1. Newton's method for computing least squares

(a)因为

$$
\begin{aligned}
\frac{\partial J(\theta)}{\partial \theta_j} &= \sum_{i=1}^{m}(\theta^T x^{(i)} - y^{(i)})x_j^{(i)} \\
\nabla J(\theta) &= \sum_{i=1}^{m}(\theta^T x^{(i)} - y^{(i)})x^{(i)}
\end{aligned}
$$

所以

$$
\frac{\partial^2 J(\theta)}{\partial \theta_k \partial \theta_j} = \frac{\partial}{\partial \theta_k}\sum_{i=1}^{m}(\theta^T x^{(i)} - y^{(i)})x_j^{(i)} = \sum_{i=1}^{m} x_k^{(i)} x_j^{(i)}
$$

注意到

$$
X = \begin{bmatrix} (x^{(1)})^T \\ (x^{(2)})^T \\ \dots \\ (x^{(m)})^T \end{bmatrix}, \vec{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \dots \\ y^{(m)} \end{bmatrix}
$$

所以

$$
\nabla^2 J(\theta) = X^T X
$$

(b)牛顿法的规则为

$$
\theta := \theta - (\nabla^2 J(\theta))^{-1} \nabla J(\theta)
$$

θ的初始值为0，所以此时

$$
\nabla J(\theta) = \sum_{i=1}^{m}(\theta^T x^{(i)} - y^{(i)})x^{(i)} = -\sum_{i=1}^{m} y^{(i)} x^{(i)} = -X^T \vec{y}
$$

所以第一步更新后

$$
\theta = (X^T X)^{-1} X^T \vec{y}
$$

2.Locally-weighted logistic regression

(a)首先推导题目中给出的梯度计算式，注意到

那么

$$
\begin{aligned}
h_\theta(x^{(i)}) &= \sigma(\theta^T x^{(i)}) \\
\sigma(x) &= \frac{1}{1+e^{-x}} \\
\sigma'(x) &= \sigma(x)(1-\sigma(x))
\end{aligned}
$$

所以

$$
\begin{aligned}
\nabla_\theta h_\theta(x^{(i)}) &= h_\theta(x^{(i)})(1-h_\theta(x^{(i)}))x^{(i)} \\
\nabla_\theta \log(h_\theta(x^{(i)})) &= \frac{1}{h_\theta(x^{(i)})} \times \nabla_\theta h_\theta(x^{(i)}) = (1-h_\theta(x^{(i)}))x^{(i)} \\
\nabla_\theta \log(1-h_\theta(x^{(i)})) &= \frac{1}{1-h_\theta(x^{(i)})} \times (-1) \times \nabla_\theta h_\theta(x^{(i)}) = -h_\theta(x^{(i)})x^{(i)}
\end{aligned}
$$

从而

$$
\begin{aligned}
\nabla_\theta \ell(\theta) &= -\lambda\theta + \sum_{i=1}^m w^{(i)} \left[y^{(i)}(1-h_\theta(x^{(i)}))x^{(i)} - (1-y^{(i)})h_\theta(x^{(i)})x^{(i)} \right] \\
&= -\lambda\theta + \sum_{i=1}^m w^{(i)} \left[(y^{(i)} - h_\theta(x^{(i)}))x^{(i)} \right]
\end{aligned}
$$

定义$z \in \mathbb{R}^m$

$$
z_i = w^{(i)}(y^{(i)} - h_\theta(x^{(i)}))
$$

那么

$$
\nabla_\theta \ell(\theta) = X^T z - \lambda\theta
$$

接着计算Hessian矩阵，首先求偏导数

$$
\begin{aligned}
\frac{\partial^2 \ell(\theta)}{\partial\theta_k \partial\theta_j} &= \frac{\partial}{\partial\theta_k}\Big(-\lambda\theta_j + \sum_{i=1}^m w^{(i)} \left[(y^{(i)} - h_\theta(x^{(i)}))x_j^{(i)} \right] \Big) \\
&= -\lambda 1\{k=j\} + \sum_{i=1}^m w^{(i)} x_j^{(i)} (-h_\theta(x^{(i)})(1-h_\theta(x^{(i)}))x_k^{(i)}) \\
&= -\lambda 1\{k=j\} - \sum_{i=1}^m w^{(i)} h_\theta(x^{(i)})(1-h_\theta(x^{(i)}))x_j^{(i)} x_k^{(i)})
\end{aligned}
$$

