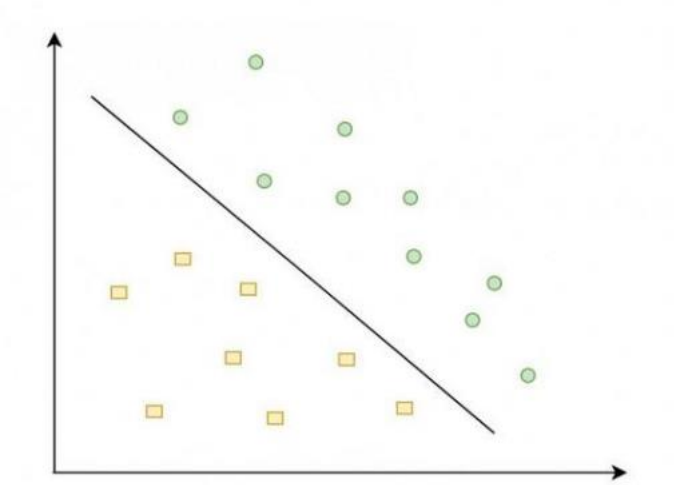
精英班系列课程

逻辑回归

逻辑回归

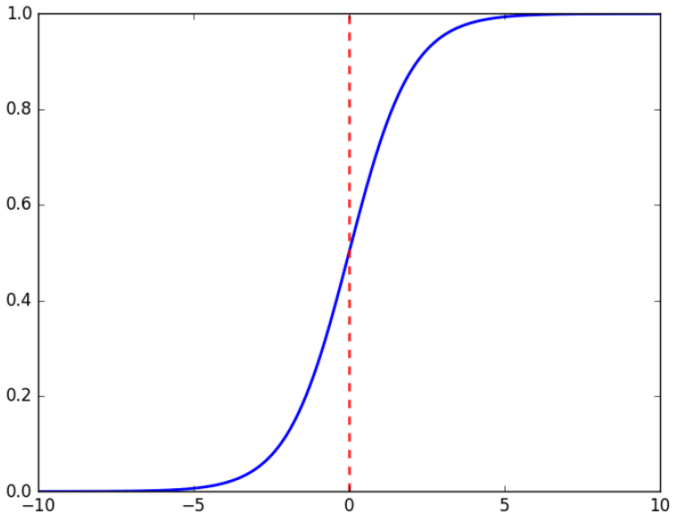
- 逻辑回归：
  - ✓ Logistic回归是概率型非线性回归模型，是研究二值型输出分类的一种多变量分析方法。通过logistic回归我们可以将二分类的观察结果y与一些影响因素[x1, x2, x3, …]建立起关系从而对某些因素条件下某个结果发生的概率进行估计并分类



逻辑回归

- Sigmoid函数：
  - ✓ 对于二分类问题，我们想要一个函数能够接受所有输入然后预测出两种类别，可以通过输出0或者1。这个函数就是sigmoid函数，它是一种阶跃函数具体的计算公式如下：

$$\sigma(z) = \frac{1}{1+e^{-z}}$$



逻辑回归

- Sigmoid函数特性：
  - ✓ sigmoid函数连续，严格单调，以(0,0.5)中心对称，是一个非常良好的阈值函数
  - ✓ 当x为0时，Sigmoid函数值为0.5，随着x的增大对应的Sigmoid值将逼近于1；而随着x的减小，Sigmoid函数会趋近于0
  - ✓ Sigmoid函数的值域范围限制在(0,1)之间，与概率值的范围是相对应的
  - ✓ Sigmoid函数的导数是其本身的函数，即f'(x)=f(x)(1−f(x))，计算方便

逻辑回归

- Logistic回归分类器(Logistic Regression Classifier)：
  - ✓ z是一个矩阵，θ是参数列向量(要求解的)，x是样本向量(给定的数据集)，hθ(x)给出了输出为1的概率

$$Z = X\theta = (\mathrm{x}_0 \quad \mathrm{x}_1 \quad \cdots \quad \mathrm{x}_n) \cdot \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{pmatrix}$$

$$g(z) = \frac{1}{1 + e^{-z}}$$

$$h_\theta(x) = g(x\theta) = \frac{1}{1 + e^{-x\theta}}$$

逻辑回归

- 如何得到合适的参数向量θ？：
  - ✓ 在已知样本x和参数θ的情况下，样本x属性为正样本(y=1)和负样本(y=0)的条件概率

$$P(y=1 \mid x;\theta) = h_\theta(x)$$
$$P(y=0 \mid x;\theta) = 1 - h_\theta(x)$$

