathbf{x}^{(i)}\right)^2 - \frac{\|\hat{\mathbf{w}}\|_1}{b} \\
&= \arg\min \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y^{(i)} - \hat{\mathbf{w}}^T\mathbf{x}^{(i)}\right)^2 + \frac{\|\hat{\mathbf{w}}\|_1}{b} \\
&= \arg\min \frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - \hat{\mathbf{w}}^T\mathbf{x}^{(i)}\right)^2 + \frac{2\sigma^2}{Nb}\|\hat{\mathbf{w}}\|_1
\end{aligned}$$

It doesn't have a closed solution. But we can see it has the same form as lassor regression.

2

(a)

$$\hat{f}(\mathbf{x}) = \underset{y\in\{0,1\}}{\arg\max} \Pr(y|\mathbf{x}) = \underset{y\in\{0,1\}}{\arg\max} \underbrace{\pi_y}_{:=\Pr(y)} p(\mathbf{x}|y;\mu_y,\mathbf{\Sigma}_y)$$

$$p(\mathbf{x}|y;\mu_y,\mathbf{\Sigma}_y) = \frac{1}{(2\pi)^{d/2}|\Sigma_y|^{1/2}} \exp\left(\frac{-(\mathbf{x}-\mu_y)^T(\Sigma_y)^{-1}(\mathbf{x}-\mu_y)}{2}\right)$$

$$d_\Sigma(\mathbf{x},\mu) = (\mathbf{x}-\mu)^T(\Sigma)^{-1}(\mathbf{x}-\mu)$$

$$\hat{f}(\mathbf{x}) = 1[\Pr(y=1|\mathbf{x}) > \Pr(y=0|\mathbf{x})] = 1\left[\ln\frac{\Pr(y=1|\mathbf{x})}{\Pr(y=0|\mathbf{x})} > 0\right]$$

$$= 1\left[\ln\frac{\pi_1}{(1-\pi_1)} + \ln\frac{p(\mathbf{x}|y=1)}{p(\mathbf{x}|y=0)} > 0\right]$$

$$= 1\left[\ln\frac{\pi_1}{(1-\pi_1)} - \frac{1}{2}\ln\frac{|\Sigma_1|}{|\Sigma_0|} - \frac{1}{2}(d_{\Sigma_1}(\mathbf{x},\mu_1) - d_{\Sigma_0}(\mathbf{x},\mu_0)) > 0\right]$$

$$\text{Because } \Sigma_1 \neq \Sigma_0, \text{ the boundry is quadratic}$$

$$\text{If } \Sigma_1 = \Sigma_0, \; d_{\mathbf{\Sigma}_1}(\mathbf{x},\mu_1) - d_{\mathbf{\Sigma}_0}(\mathbf{x},\mu_0) = \mathbf{x}^T(\mathbf{\Sigma}_1)^{-1}(\mu_1-\mu_0) - \frac{1}{2}\mu_1^T(\mathbf{\Sigma}_1)^{-1}\mu_1 + \frac{1}{2}\mu_0^T(\mathbf{\Sigma}_1)^{-1}\mu_0$$

So the boundry becomes linear.

(b)

$$For\ (\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix})where\ (x_k \in \{0,1\}\ ),\ assume$$

$$\Pr(\mathbf{x}|y) = \prod_{k=1}^{d} \Pr[x_k|y]$$

Then we can make the classifier a Gaussian Naive Bayes Classifier.

Bayes Classifier:

$$\hat{f}(\mathbf{x}) = \underset{y\in\mathcal{Y}}{\arg\max} \Pr(y|\mathbf{x}) = \underset{y\in\mathcal{Y}}{\arg\max} \prod_{k=1}^{d} \Pr[x_k|y] \cdot \Pr[y]$$

$$\text{Because every pair of } \mathrm{x_i} \text{ and } \mathrm{x_j} \text{ is independent on the condition y, } \Sigma \text{ is diagonal matrix.}$$