记$D \in \mathbb{R}^{m\times m}$为对角阵，其中

$$
D_{ii} = -w^{(i)} h_\theta(x^{(i)})(1-h_\theta(x^{(i)}))
$$

那么

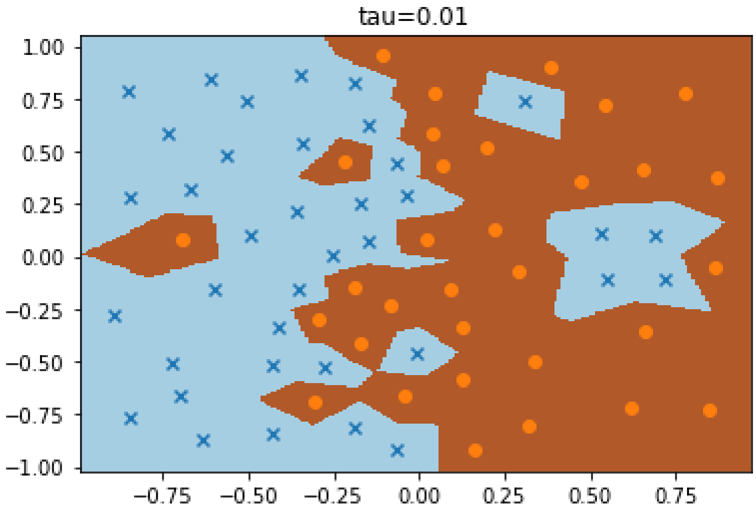
$$
H = X^T D X - \lambda I
$$

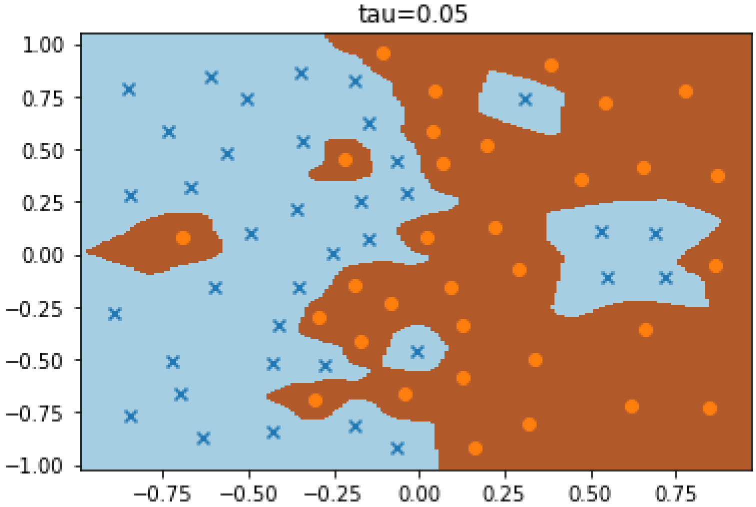
代码见(b)

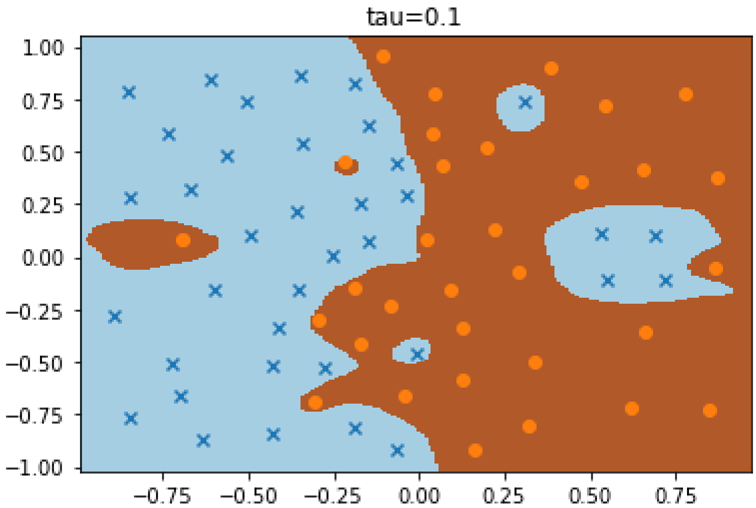
(b)