逻辑回归

- 如何得到合适的参数向量θ？：
  - ✓ 将两个公式合并，得单个样本属于y的概率为：

$$P(h_\theta(\mathrm{x}),\mathrm{y}) = h_\theta(x)^y(1-h_\theta(x))^{(1-y)}$$

逻辑回归

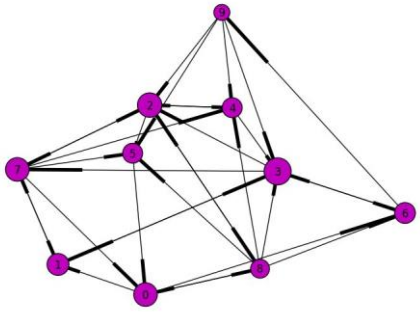
- 最大似然估计：
  - ✓ 假定样本与样本之间相互独立，那么整个样本集生成的概率即为所有样本生成概率的乘积，并对结果取对数：

$$L(\theta) = \sum_{i=1}^{m} [y^{(i)} \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))]$$

其中，m为样本的总数，y(i)表示第i个样本的类别，x(i)表示第i个样本，θ是多维向量，x(i)也是多维向量

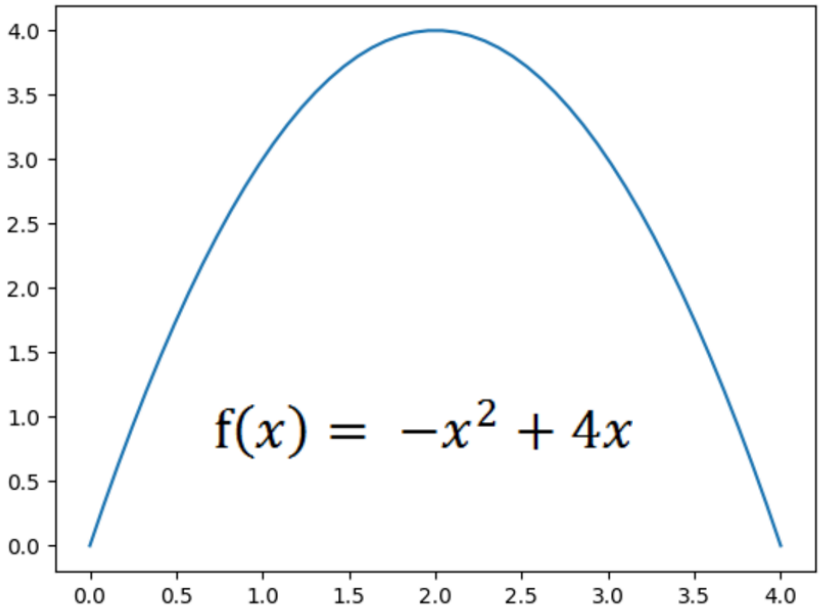
满足L(θ)的最大的θ值即是我们需要求解的模型

梯度上升法



梯度上升法

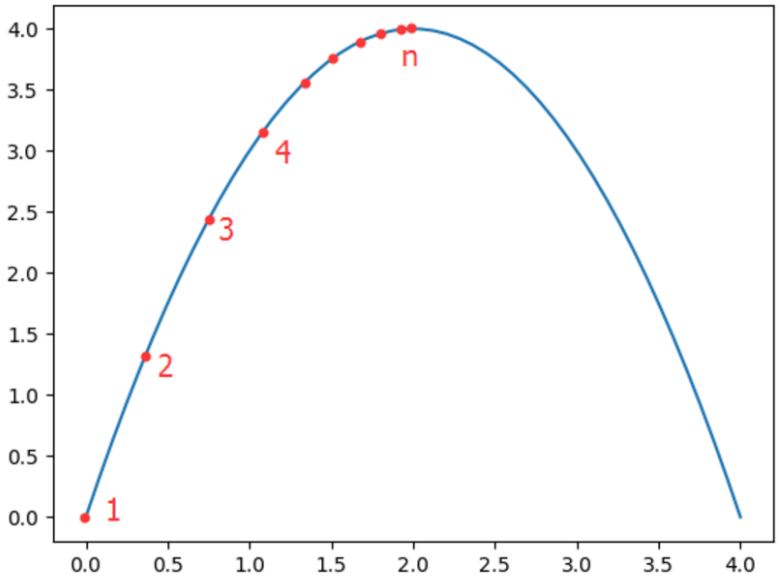
- 梯度上升法：
  - ✓ 求最大值，使用梯度上升算法



梯度上升法

- 梯度上升法：
  - ✓ 最优化算法：梯度上升算法

$$x_{i+1} = x_i + \alpha \frac{\partial f(x_i)}{x_i}$$



梯度上升法

- L(θ)函数的极大值：
  - ✓ 迭代更新公式：

$$\theta_j := \theta_j + \alpha \frac{\partial L(\theta)}{\theta_j}$$

$$\mathrm{L}(\theta) = \sum_{i=1}^{m} [y^{(i)} \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))]$$

$$h_\theta(x) = g(x\theta) = \frac{1}{1 + e^{-x\theta}}$$

梯度上升法

- L(θ)函数的极大值：
  - ✓ 求解L(θ)对θ的偏导：

$$\frac{\partial}{\theta_j}L(\theta)=\frac{\partial L(\theta)}{\partial g(x\theta)}*\frac{\partial g(x\theta)}{\partial x\theta}*\frac{\partial x\theta}{\partial \theta_j}$$

