数据挖掘作业一

高阳 15331089 数媒

1. 线性回归

建立 m 与其它变量的多元线性回归方程：

$m(\theta) = \theta_0 + \theta_1 * p + \theta_2 * c + \theta_3 * e + \theta_4 * ch$

（1）利用梯度下降法，算出经过第一次迭代后的 θ（θ 的初始值全为 0，α = 1，m = 5，并设 y 为测试样本中的输出值）。

解：

开销函数为：

$J(\theta) = (1 / 2m) * \sum((m(\theta) - y) \wedge 2)$

设 J(θ)对每个 θ 的偏导数分别为 J’(θj)，则 θ 迭代公式为：

$\theta j := \theta j - \alpha * J'(\theta j)$

化简得：

$\theta j := \theta j - 0.2 * \sum((m(\theta) - y) * xj)$

其中 xj 依次为：1, p, c, e, ch

经计算，第一次迭代后 θ 的值分别为：

$\theta 0 := -0.2 * \sum(-y) = 93$

$\theta 1 := -0.2 * \sum(-y * x1) = 8376$

$\theta 2 := 6864.6$

$\theta 3 := 8059.8$

$\theta 4 := 8501.8$

（2）第一次迭代后的 θ 值设为 θ(1)，则 θ(1) > θ(0)

带入 J(θ)中，J(θ(1)) > J(θ(0))

所以不能使线性回归中的代价函数下降，因为 α 太大

（3）当 α * J’(θj)趋于 0 时，代价函数趋于最优值

所以有：$(\alpha / m) * \sum((m(\theta j) - y) * xj) \to 0$

对 θ_0，$\alpha * 93 \to 0$，有 $\alpha \approx 0.01$

或者可以预设 α 的几个可能的值 0.01,0.03,0.1,0.3，依次计算代价函数来确定目前最优的学习率 α 。

（4）利用标准方程求出最优的多元线性回归方程，并预计数学成绩。

$$X = \begin{bmatrix} 1 & 87 & 72 & 83 & 90 \\ 1 & 89 & 76 & 88 & 93 \\ 1 & 89 & 74 & 82 & 91 \\ 1 & 92 & 71 & 91 & 89 \\ 1 & 93 & 76 & 89 & 94 \end{bmatrix}$$

$y = [89\ 91\ 93\ 95\ 97]^T$

则 $\theta = (X^TX)^{-1}X^Ty$，通过 matlab 计算 pinv(x' * x) * x' * y 得：

```
>> pinv(x'*x)*x'*y

ans =

  -19.5000
    1.6875
    0.3750
   -0.3125
   -0.4375
```

$\theta_0 = -19.50$, $\theta_1 = 1.69$, $\theta_2 = 0.38$, $\theta_3 = -0.31$, $\theta_4 = -0.48$

当 p = 88,c = 73, e = 87, ch = 92 时，m = 85.83

（5）令 $\lambda = 1$，利用标准方程求出最优的 L2 正则化多元线性回归方程，并与（4）比较结果。

由题：$J(\theta_j) = (1 / 2m) * (\sum(m(\theta_j) - y) ^ 2 + \sum \theta_j ^2)$

利用上面的矩阵，此时 $\theta = (X^TX + c)^{-1}X^Ty$

这里 c = [0 0 0 0 0

0 1 0 0 0

0 0 1 0 0

0 0 0 1 0

0 0 0 0 1]

通过 matlab 计算得：

```
>> pinv(x'*x + c)*x'*y

ans =

  -19.9885
    1.4734
    0.0694
   -0.2257
   -0.0568
```

θ0 = -19.99, θ1 = 1.47, θ2 = 0.07, θ3 = -0.23, θ4 = -0.06

当 p = 88,c = 73, e = 87, ch = 92 时， m = 88.95

L2 正则化多元线性回归方程更好。

2. 逻辑回归

（1）使用逻辑回归或实现求解 L2 逻辑回归分析的梯度下降算法，求出最优的逻辑回归模型。

设 hθ(x) = g(θTx), g(z) = 1 / (1 + e ^(-z))

