Exercise 1:

1. The methods not working in this case is Logistic regression, the error message is:

```
/Users/houningzhi/Library/Python/3.8/lib/python/site-packages/statsmodels/base/model.py:607: Converg
enceWarning: Maximum Likelihood optimization failed to converge. Check mle_retvals
  warnings.warn("Maximum Likelihood optimization failed to "
```

The loss function for Logit is $L_W(X_i, Y_i) = -Y_i \log P_i - (1-Y_i)\log(1-P_i)$, where $P_i = \frac{1}{1+\exp(-\langle X_i, W\rangle)}$

when data is linearly separable $\langle X_i, W\rangle > 0$ when $Y_i = 1$, $\langle X_i, W\rangle < 0$ when $Y_i = 0$

We can compute: $L_{2W}(X_i, Y_i)$, If $Y_i = 0$, $L_W(X_i, Y_i) = -\log(1-P_i)$, $\frac{1}{1+\exp(-\langle X_i, 2W\rangle)} < \frac{1}{1+\exp(\langle X_i, W\rangle)}$

$L_{2W}(X_i, Y_i) > L_W(X_i, Y_i)$, so which means loss gets smaller if $C \cdot W$, C increases

So the loss never converges to a minimum value, that's exactly what error said.

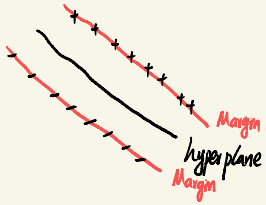
Compare Soft Margin SVC and Logistic regression

(1) Similarity: Both use for classification task, Both Logistic Regression and Soft Margin SVM)

(2) Difference: The loss function is different: Logit models the probability that a data point belongs to a class, SVC focus on maximizing margin instead.

2. By python code, there are 2000 values ≤ 1, all $y\hat{y} = 1$, so:

For data set A, the sketch is like:



$W = \sum \alpha_i Y_i X_i$ where: $Y_i \hat{Y}_i = 1$: (By complementary slack ness, $\alpha_i = 0$ or $Y_i \hat{Y}_i = 1$)

By the python result: There are 2 support vectors in the SVM, so we require 2 points

For data set B, the failed model is Hard margin SVC, Because the data set is not linearly separable. I need 192 support vectors

Exercises:

1. The primal form is $\min_{w\in\mathbb{R}^d,\, b\in\mathbb{R}} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \max\{|y_i-(w^Tx+b)|-\varepsilon, 0\}$

$$= \min_{w\in\mathbb{R}^d,\, b\in\mathbb{R}} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \max\{\gamma_i-(w^Tx+b)-\varepsilon, (w^Tx+b)-\gamma_i-\varepsilon, 0\}$$

$$= \min_{w\in\mathbb{R}^d,\, b\in\mathbb{R}} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \gamma_i \;,\; s.t \quad y_i-(w^Tx+b)-\varepsilon \le \gamma_i,\ \gamma_i \ge 0,\ (w^Tx+b)-\gamma_i-\varepsilon \le \gamma_i$$

Take Lagrangian dual to remove constraint, and by strong duality.

$$= \max_{\alpha,\beta,d\in\mathbb{R}^n,\ \alpha,\beta,d\ge 0} \min_{w\in\mathbb{R}^d,\, b\in\mathbb{R}} \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} C\gamma_i + \alpha_i(y_i-(w^Tx+b)-\varepsilon-\gamma_i) - \beta_i\cdot\gamma_i + d_i(w^Tx+b-\gamma_i-\varepsilon-\gamma_i)$$

set gradient to 0: $\frac{\partial}{\partial b} = \sum_i d_i-\alpha_i = 0$ $\frac{\partial}{\partial w} = \underbrace{w^*}_{\sum_i} - \alpha_i\cdot x + d_i\cdot x = 0$ $\frac{\partial}{\partial \gamma_i} = C-\alpha_i-\beta_i-d_i = 0 \Rightarrow w = \sum_i (\alpha_i-d_i)\cdot x_i$

$$= \max_{\alpha,\beta,d\in\mathbb{R}^n,\ \alpha,\beta,d\ge 0} \min_{w\in\mathbb{R}^d,\, b\in\mathbb{R}} \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n}(d_i-\alpha_i)(w^Tx+b-y_i) + \underbrace{\sum_{i=1}^{n}(C-\alpha_i-\beta_i-d_i)\cdot\gamma_i}_{0} - \sum_{i=1}^{n}(\alpha_i+d_i)\cdot\varepsilon$$

