CS 567 HW 1 Congyue Wang 4566223937

1.a)

$$\mathrm{f(x)} = \frac{x^{\alpha-1}(1-x)^{\beta}}{B(\alpha,\beta)}$$

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,dt$$

$$\beta = 1 \quad \mathrm{B}(\alpha,\beta) = \frac{1}{a} \quad f(x) = ax^{a-1}$$

$$\mathrm{L} = \prod_{i=1}^{n} f(X_i|\alpha)$$

$$\ln(\mathrm{L}) = \mathrm{n}\ln(\alpha) + (\alpha-1)\sum_{i=1}^{n}\ln(Xi)$$

Maximum likelihood of a

$$\frac{\partial(\ln(\mathrm{L}))}{\partial\alpha} = \frac{n}{a} + \sum_{i=1}^{n}\ln(X_i) = 0$$

$$\alpha = -\frac{n}{\sum_{i=1}^{n}\ln(X_i)}$$

Univariate Gaussian: $X \sim \mathcal{N}(\mu,\sigma^2)$ if

$$f(x) = \mathcal{N}(\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

for $\mu \in R$, $\sigma > 0$.

Multivariate Gaussian: $\mathbf{X} \sim \mathcal{N}(\mu,\Sigma)$ if

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right)$$

for $\mu \in R^D$, Σ is a $D \times D$ symmetric positive definite matrix. Notice that $E[\mathbf{X}] = \mu$ and $Var[\mathbf{X}] = \Sigma$.

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)$$

In this case

$$L = \frac{1}{2\pi^{d/2}\prod_{i=1}^{n}\theta_i^{1/2}}\exp(-\frac{1}{2}\sum_{i=1}^{n}\frac{1}{\theta_i}(x_i-\theta_i)^2)$$

$$\ln(L) = \text{const} - \frac{1}{2}\sum_{i=1}^{n}\ln(\theta_i) - \frac{1}{2}\sum_{i=1}^{n}\frac{1}{\theta_i}(x_i-\theta_i)^2$$

$$\frac{\partial \ln(L)}{\partial\theta_i} = -\frac{1}{2}\frac{1}{\theta_i} - \frac{1}{2}\left(-\frac{1}{\theta_i^{2}}(x_i-\theta_i)^2 - \frac{2}{\theta_i}(x_i-\theta_i)\right) = 0$$

$$\theta_i = \sqrt{x_i^2 + \frac{1}{4}} - \frac{1}{2}$$

b)

- Likelihood of one training sample (x_n, y_n)

$$p(y_n|x_n) = N(w_0 + w_1 x_n, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{[y_n-(w_0+w_1x_n)]^2}{2\sigma^2}}$$

$$\log P(\mathcal{D}) = \log\prod_{n=1}^{N} p(y_n|x_n) = \sum_n \log p(y_n|x_n)$$

- Maximize over w_0 and w_1

$$\max\ \log P(\mathcal{D}) \Leftrightarrow \min\ \sum_n [y_n - (w_0 + w_1 x_n)]^2$$

$$\frac{\partial L}{\partial\omega_0} = -2\sum_{i=1}^{n}(Y - \omega_0 - \omega^T x) = 0$$

$$\omega_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - w^T\sum_{i=1}^{n} x_i = \bar{y} - \omega^T\bar{x}$$

- Maximize over w_0 and w_1

$$\max\ \log P(\mathcal{D}) \Leftrightarrow \min\ \sum_n [y_n - (w_0 + w_1 x_n)]^2$$

$\sum_{i=1}^{n}(y_i - \bar{y} + \omega^T \bar{x} - \omega^T x_i)^2 = \sum_{i=1}^{n}(y_i - \bar{y} - \omega^T(x_i - \bar{x}))^2 =$

$$\sum_{i=1}^{n}(y_i - \bar{y} - (x_i - \bar{x})^T \omega)^2$$

$$\frac{\partial \mathrm{L}}{\partial \omega} = 2(x_i - \bar{x}) \quad \sum_{i=1}^{n}(y_i - \bar{y} - (x_i - \bar{x})^T \omega) \quad = 0$$

$$\omega = [\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T]^{-1}[\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) \quad]$$

c)