则 hθ(x) = 1 / (1 + e ^ (-θTx))

逻辑回归的代价函数（其中 m = 40）：

$$
\begin{aligned}
J(\theta) &= \frac{1}{m}\sum_{i=1}^{m}\mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)}) \\
&= -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)}\log h_\theta(x^{(i)}) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))]
\end{aligned}
$$

则 θj := θj - α * ∑((h(θj) - y) * xj)

调用 matlab 自带的逻辑回归函数：

先在 txt 文件中写入测试数据，再读入数据使用函数进行计算（其中第一列为目标结果，后四列为测试的四项指标）：

data.txt - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

1	1	1	0	1
0	0	1	0	0
1	1	0	1	1
0	0	0	0	1
1	1	1	0	1
0	1	0	1	1
1	1	0	0	0
0	1	0	1	1
1	1	0	1	1
0	0	0	0	0
1	1	1	0	1
0	0	0	0	0
1	1	0	0	1

```
>> load data.txt;
>> x = data(:, 2:5);
>> y = data(:, 1);
>> b = glmfit(x, y, 'binomial','link','logit');
>> p = glmval(b,x,'logit');
```

（2）找出最直接的影响因素。

由上面最优的逻辑回归模型可知，θ 最 λ 大的系数的特征影响最大。

（3）求解 L2 正则化逻辑回归分析的梯度下降算法，并求当平衡系数为 1 时最优正则化逻辑回归模型。

$J(\theta) = -(1/m) * [\sum(y * \log(h_\theta(x)) + (1-y) * \log(1 - h_\theta(x)))] + (\lambda/2m) * \sum(\theta^2)$

则对 θ 的梯度下降：

$\theta_0 := \theta_0 - (\alpha/m) * \sum(h_\theta(x) - y) * x_0$

$\theta_j := \theta_j - \alpha[(1/m) * \sum(h_\theta(x) - y) * x_j + (\lambda/m) * \theta_j]$ $(j = 1, 2, 3...)$

当平衡系数 $\lambda = 1$ 时，

θ_0 不变；

$\theta_j := \theta_j - \alpha[(1/m) * \sum(h_\theta(x) - y) * x_j + (1/m) * \theta_j]$ $(j = 1, 2, 3...)$

3. 支持向量机

令+ := y = 1

- := y = 0

（1）

X = [1 1 1

1 2 2

1 2 0

1 0 0

1 1 0

1 0 1]

Y = [1 1 1 0 0 0]T

设 h θ (x) = 1 / (1 + e ^ * (- θ T x)

当 y = 1 时，令 h θ (x)≈1，则 θ Tx >= 1

当 y = 0 时，令 h θ (x)≈0，则 θ Tx <= -1

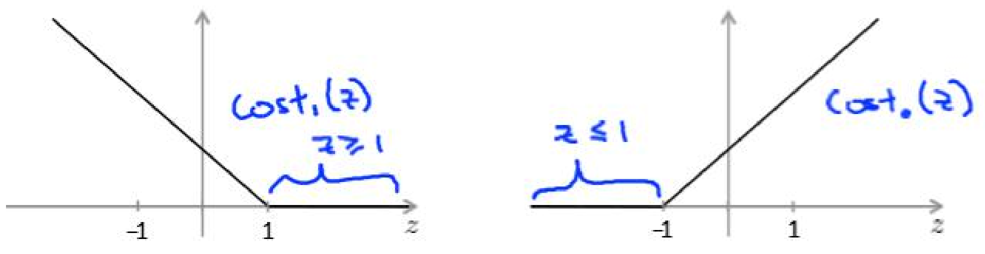
则逻辑回归方程：

$$\min_{\theta} \frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)}\left(-\log h_\theta(x^{(i)})\right)+(1-y^{(i)})\left((-\log(1-h_\theta(x^{(i)}))\right)\right]+\frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$