$$= \max_{\alpha,\beta,d\in\mathbb{R}^n,\ \alpha,\beta,d\ge 0} \min_{w\in\mathbb{R}^d,\, b\in\mathbb{R}} \frac{1}{2}\|w\|^2 + \underbrace{\sum_{i=1}^{n}(d_i-a_i)b}_{0} + \underbrace{w^T\sum_{i=1}^{n}(d_i-a_i)\cdot x}_{-\|w\|^2} - \sum_{i=1}^{n}(\alpha_i+d_i)\cdot\varepsilon - \sum_{i=1}^{n}(d_i-a_i)\cdot y_i$$

$$= \max_{\substack{\alpha,d\in\mathbb{R}^n \\ C\ge\alpha,d\ge 0}} \min_{w\in\mathbb{R}^d,\, b\in\mathbb{R}} -\frac{1}{2}\|w\|^2 - \sum_{i=1}^{n}(\alpha_i+d_i)\cdot\varepsilon - \sum_{i=1}^{n}(d_i-\alpha_i)\cdot y_i$$

2) The gradient is:

If $\varepsilon+(w^Tx+b) < y_i$, then $\frac{\partial}{\partial w}\left(C\sum_{i=1}^{n}\max\{y_i-(w^Tx_i+b)-\varepsilon, 0\}\right) = -Cx_i$

$$\frac{\partial}{\partial b}\left(C\sum_{i=1}^{n}\max\{y_i-(w^Tx_i+b)-\varepsilon, 0\}\right) = -C$$

If $-\varepsilon \le y_i-(w^Tx+b) \le \varepsilon$, then $\frac{\partial}{\partial w}\left(C\sum_{i=1}^{n}\max\{y_i-(w^Tx_i+b)-\varepsilon, 0\}\right) = 0$

$$\frac{\partial}{\partial b}\left(C\sum_{i=1}^{n}\max\{y_i-(w^Tx_i+b)-\varepsilon, 0\}\right) = 0$$

$w^Tx+b-\varepsilon \ge y_i$, then $\frac{\partial}{\partial w}\left(C\sum_{i=1}^{n}\max\{(w^Tx_i+b)-y_i-\varepsilon, 0\}\right) = Cx_i$

$$\frac{\partial}{\partial b}\left(C\sum_{i=1}^{n}\max\{(w^Tx_i+b)-y_i-\varepsilon, 0\}\right) = C$$

(3) $p^{\eta}(w) = \arg\min_{z} \frac{1}{2\eta}\|z-w\|_2^2 + \frac{1}{2}\|z\|_2^2$

$= \arg\min_{z} \frac{1}{2\eta}\cdot\|z\|_2^2 - \frac{1}{\eta} z\cdot w + \frac{1}{2\eta}\|w\|_2^2 + \frac{1}{2}\|z\|_2^2$.

So that $\frac{\partial p^{\eta}(w)}{\partial z} = z + \frac{1}{\eta}z - \frac{1}{\eta}w$, set gradient to 0, then

$z = \frac{1}{1+\eta}\cdot w$.

(4) The output of the traing output by selecting appropriate step size

```
the Training loss is 610.554899840419
the Training error is 610.0421934063855
the Test error is 776.4545373565409
0     -0.021302
1     -0.088689
2     -0.005106
3      0.101048
4      0.035666
         ...
195    0.009747
196   -0.007770
197   -0.079203
198    0.028541
199   -0.007645
Length: 200, dtype: float64 0.006
kous-MacBook-Air:A2 houningzhi$
```

Exercise 3:

(1) $k(x,y) = e^{-\alpha(x-y)^2}$, For Taylor expansion, $e^x = \frac{x^0}{0!} + \frac{x^1}{1!} + \frac{x^2}{2!} + \cdots\cdots$

$= e^{-\alpha(x^2-2xy+y^2)}$

$= e^{-\alpha(x^2+y^2)} \cdot e^{2\alpha xy}$ (Taylor expansion on $e^{2\alpha xy}$)

$= e^{-\alpha(x^2+y^2)} \cdot \left[1 + \frac{2\alpha xy}{1!} + \frac{(2\alpha xy)^2}{2!} + \cdots\cdots\right]$

$= e^{-\alpha x^2} \cdot e^{-\alpha y^2} \left[1, \sqrt{\frac{2\alpha}{1!}}x, \sqrt{\frac{(2\alpha)^2}{2!}}x^2, \cdots\cdots\right]^T \left[1, \sqrt{\frac{2\alpha}{1!}}y, \sqrt{\frac{(2\alpha)^2}{2!}}y^2, \cdots\right]$

where clearly we can observe that:

$\phi(x) = e^{-\alpha x^2}\left[1, \sqrt{\frac{2\alpha}{1!}}x, \sqrt{\frac{(2\alpha)^2}{2!}}x^2, \cdots\cdots\right]$, thus that $k(x,y) = \phi(x)^T\phi(y)$

I'd love to solve the dual representation, because $\phi(x)$ maps x to an infinite dimension vector. It's impossible to solve primal form, but in dual form, we only need $k(x,y)$.