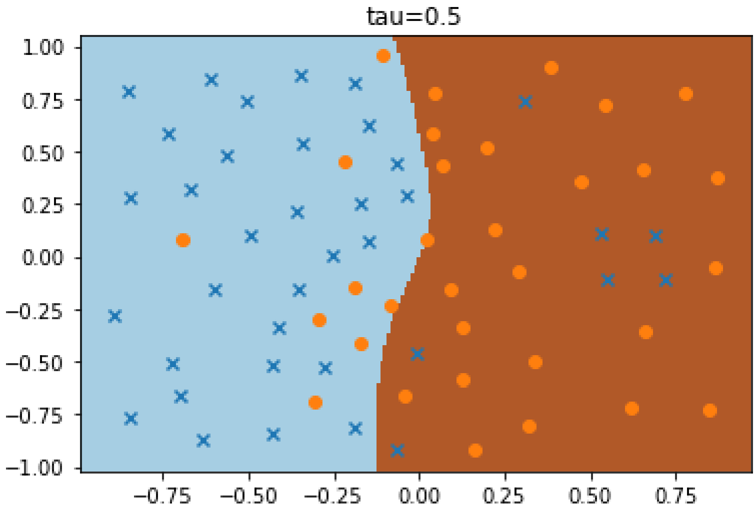
```python
# -*- coding: utf-8 -*-
"""
Created on Mon Jan 28 16:12:53 2019

@author: qinzhen
"""
import numpy as np
import matplotlib.pyplot as plt

Lambda = 0.0001
threshold = 1e-6

#读取数据
def load_data():
    X = np.loadtxt('data/x.dat')
    y = np.loadtxt('data/y.dat')

    return X, y

#定义h(theta, X)
def h(theta, X):
    return 1 / (1 + np.exp(- X.dot(theta)))

#计算
def lwlr(X_train, y_train, x, tau):
    #记录数据维度
    m, d = X_train.shape
    #初始化
    theta = np.zeros(d)
    #计算权重
    norm = np.sum((X_train - x) ** 2, axis=1)
    W = np.exp(- norm / (2 * tau ** 2))
    #初始化梯度
    g = np.ones(d)

    while np.linalg.norm(g) > threshold:
        #计算h(theta, X)
        h_X = h(theta, X_train)
        #梯度
        z = W * (y_train - h_X)
        g = X_train.T.dot(z) - Lambda * theta
        #Hessian矩阵
        D = - np.diag(W * h_X * (1 - h_X))
        H = X_train.T.dot(D).dot(X_train) - Lambda * np.eye(d)

        #更新
        theta -= np.linalg.inv(H).dot(g)

    ans = (theta.dot(x) > 0).astype(np.float64)
    return ans

#作图
def plot_lwlr(X, y, tau):
```