$$\mathrm{E}\left[\widehat{f(x)}\right] = \frac{1}{nh}\sum_{i=1}^{n} E[K(\frac{x - x_i}{h})] = \frac{1}{h}\int K(\frac{x-t}{h}) f(t) dt$$

when h->0

$$\mathrm{f(x - hz) = f(x) - f'(x)hz} + \frac{1}{2} f''(x) h^2 z^2$$

$$\int K(z) f(x - hz) dz$$

$$= \int \begin{matrix} K(z)\left(\mathrm{f(x) - f'(x)hz} + \frac{1}{2} f''(x) h^2 z^2\right) dz \\ = f(x)\int K(z) dz + f'(x) h \int K(z)\, z dz + \\ \frac{1}{2} f(\mathrm{x})\mathrm{h}^2 \int \mathrm{K(z)\, z^2 dz} = \mathrm{f(x)} + \frac{1}{2}\mathrm{f}(x) h^2 \int K(z)\, z^2 dz \end{matrix}$$

$$bias = \frac{1}{2}\mathrm{f}(x) h^2 \int K(z)\, z^2 dz$$

2 a)

$$\bar{x} = 14.1 \; \bar{y} = 21.1$$

$$\sigma_x = 20.1 \; \sigma_y = 27.3$$

After normalization

X=[0.044 -1.0485 -0.899 0.740 0.8895 1.1379 0.1938 1.2870 -1.098 -1.2472]

Y=[1.0231 0.6197 0.9498 0.1063 0.5464 0.8031 -0.4437 -1.8005 -1.4705 -0.3337]

Math {(0.044,1.0231)(-1.0485, 0.6197)(-0.899, 0.9498)}

EE {(0.740,0.1063)(0.8895, 0.5464)(1.1379, 0.8031)}

CS{(0.1938, -0.4437)(1.2870, -1.8005)(-1.098, -1.4705) (-1.2472, -0.3337)}

For (9,18) after normalization become (-0.2534,-0.1137)

L1 distance =min($|\mathrm{x} - x_n| + |\mathrm{y} - y_n|$)

The list of L1 distance is

1.4349	1.5285	1.7094	1.2138	1.8030
2.3081	0.7773	3.2273	2.2015	1.2138

L1 distance and K=1 the nearest neighbor is CS

K=3 the nearest neighbor is CS

L2 distance =min($\sqrt{(x - x_n)^2 + (y - y_n)^2}$)

The list of L2 distance is

1.1752	1.0817	1.2443	1.0179	1.3198
1.6662	0.5558	2.2844	1.5983	1.0179

L2 distance and K=1 the nearest neighbor is CS

K=3 the nearest neighbor is CS

So they are the same.

b)

given the fact that $\sum_c K_c = K \ \sum_c N_c = N$

P(x)=$\frac{K}{NV}$

Using Bayes rule

$$\mathrm{P(Y = c|x)} = \frac{P(x|Y = c)P(Y = c)}{P(x)} = \frac{K_c}{K}$$

3 a)

As an independent noise is added to each Xn we have

$$\mathrm{y}(x_n, \omega) = \omega_0 + \sum_{d=1}^{D} \omega_d (x_{nd} + \epsilon_{nd})$$

$$= \omega_0 + \sum_{d=1}^{D} \omega_d x_{nd} + \sum_{d=1}^{D} \omega_d \epsilon_{nd}$$

ϵ_{nd} independent across both n and d

$$\mathrm{E} = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\omega) - t_n\}^2$$

$$= \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\omega) + \sum_{d=1}^{D}\omega_d\epsilon_{nd} - t_n\}^2$$

$$= \frac{1}{2}\sum_{n=1}^{N}\{(y(x_n,\omega) - t_n)^2 + 2(y(x_n,\omega) - t_n)\left(\sum_{d=1}^{D}\omega_d\,\epsilon_{nd}\right) + (\sum_{d=1}^{D}\omega_d\epsilon_{nd})^2\}$$

E[$(\sum_{d=1}^{D}\omega_d\epsilon_{nd})^2]$ = $\mathrm{E}\left[\left(\sum_{d=1}^{D}\sum_{d=1}^{D}\omega_d\omega_{d'}\,\epsilon_{nd}\epsilon_{nd'}\right)\right] =$