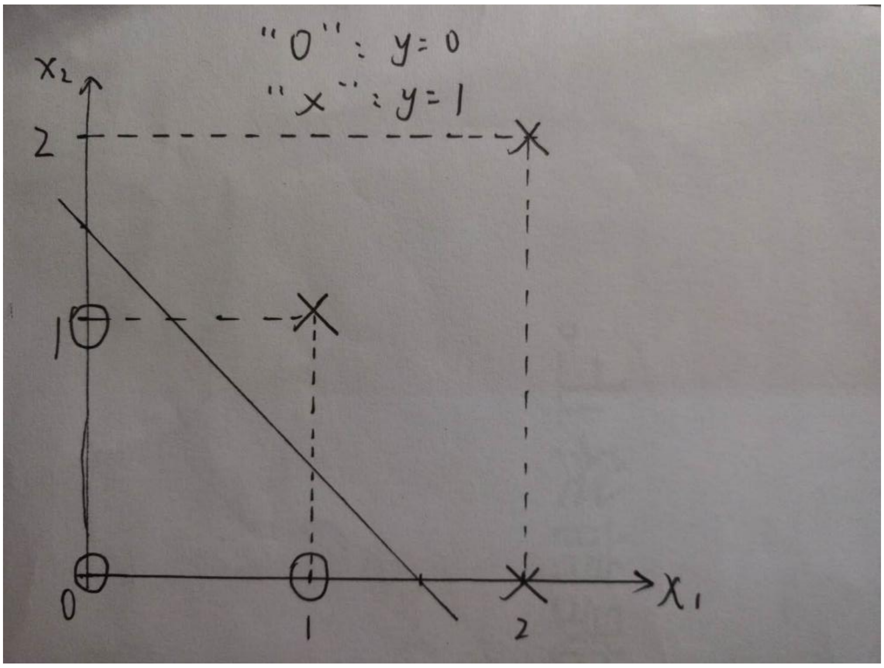
其支持向量机为：

$$\min_{\theta} C\sum_{i=1}^{m}\left[y^{(i)}cost_1(\theta^T x^{(i)})+(1-y^{(i)})cost_0(\theta^T x^{(i)})\right]+\frac{1}{2}\sum_{i=1}^{n}\theta_j^2$$

其中 C = 1 / λ， cost1,cost0 如下：



画出 6 个训练样本及决策边界：



求最优超平面（决策边界，令θ0 = 0）：

即求 min(0.5 * (θ1 + θ2) ^ 2) = min(0.5 * (||θ||) ^ 2)

又θTx = θ1x1 + θ2x2 = p * ||θ|| (p 为向量 x 在向量θ上的投影)

则：

$$\|\theta\| \cdot p^{(i)} \geq 1 \quad \text{if } y^{(i)} = 1$$
$$\|\theta\| \cdot p^{(i)} \leq -1 \quad \text{if } y^{(i)} = 0$$

解||θ|| = √2 / 4

得出最优超平面方程为：-1.5 + x1 +x2 = 0

（2）因为新增的训练样本正确分类并远离最优超平面，所以不会对最优超平面产生影响。而线性回归每一点都参与决策，都对结果产生影响。

（3）支持向量：(1, 1),(2, 0), (1, 0), (0, 1)

两个异类支持向量到最优超平面的距离均为√2 / 4，之和为√2 / 2

（4）设约束最优化问题：

$$\min_{\theta} \frac{1}{2}\sum_{j=1}^{n}\theta_j^2$$

$$\text{s.t. } \|\theta\|\cdot p^{(i)} \geq 1 \quad \text{if } y^{(i)}=1$$

$$\|\theta\|\cdot p^{(i)} \leq -1 \quad \text{if } y^{(i)}=0$$

引入拉格朗日乘子：

L(θ，αi) = 0.5 * (||θ||) ^ 2 + ∑(αi* ||θ|| * pi)

将上面拉格朗日函数对θ求导并令导数为 0 得：

θ = ∑(αi * pi)

带入拉格朗日函数中：

Max θ(α) = ∑(αi) - 0.5 * ∑∑(αi * αj * piT * pj)

即知道α或θ中的一个，另一个就能求出来。

求对α的极大值，即关于对偶变量α的优化问题。当求得最优的α’后，就可以带入上面的公式，导出θ’。