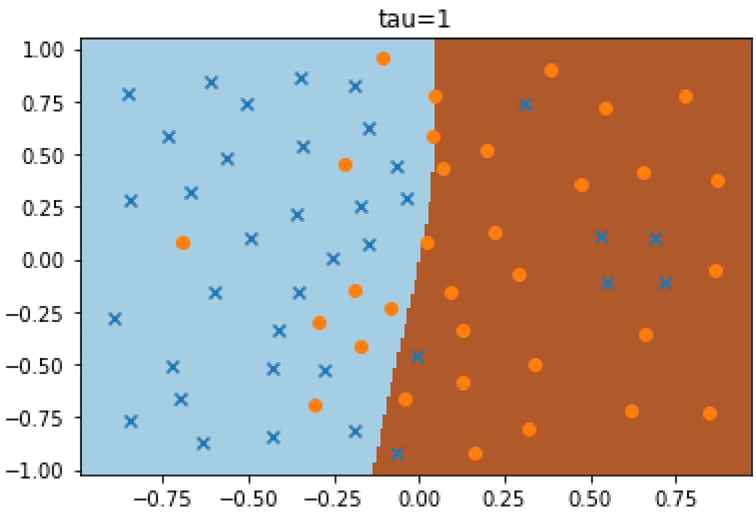
```
    x_min, x_max = X[:, 0].min() - .1, X[:, 0].max() + .1
    y_min, y_max = X[:, 1].min() - .1, X[:, 1].max() + .1
    xx, yy = np.meshgrid(np.arange(x_min, x_max, 0.01),
                         np.arange(y_min, y_max, 0.01))
    d = xx.ravel().shape[0]
    Z = np.zeros(d)
    data = np.c_[xx.ravel(), yy.ravel()]

    for i in range(d):
        x = data[i, :]
        Z[i] = lwlr(X, y, x, tau)

    plt.pcolormesh(xx, yy, Z, cmap=plt.cm.Paired)
    X0 = X[y == 0]
    X1 = X[y == 1]
    plt.scatter(X0[:, 0], X0[:, 1], marker='x')
    plt.scatter(X1[:, 0], X1[:, 1], marker='o')
    plt.title("tau="+str(tau))
    plt.show()

Tau = [0.01, 0.05, 0.1, 0.5, 1, 5]
X, y = load_data()
for tau in Tau:
    plot_lwlr(X, y, tau)
```

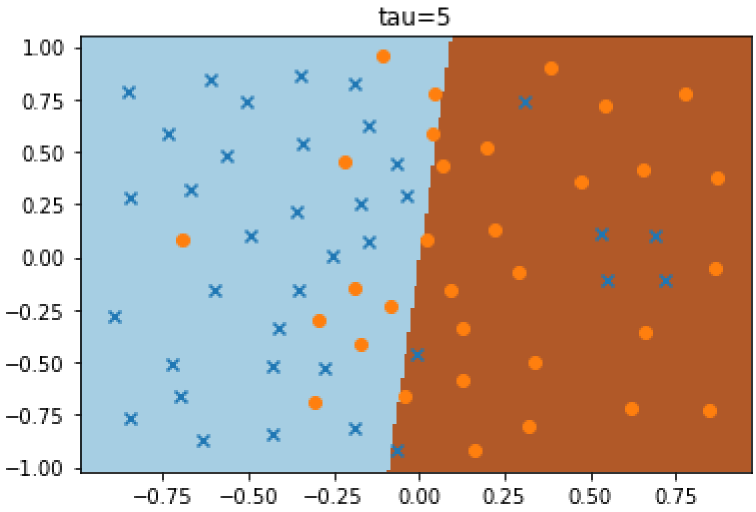



tau=0.05
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75


tau=0.1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75


tau=0.5
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75





参数τ越大，边界越平滑，如果是unweighted的形式，相当于$\tau \to \infty$，所以可以推断出unweighted的边界类似于$\tau = 5$时的情形。

备注：这里和标准答案的图不一样是因为答案中的代码为τ，实际应该是τ^2

3.Multivariate least squares

(a)注意到

$$X\Theta = \begin{bmatrix} (x^{(1)})^T\Theta \\ (x^{(2)})^T\Theta \\ \ldots \\ (x^{(m)})^T\Theta \end{bmatrix}$$

$$X\Theta - Y = \begin{bmatrix} (x^{(1)})^T\Theta - y^{(1)} \\ (x^{(2)})^T\Theta - y^{(2)} \\ \ldots \\ (x^{(m)})^T\Theta - y^{(m)} \end{bmatrix}$$

所以

$$
\begin{aligned}
(X\Theta-Y)^T(X\Theta-Y)_{ii} &= ((x^{(i)})^T\Theta-y^{(i)})^T((x^{(i)})^T\Theta-y^{(i)}) \\
&= (\Theta^T x^{(i)}-y^{(i)})^T(\Theta^T x^{(i)}-y^{(i)}) \\
&= \sum_{j=1}^{p}\Big((\Theta^T x^{(i)})_j-y_j^{(i)}\Big)^2
\end{aligned}
$$

$$
J(\Theta)=\frac{1}{2}\operatorname{tr}((X\Theta-Y)^T(X\Theta-Y))
$$

(b)

$$
\begin{aligned}
J(\Theta) &= \frac{1}{2}\operatorname{tr}((X\Theta-Y)^T(X\Theta-Y)) \\
&= \frac{1}{2}\operatorname{tr}(\Theta^TX^TX\Theta-Y^TX\Theta-\Theta^TX^TY+Y^TY) \\
&= \frac{1}{2}\operatorname{tr}(\Theta^TX^TX\Theta-2Y^TX\Theta+Y^TY)
\end{aligned}
$$