$\left(\sum_{d=1}^{D}\sum_{d=1}^{D}\omega_d\omega_{d'}E[\,\epsilon_{nd}\epsilon_{nd'}]\right)=\sum_{d=1}^{D}\omega_d^2$

$$\mathrm{E}=\frac{1}{2}\sum_{n=1}^{N}\left\{(y(x_n,\omega)-t_n)^2+\sum_{d=1}^{D}\omega_d^2\right\}=E(\omega)_D+\frac{N}{2}\sum_{d=1}^{D}\omega_d^2$$

$$\mathrm{E}[\widetilde{x_n}]=\begin{bmatrix}\mathrm{E}[\widetilde{x_{n,1}}]\\ \vdots\\ \mathrm{E}[\widetilde{x_{n,d}}]\end{bmatrix}=\begin{bmatrix}x_{n,1}(1-\delta)\\ \vdots\\ x_{n,d}(1-\delta)\end{bmatrix}=x_n$$

$$\mathrm{var}[w^T\widetilde{x_n}]=\sum_{d=1}^{D}\omega_d^2\,var[\widetilde{x_n}]=\sum_{d=1}^{D}\omega_d^2\frac{x_{n,d}^2}{(1-p)^2}p(1-p)$$

$$=\frac{p}{1-p}\sum_{d=1}^{D}\omega_d^2\,x_{n,d}^2$$

4 a)

By using the formula of conditional entropy we get

Outlook 0.48 temp 0.71 humidity 0.54 wind 0.74

So outlook should be the first split

b)