$$h_\theta(x)=g(x\theta)=\frac{1}{1+e^{-x\theta}}$$

$$\frac{\partial L(\theta)}{\partial g(x\theta)}=y*\frac{1}{g(x\theta)}+(y-1)*\frac{1}{1-g(x\theta)}$$

梯度上升法

- L(θ)函数的极大值：

$$\frac{\partial g(x\theta)}{\partial x\theta} = g(x\theta)(1 - g(x\theta))$$

$$\frac{\partial x\theta}{\theta_j} = \frac{\partial(\theta_1 x_1 + \theta_2 x_2 + \cdots \theta_n x_n)}{\partial \theta_j} = x_j$$

$$\frac{\partial}{\theta_j} L(\theta) = (y - h_\theta(x)) x_j$$

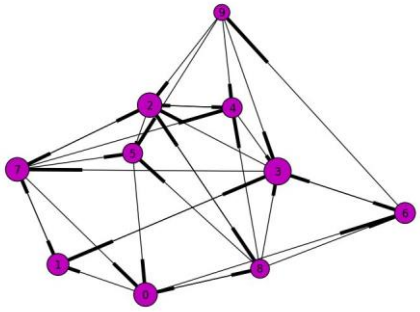
最后可得：$\theta_j := \theta_j + \alpha \sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)})) x_j^{(i)}$

矩阵形式：$\theta := \theta + \alpha X^T(\vec{y} - g(X\theta))$

梯度上升法

- Logistic回归梯度上升法python实现

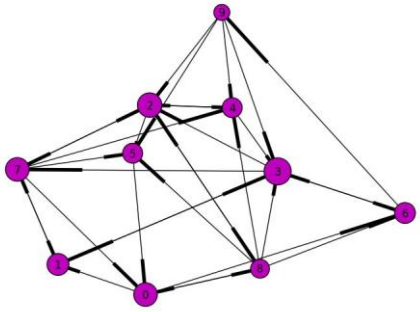
使用sklearn构建Logistic回归分类器



随机梯度上升

- 使用sklearn构建Logistic回归分类器

从疝气病症状预测病马的死亡率



实战

- 从疝气病症状预测病马的死亡率
  - ✓ 数据包含了368个样本和28个特征。这种病不一定源自马的肠胃问题，其他问题也可能引发马疝病。该数据集中包含了医院检测马疝病的一些指标，有的指标比较主观，有的指标难以测量，例如马的疼痛级别
  - ✓ 利用Logistic回归预测病马的生死

Talk is cheap

Show me the

CODE