注意到

$$
\nabla_X\operatorname{tr}(AXB)=A^TB^T,\nabla_X\operatorname{tr}(X^TAX)=(A+A^T)X
$$

所以

$$
\nabla_\Theta J(\Theta)=\frac{1}{2}(2X^TX\Theta-2X^TY)=X^TX\Theta-X^TY
$$

令上式为0可得

$$
\Theta=(X^TX)^{-1}X^TY
$$

(c)如果化为p个独立的最小二乘问题，则

$$
\theta_j=(X^TX)^{-1}X^TY_{:,j}
$$

其中$Y_{:,j}$为Y的第j列，从而

$$
\Theta=[\theta_1,\ldots,\theta_p]
$$

4.Naive Bayes

(a)不难看出

$$
p(x|y=k)=\prod_{j=1}^{n}(\phi_{j|y=k})^{x_j}(1-\phi_{j|y=k})^{1-x_j}
$$

所以

$$
\begin{aligned}
\ell(\varphi) &= \log \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}; \varphi) \\
&= \sum_{i=1}^{m} \log p(x^{(i)}, y^{(i)}; \varphi) \\
&= \sum_{i=1}^{m} \log p(x^{(i)} | y^{(i)}) p(y^{(i)}) \\
&= \sum_{i=1}^{m} \log \prod_{j=1}^{n} (\phi_{j|y=y^{(i)}})^{x_j^{(i)}} (1-\phi_{j|y=y^{(i)}})^{1-x_j^{(i)}} (\phi_y)^{y^{(i)}} (1-\phi_y)^{1-y^{(i)}} \\
&= \sum_{i=1}^{m} \sum_{j=1}^{n} \Big(x_j^{(i)} \log(\phi_{j|y=y^{(i)}}) + (1-x_j^{(i)}) \log(1-\phi_{j|y=y^{(i)}}) \Big) + \sum_{i=1}^{m} \Big(y^{(i)} \log \phi_y + (1-y^{(i)}) \log(1-\phi_y) \Big)
\end{aligned}
$$

(b)先关于$\phi_{j|y=k}$求梯度

$$
\begin{aligned}
\nabla_{\phi_{j|y=k}} \ell(\varphi) &= \sum_{i=1}^{m} \Big(x_j^{(i)} \frac{1}{\phi_{j|y=y^{(i)}}} 1\{y^{(i)} = k\} + (1-x_j^{(i)}) \frac{1}{1-\phi_{j|y=y^{(i)}}} (-1) 1\{y^{(i)} = k\} \Big) \\
&= \sum_{i=1}^{m} \frac{1\{y^{(i)} = k\}}{\phi_{j|y=y^{(i)}} (1-\phi_{j|y=y^{(i)}})} \Big(x_j^{(i)} (1-\phi_{j|y=y^{(i)}}) - (1-x_j^{(i)}) \phi_{j|y=y^{(i)}} \Big) \\
&= \frac{1}{\phi_{j|y=k} (1-\phi_{j|y=k})} \sum_{i=1}^{m} 1\{y^{(i)} = k\} \Big(x_j^{(i)} - \phi_{j|k} \Big)
\end{aligned}
$$

令上式为0可得

$$
\begin{gathered}
\sum_{i=1}^{m} 1\{y^{(i)} = k\} \Big(x_j^{(i)} - \phi_{j|k} \Big) = 0 \\
(\sum_{i=1}^{m} 1\{y^{(i)} = k\}) \phi_{j|k} = \sum_{i=1}^{m} 1\{y^{(i)} = k\} x_j^{(i)} = \sum_{i=1}^{m} 1\{y^{(i)} = k \wedge x_j^{(i)} = 1\} \\
\phi_{j|k} = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = k \wedge x_j^{(i)} = 1\}}{\sum_{i=1}^{m} 1\{y^{(i)} = k\}}
\end{gathered}
$$