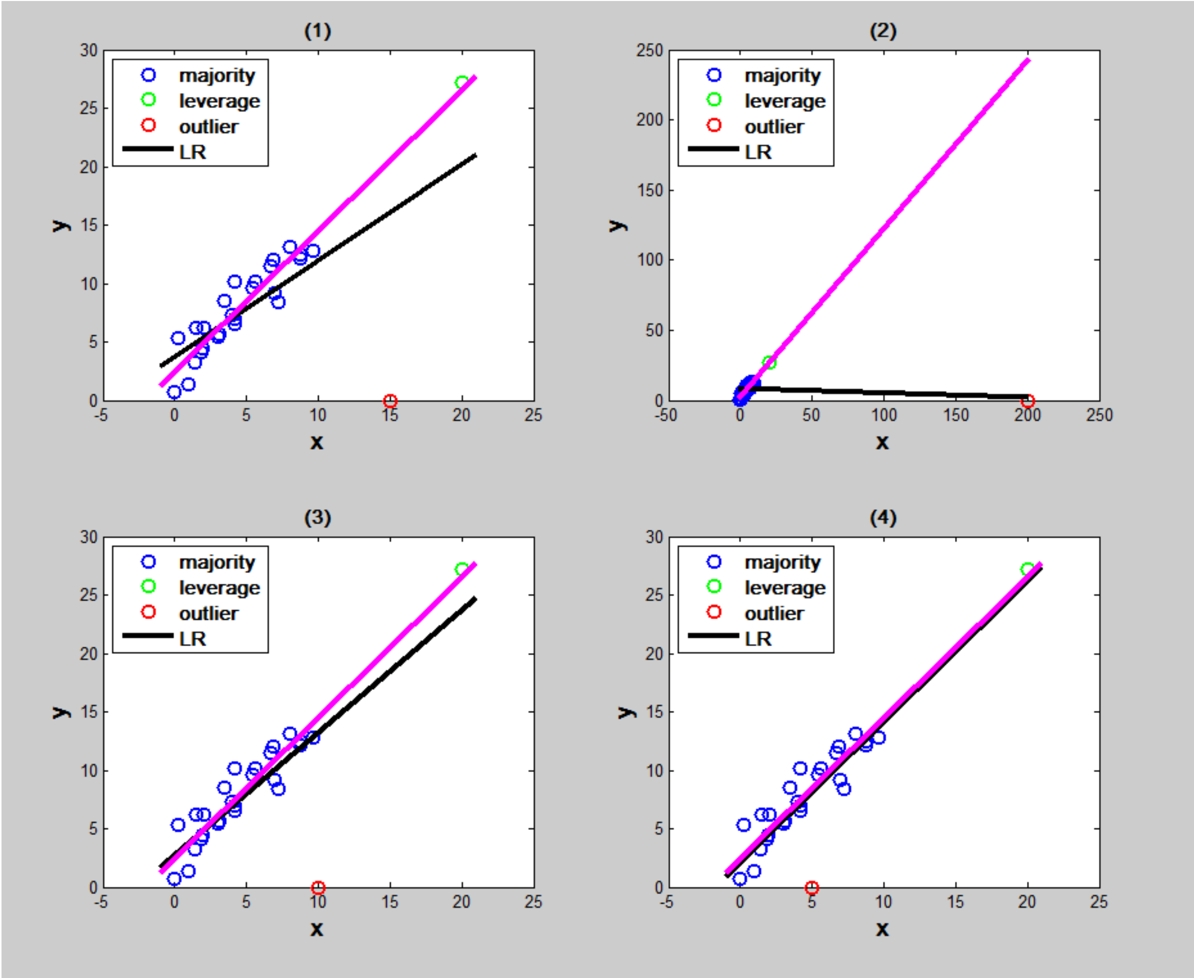
$$\mathrm{f}(p_k)=1-p_k+\log(p_k)$$

$$\mathrm{f}'=-1+\frac{1}{p_k ln2}\quad 0\le p_k\le 1\ \ ln2=0.693\ f'\ge 0$$

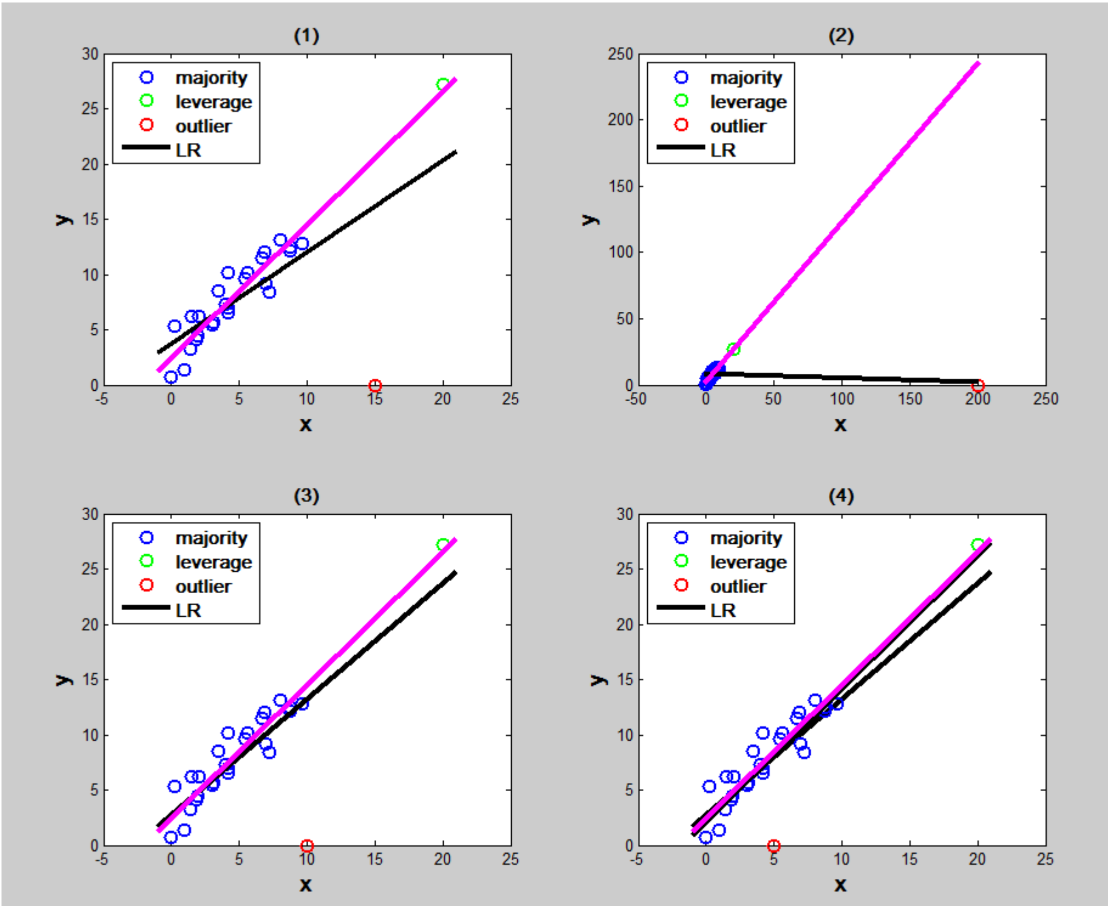
f increase for $0\le p_k\le 1$ if $p_k=1\ f(1)=0$ so if $p_k\le 1$