$$\Sigma_{\mathrm{i,i}} = \mathrm{Var}(\mathrm{x_i}) = \sigma_i^2$$

$$\Sigma_{ii}^{-1} = \frac{1}{\sigma_i^2}$$

$$
\begin{aligned}
\hat{f}(\mathbf{x}) = \underset{y\in\{0,1\}}{\arg\max}\Pr(y|\mathbf{x}) &= \underset{y\in\{0,1\}}{\arg\max}\ \pi_y \frac{1}{(2\pi)^{d/2}|\Sigma_y|^{1/2}}\exp\left(\frac{-(\mathbf{x}-\mu_y)^T(\Sigma_y)^{-1}(\mathbf{x}-\mu_y)}{2}\right) \\
&= \underset{y\in\{0,1\}}{\arg\max}\ \pi_y \frac{1}{(2\pi)^{d/2}\prod_{k=1}^{d}\sigma_k}\exp\left(\frac{-\sum_{k=1}^{d}\frac{(x_k-\mu_{yk})^2}{\sigma_i^2}}{2}\right) \\
&= \underset{y\in\{0,1\}}{\arg\max}\ \pi_y \prod_{k=1}^{d}\frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{(x_k-\mu_{yk})^2}{2\sigma_i^2}\right) \\
&= \underset{y\in\{0,1\}}{\arg\max}\prod_{k=1}^{d}\Pr[x_k|y]\Pr[y]
\end{aligned}
$$

We can see two classifiers are equavalent.

3

(a)

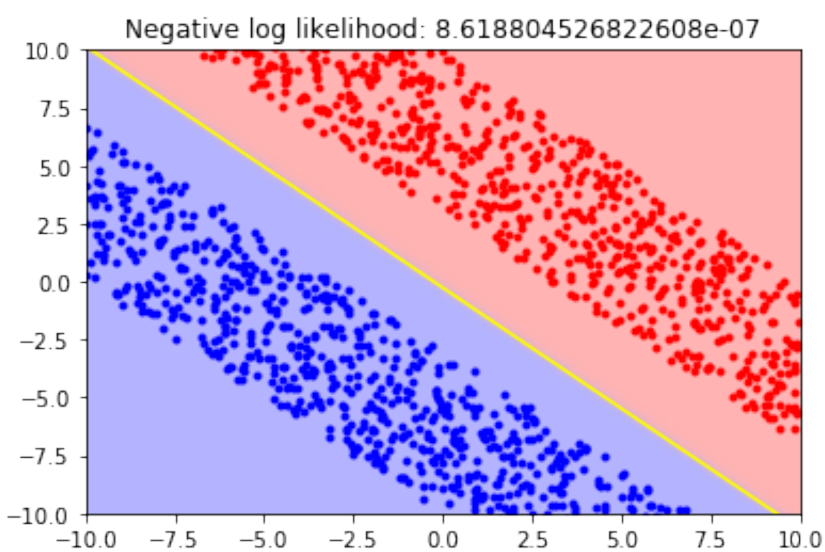
$$
\text{Because } \lambda \text{ is very large, } \sum_{i=1}^{N}\log\left(P\left(y^{(i)}|x^{(i)};w_0,w_1,w_2\right)\right)-\lambda\cdot w_j^2\approx-\lambda\cdot w_j^2
$$

This function reach maximum when wj=0. From figure 1 we can see that when w2=0, the train error is smallest.

When w1=0, the train error is largest.

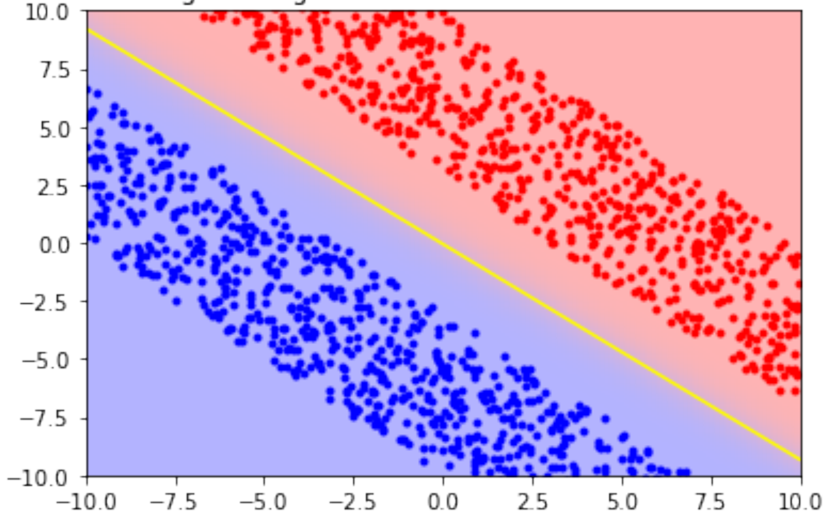
(b)