从而

$$
\begin{aligned}
\phi_{j|0} &= \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 0 \wedge x_j^{(i)} = 1\}}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}} \\
\phi_{j|1} &= \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1 \wedge x_j^{(i)} = 1\}}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}
\end{aligned}
$$

关于ϕ_y求梯度可得

$$
\begin{aligned}
\nabla_{\phi_y}\ell(\varphi) &= \sum_{i=1}^{m}\nabla_{\phi_y}\Big(y^{(i)}\log\phi_y + (1-y^{(i)})\log(1-\phi_y)\Big)\\
&= \sum_{i=1}^{m}\Big(y^{(i)}\frac{1}{\phi_y} - (1-y^{(i)})\frac{1}{1-\phi_y}\Big)\\
&= \frac{1}{\phi_y(1-\phi_y)}\sum_{i=1}^{m}\Big(y^{(i)}(1-\phi_y) - (1-y^{(i)})\phi_y\Big)\\
&= \frac{1}{\phi_y(1-\phi_y)}\sum_{i=1}^{m}\Big(y^{(i)} - \phi_y\Big)
\end{aligned}
$$

令上式为0可得

$$
\phi_y = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}{m}
$$

(c)

$$
\begin{aligned}
p(y=k|x) &= \frac{p(y=k,x)}{p(x)}\\
&= \frac{p(y=k,x)}{p(x|y=1)p(y=1)+p(x|y=0)p(y=0)}\\
&= \frac{p(x|y=k)p(y=k)}{p(x|y=1)p(y=1)+p(x|y=0)p(y=0)}\\
&= \frac{\phi_y^k(1-\phi_y)^{1-k}\prod_{j=1}^{n}(\phi_{j|y=k})^{x_j}(1-\phi_{j|y=k})^{1-x_j}}{\phi_y\prod_{j=1}^{n}(\phi_{j|y=1})^{x_j}(1-\phi_{j|y=1})^{1-x_j} + (1-\phi_y)\prod_{j=1}^{n}(\phi_{j|y=0})^{x_j}(1-\phi_{j|y=0})^{1-x_j}}
\end{aligned}
$$

所以

$$
\begin{aligned}
\frac{p(y=1|x)}{p(y=0|x)} &= \frac{\phi_y\prod_{j=1}^{n}(\phi_{j|y=1})^{x_j}(1-\phi_{j|y=1})^{1-x_j}}{(1-\phi_y)\prod_{j=1}^{n}(\phi_{j|y=0})^{x_j}(1-\phi_{j|y=0})^{1-x_j}}\\
&= \frac{\phi_y}{1-\phi_y}\Big(\prod_{j=1}^{n}\frac{1-\phi_{j|y=1}}{1-\phi_{j|y=0}}\Big)\exp\Big(\sum_{j=1}^{n}x_j\ln(\frac{\phi_{j|y=1}(1-\phi_{j|y=0})}{\phi_{j|y=0}(1-\phi_{j|y=1})})\Big)
\end{aligned}
$$

所以

$$
\frac{p(y=1|x)}{p(y=0|x)} \geq 1
$$

等价于

化简可得

$$
\begin{aligned}
\frac{\phi_y}{1-\phi_y}\Big(\prod_{j=1}^{n}\frac{1-\phi_{j|y=1}}{1-\phi_{j|y=0}}\Big)\exp\Big(\sum_{j=1}^{n}x_j\ln(\frac{\phi_{j|y=1}(1-\phi_{j|y=0})}{\phi_{j|y=0}(1-\phi_{j|y=1})})\Big)\geq 1\\
\exp\Big(\sum_{j=1}^{n}x_j\ln(\frac{\phi_{j|y=1}(1-\phi_{j|y=0})}{\phi_{j|y=0}(1-\phi_{j|y=1})})\Big)\geq \frac{1-\phi_y}{\phi_y}\prod_{j=1}^{n}\frac{1-\phi_{j|y=0}}{1-\phi_{j|y=1}}\\
\sum_{j=1}^{n}x_j\ln\Big(\frac{\phi_{j|y=1}(1-\phi_{j|y=0})}{\phi_{j|y=0}(1-\phi_{j|y=1})}\Big)\geq \ln\Big(\frac{1-\phi_y}{\phi_y}\prod_{j=1}^{n}\frac{1-\phi_{j|y=0}}{1-\phi_{j|y=1}}\Big)\\
\sum_{j=1}^{n}x_j\ln\Big(\frac{\phi_{j|y=1}(1-\phi_{j|y=0})}{\phi_{j|y=0}(1-\phi_{j|y=1})}\Big)-\ln\Big(\frac{1-\phi_y}{\phi_y}\prod_{j=1}^{n}\frac{1-\phi_{j|y=0}}{1-\phi_{j|y=1}}\Big)\geq 0
\end{aligned}
$$