$$\mathrm{f}\le 0$$

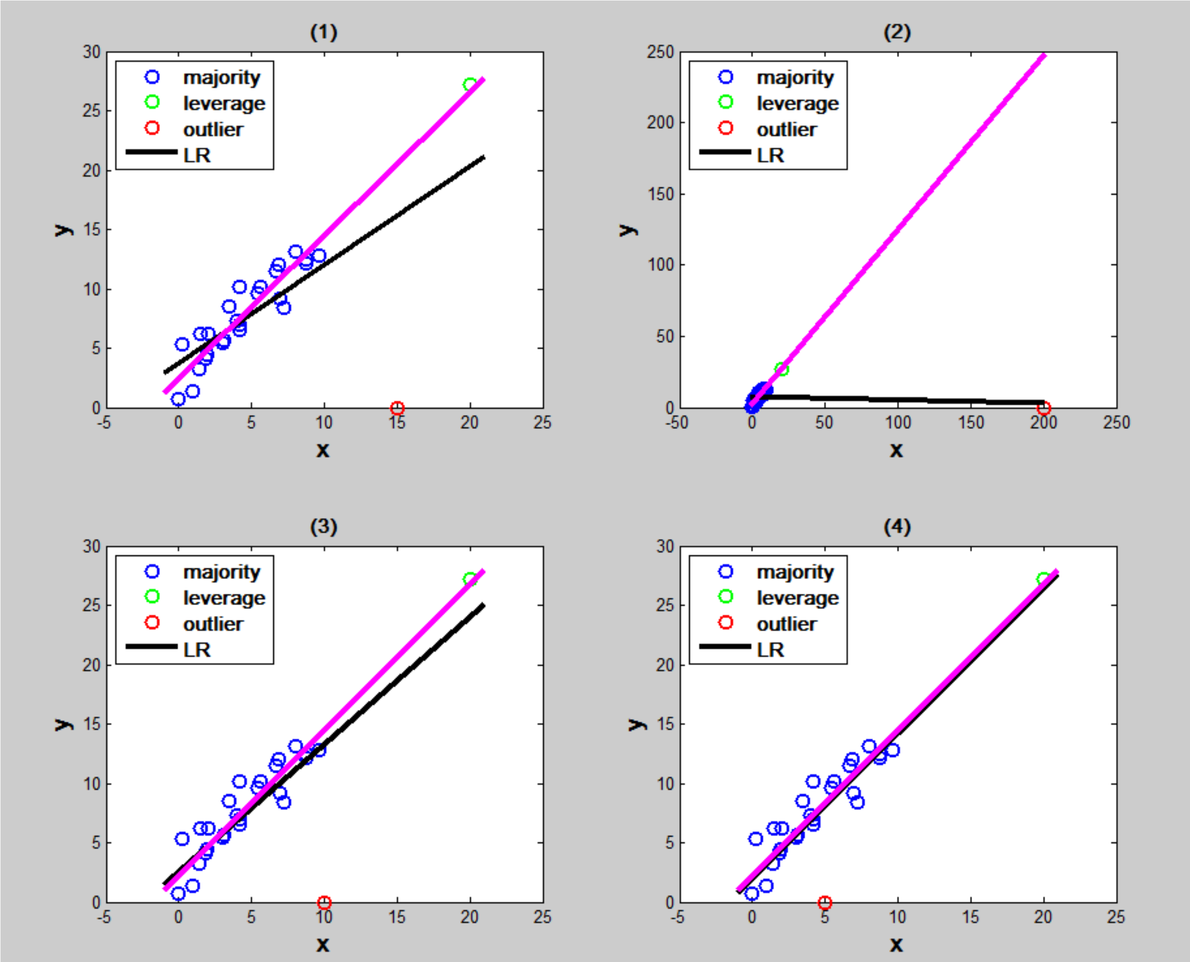
5 a) LR without lampda



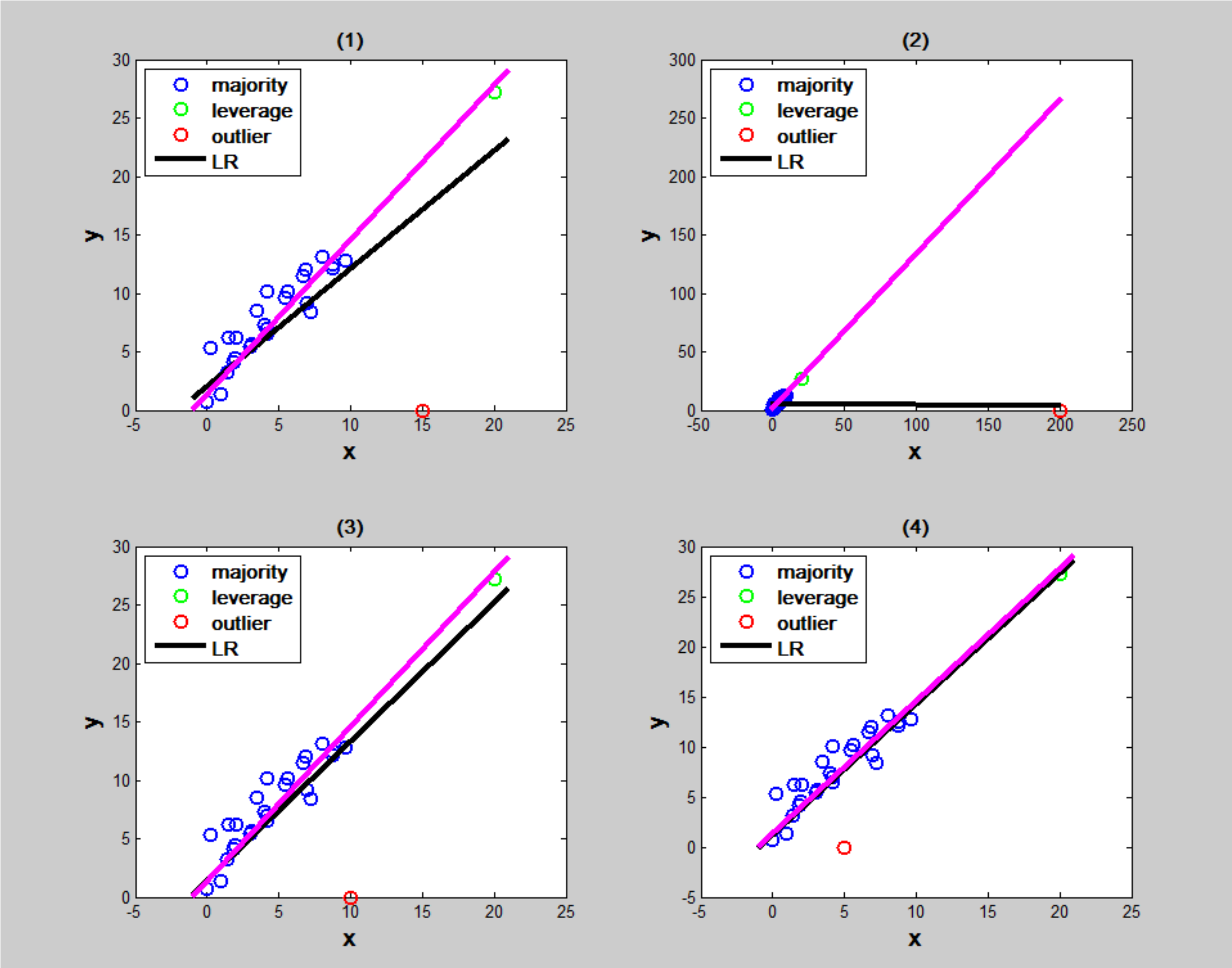