1. Without regularization

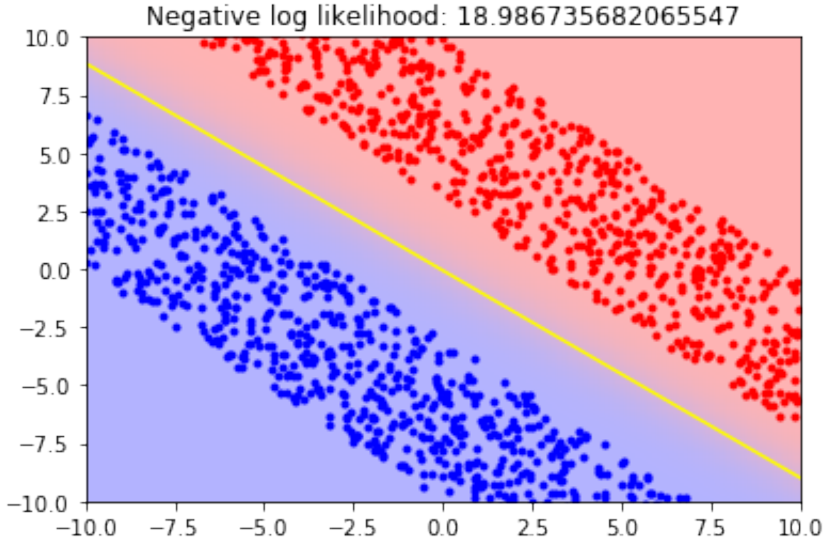


2. Regularize x1 (the order of pictures is the lambda from 1 to 100000)

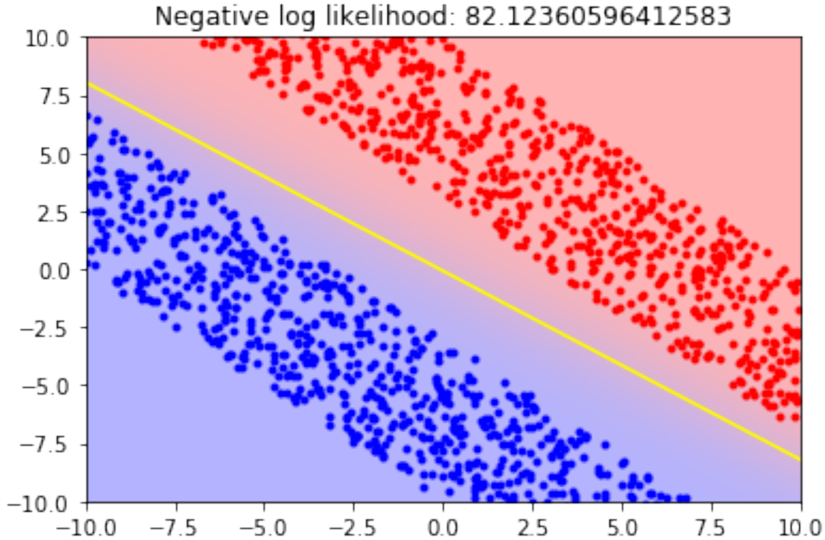
Negative log likelihood: 3.6571731512123398