令

$$
\theta_0=-\ln\Big(\frac{1-\phi_y}{\phi_y}\prod_{j=1}^{n}\frac{1-\phi_{j|y=0}}{1-\phi_{j|y=1}}\Big),\theta_j=\ln(\frac{\phi_{j|y=1}(1-\phi_{j|y=0})}{\phi_{j|y=0}(1-\phi_{j|y=1})})
$$

所以

$$
\frac{p(y=1|x)}{p(y=0|x)}\geq 1
$$

等价于

$$
\theta^T\begin{bmatrix}1\\x\end{bmatrix}\geq 0
$$

5.Exponential family and the geometric distribution

(a)

$$
\begin{aligned}
p(y;\phi)&=(1-\phi)^{y-1}\phi\\
&=\frac{\phi}{1-\phi}(1-\phi)^y\\
&=\exp(y\ln(1-\phi)-\ln(\frac{1-\phi}{\phi}))
\end{aligned}
$$

所以

$$
b(y)=1,\eta=\ln(1-\phi),T(y)=y,a(\eta)=\ln(\frac{1-\phi}{\phi})
$$

化简可得

$$
\begin{gathered}
e^\eta=1-\phi,\phi=1-e^\eta\\
a(\eta)=\ln(\frac{e^\eta}{1-e^\eta})
\end{gathered}
$$

综上

$$
b(y) = 1 \\
\eta = \ln(1-\phi) \\
T(y) = y \\
a(\eta) = \ln(\frac{e^\eta}{1-e^\eta})
$$

(b)

$$
\mathbb{E}[y|x;\theta] = \frac{1}{\phi} = \frac{1}{1-e^\eta}
$$

(c)由(b)可得

$$
\phi = 1-e^\eta
$$

带入

$$
p(y;\phi) = \exp(y\ln(1-\phi) - \ln(\frac{1-\phi}{\phi}))
$$

可得

$$
p(y;\phi) = \exp(y\eta - \ln(\frac{e^\eta}{1-e^\eta})) = \exp(y\eta - \eta + \ln(1-e^\eta))
$$

这里

$$
\eta = \theta^T x
$$

所以对数似然函数为

$$
\log p(y^{(i)}|x^{(i)};\theta) = y^{(i)}\theta^T x^{(i)} - \theta^T x^{(i)} + \ln(1-e^{\theta^T x^{(i)}})
$$

关于θ_j求偏导可得

$$
\begin{aligned}
\frac{\partial \log p(y^{(i)}|x^{(i)};\theta)}{\partial \theta_j} &= y^{(i)}x_j^{(i)} - x_j^{(i)} + \frac{1}{1-e^{\theta^T x^{(i)}}}(-e^{\theta^T x^{(i)}})x_j^{(i)} \\
&= (y^{(i)} - 1 - \frac{e^{\theta^T x^{(i)}}}{1-e^{\theta^T x^{(i)}}})x_j^{(i)} \\
&= (y^{(i)} - \frac{1}{1-e^{\theta^T x^{(i)}}})x_j^{(i)}
\end{aligned}
$$

所以

$$
\nabla_\theta \log p(y^{(i)}|x^{(i)};\theta) = (y^{(i)} - \frac{1}{1-e^{\theta^T x^{(i)}}})x^{(i)}
$$

所以随机梯度上升的更新规则为

$$\theta := \theta + \alpha\left(y^{(i)} - \frac{1}{1 - e^{\theta^T x^{(i)}}}\right)x^{(i)}$$