LR with lampda=0.1



Lampda=1

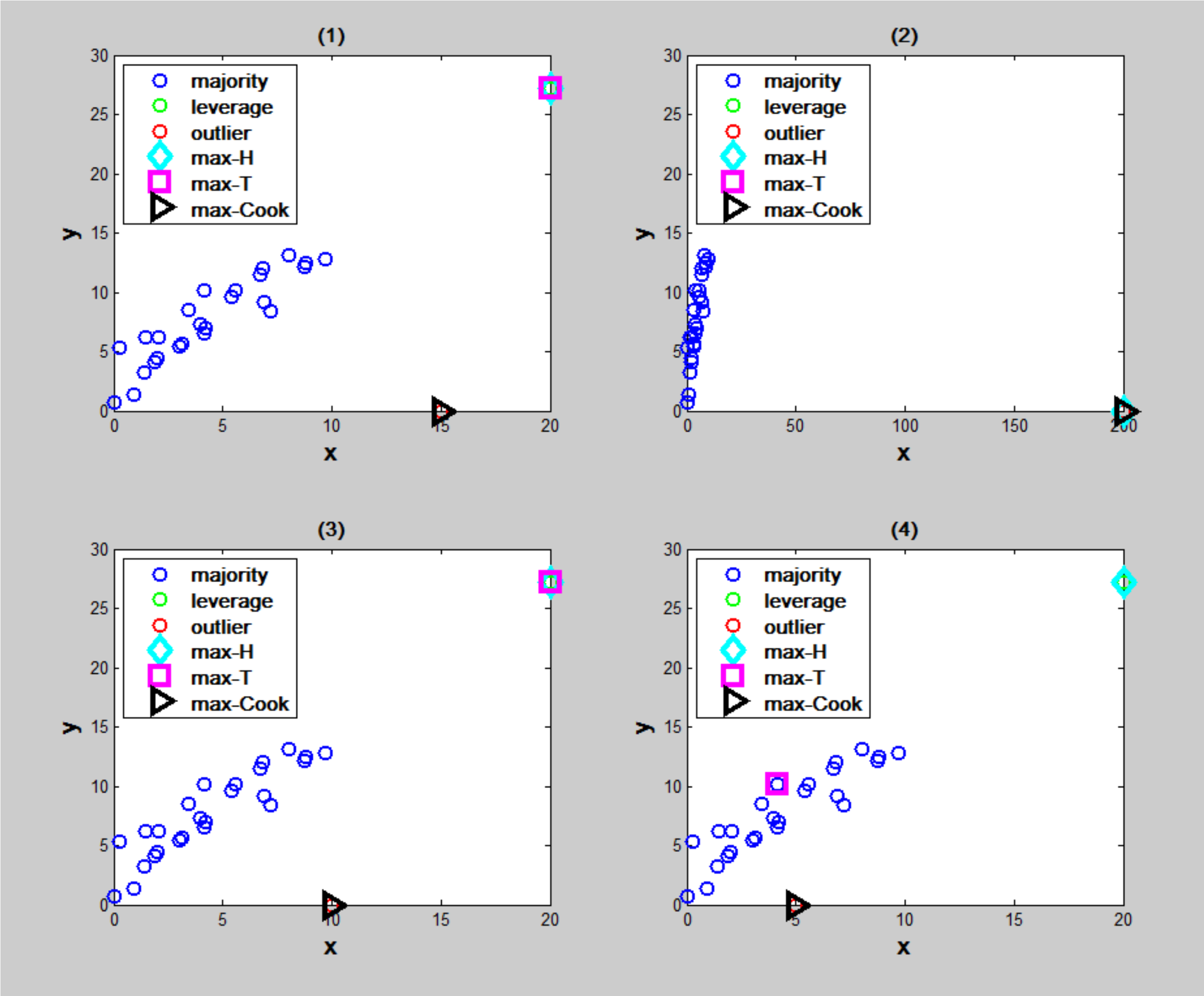


Lampda=10



From the figure we can see that there is no big difference.

1) Cook's distance, 2) leverage and 3) studentized residual.



It seems that it can predict outlier sample.