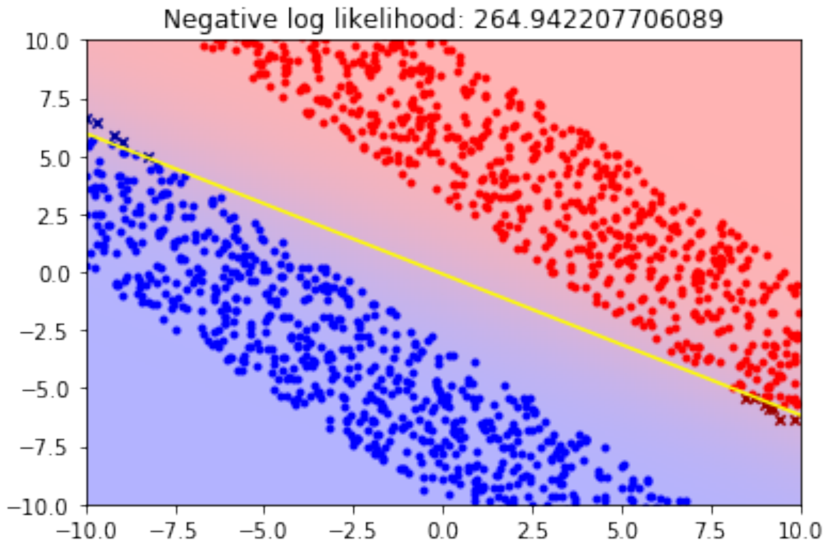


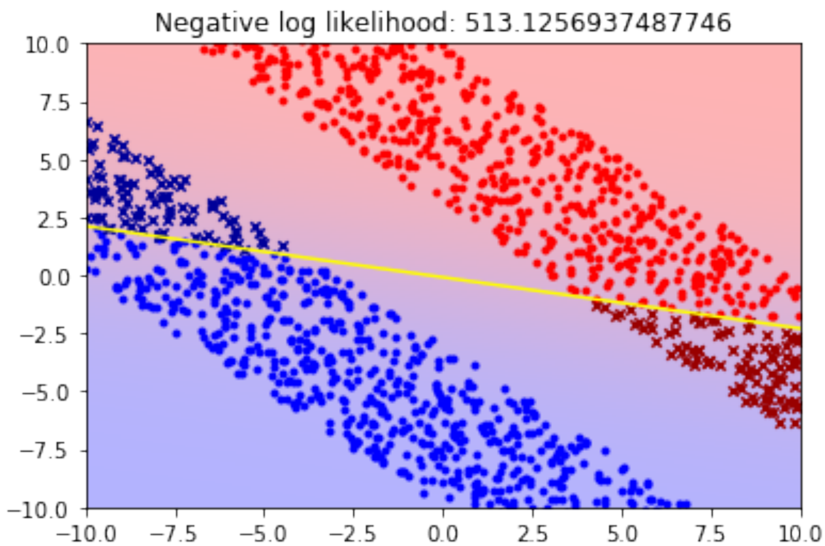
Negative log likelihood: 18.986735682065547