(2) It's a kernel, we know that:

$k(x,y) = \frac{1}{1-xy}$,

$= 1 + xy + x^2y^2 + \cdots\cdots$, by Taylor expansion

$= [1, x, x^2, \cdots]^T[1, y, y^2, \cdots]$

$= \phi(x)^T \cdot \phi(y)$, So:

The feature map $\phi(x) = [1, x, x^2, \cdots]^T$.

(3). It's not a kernel, we can give a counter example:

$x = [0.1 \ 10 \ 100]$ So the matrix k looks like:

$$k = \begin{bmatrix} \log(1.01) & \log(2) & \log(11) \\ \log(2) & \log(101) & \log(1001) \\ \log(11) & \log(1001) & \log(10001) \end{bmatrix}$$, If it's a kernel, k must be PSD. given $v = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$

$v^Tk = [\log(1.01)+\log(2)-\log(11), \log(2)+\log(101)+\log(1001), \log(11)+\log(1001)+\log(10001)]$

$v^Tkv = \log(1.01)+2\log(2)+\log(101)+\log(10001)-2\log(11)-2\log(1001)$

$= -1.47291\cdots < 0$

So k is not PSD, so k is not a kernel

(4) It's not a kernel, we can give a counter example:

$x = [\frac{\pi}{2}, \frac{3\pi}{2}]$, So that $k = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$, let $v = [1, -1]$, So we compute v^Tkv:

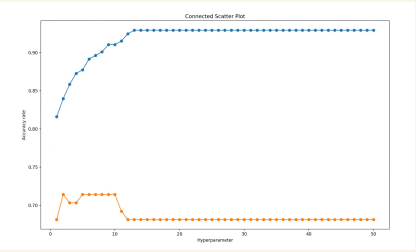
$v^Tk = [-2, 2]$ $v^Tkv = -4 < 0$

So that k is not PSD, contradiction; So $k(x,y)$ is not a kernel

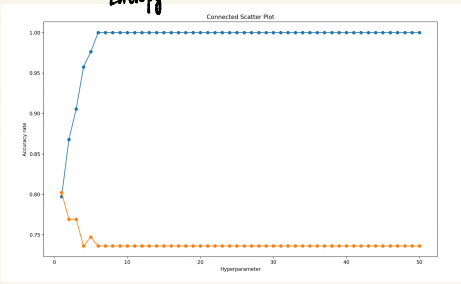
Exercise 4: Hyperparameter: depth ranges from 1 to 50

(a)

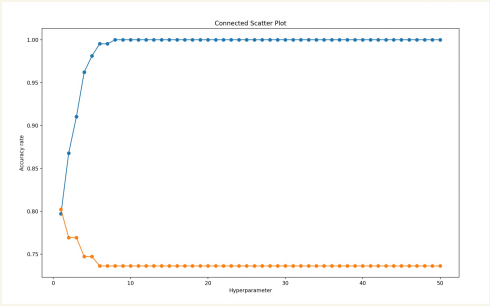
Misclassification error



Entropy



Gini-Index



Analysis: These three plots shows how accuracy changes for test data when maximum depth for tree increases.

Generally: The traing accuracy can reach 1 finally, which makes sense, If we have many leaves, we can classify 1 data on one leave to be 100% accurate on Training data. which is also overfitting.

In contrast, the test accuracy decreases gradually due to the overfitting

(b) The result for bagging is better than the result with out (72%)

```
101
101
the median is 0.8131868131868132, the maximum is 0.8351648351648352, the minimum is 0.7912087912087912
kous-MacBook-Air:A2 houningzhi$
```

The result for random forest and bagging is:

```
101
101
the median is 0.8241758241758241, the maximum is 0.8461538461538461, the minimum is 0.8021978021978022
kous-MacBook-Air:A2 houningzhi$
```

It's better than any result before!