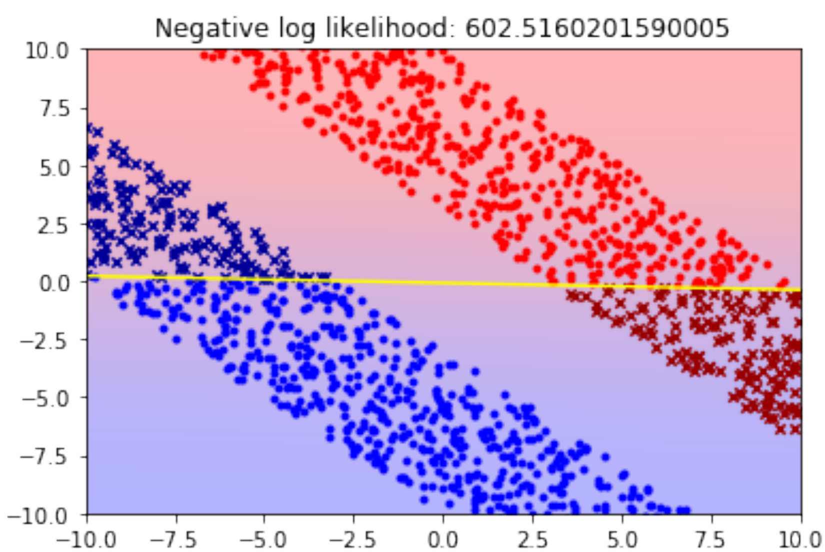


Negative log likelihood: 82.12360596412583

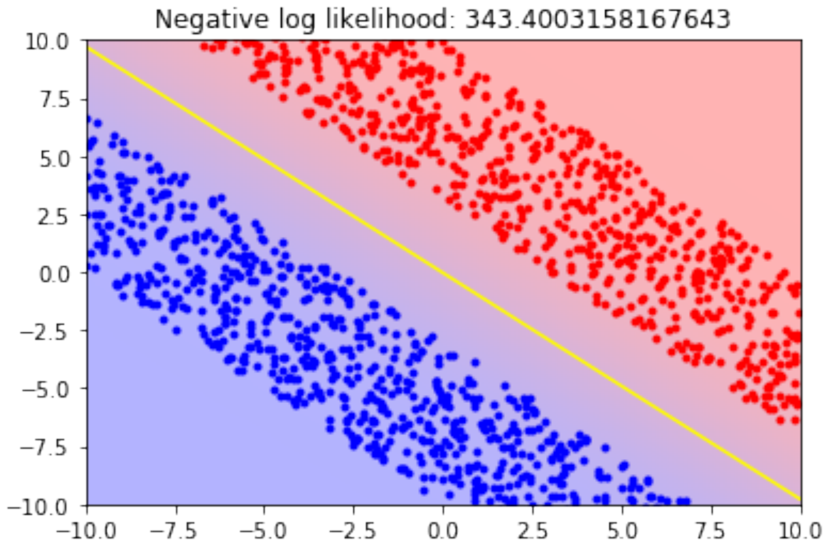


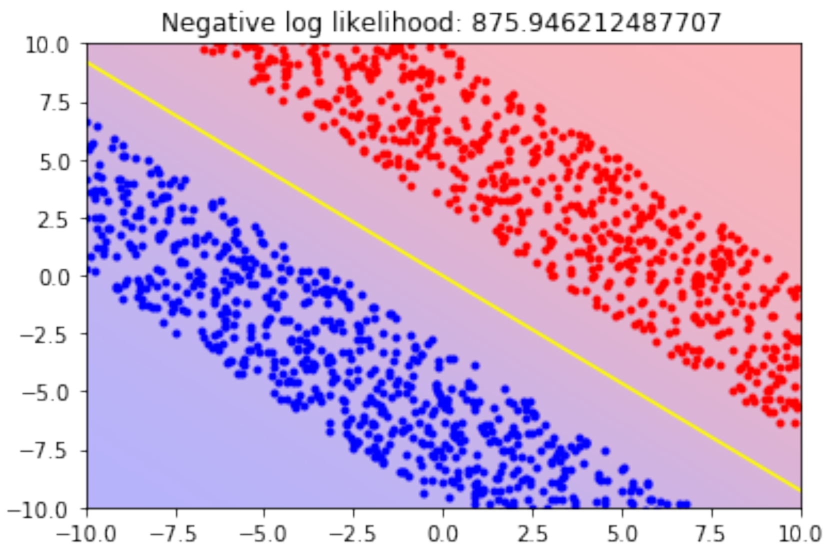


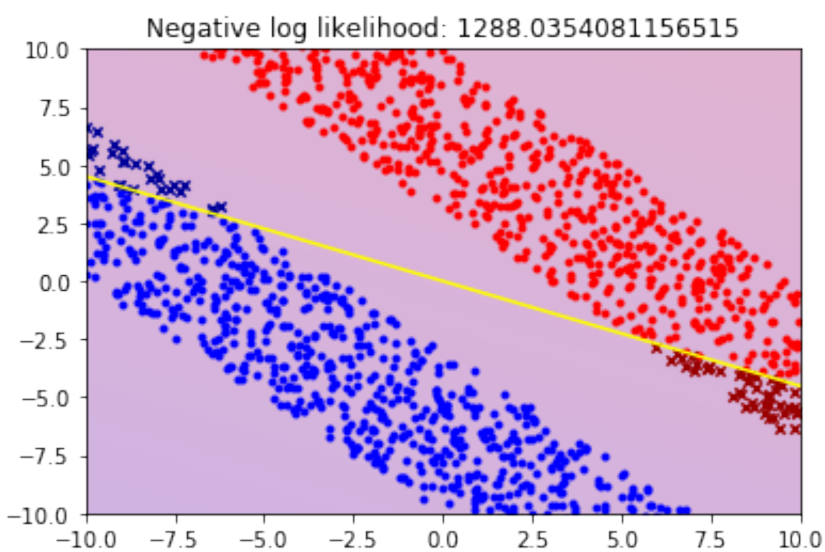


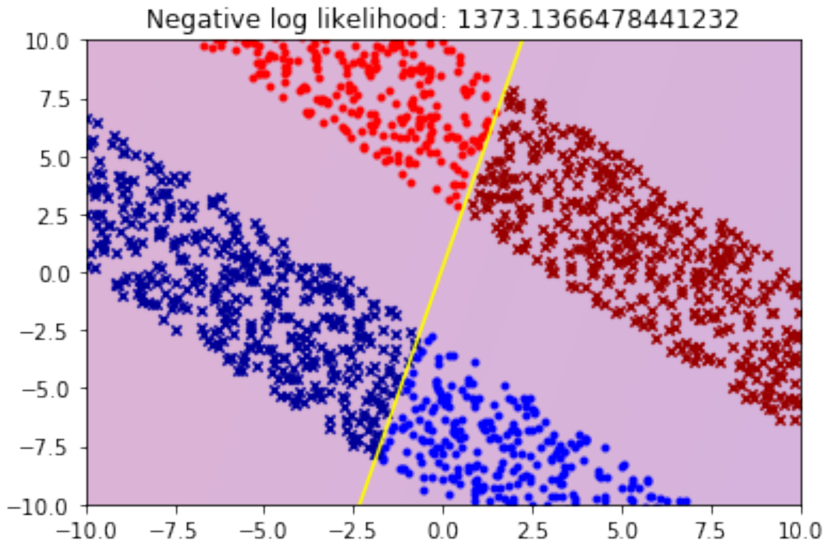


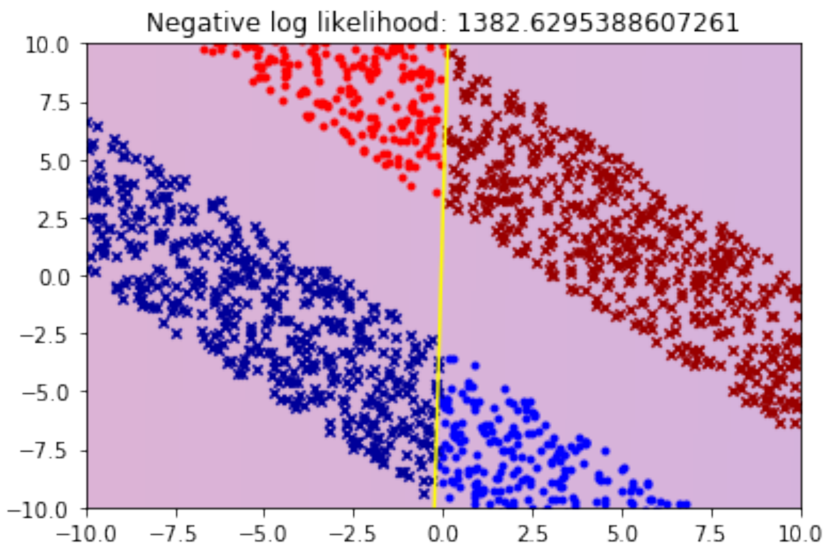
3. Regularize x2 (the order of pictures is the lambda from 1000 to 100000000)

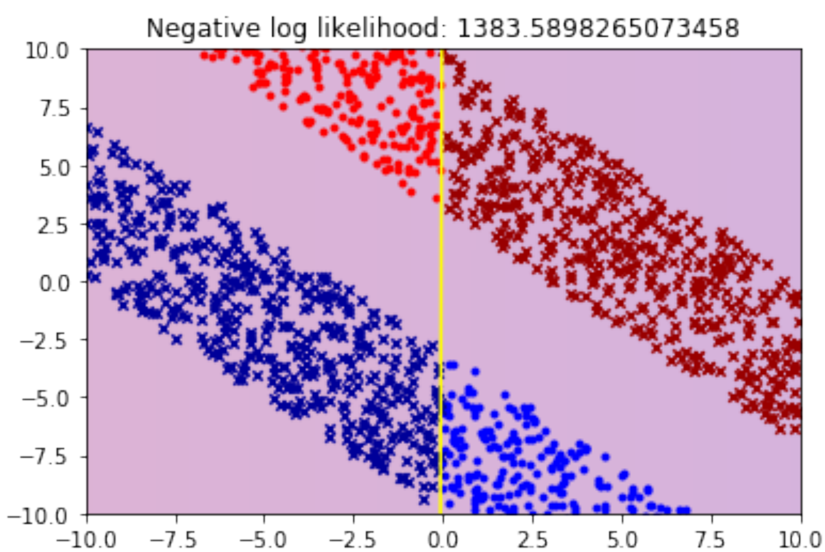

Negative log likelihood: 343.4003158167643
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0


Negative log likelihood: 875.946212487707
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0


Negative log likelihood: 1288.0354081156515
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0


Negative log likelihood: 1373.1366478441232


Negative log likelihood: 1382.6295388607261


Negative log likelihood: 1383.5898265073458

4