开课吧
kaikeba.com
慧科集团旗下企业

逻辑回归

机器学习算法之二

- 1 分类问题
- 2 sigmod函数
- 3 理论推导
- 4 softmax回归

logistic 回归是什么

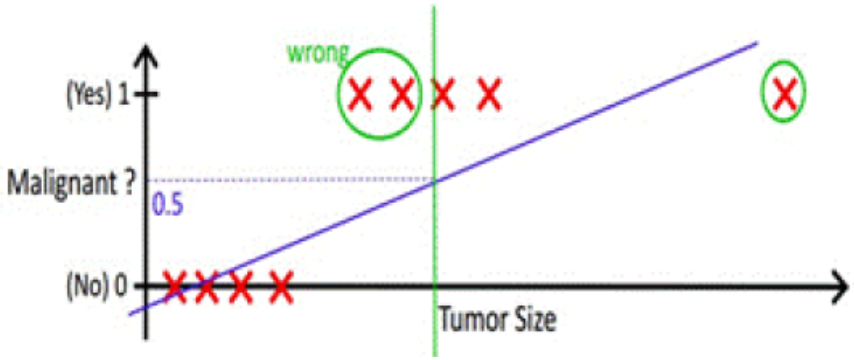
- 工业中的分类问题更多。
- 用户分级
- 是否放款
- 是否点击
- 是否配对
- 是否推荐

分类？

or

回归？

回归作为分类可行么？



用户	购买频次，浏览频次，时间，地理位置 ...
品类	销量，购买用户，浏览用户 ...
交叉	购买频次，浏览频次，购买间隔 ...

- logistic回归是一种广义线性模型（generalized linear model），因此与多重线性回归分析有很多相同之处。它们的模型形式基本上相同，都具有 wx+b，其中w和b是待求参数，其区别在于他们的因变量不同，多重线性回归直接将wx+b作为因变量，即y =wx+b，而logistic回归则通过函数L将wx+b对应一个隐状态p，p =L(wx+b),然后根据p 与1-p的大小决定因变量的值。如果L是logistic函数，就是logistic回归。

- 指数分布族

$$p(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$$

T:充分统计量
η :自然参数
$a(\eta)$被称为对数配分函数，实际上是归一化因子

我们见过的大多数分布都属于指数分布族，Bernoulli伯努利分布、Gaussian高斯分布、multinomial多项分布、Poisson泊松分布、gamma分布、指数分布、Dirichlet分布

- **指数分布族**
  - 正态分布明显指数分布

$$
\begin{aligned}
p(y;\mu) &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(y-\mu)^2\right) \\
&= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y^2\right)\cdot\exp\left(\mu y-\frac{1}{2}\mu^2\right)
\end{aligned}
$$

$$
\begin{aligned}
\eta &= \mu \\
T(y) &= y \\
a(\eta) &= \mu^2/2 \\
&= \eta^2/2 \\
b(y) &= (1/\sqrt{2\pi})\exp(-y^2/2)
\end{aligned}
$$

- **指数分布族**
  - 伯努力分布

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

$$
\eta = \log(\phi/(1-\phi))
$$

$$
\begin{aligned}
T(y) &= y \\
a(\eta) &= -\log(1-\phi) \\
&= \log(1+e^{\eta}) \\
b(y) &= 1
\end{aligned}
$$

参数Φ指的是y=1的概率，
即事件发生的概率

- **广义线性模型三个假设：**
- 1. 我们的目标是求条件概率$p(y \mid x; \theta)$
- 2. 条件概率分布服从指数分布族
- 3. 指数族分布中的η和输入变量x的关联是线性的：$\eta = \theta^T x$

- 我们需要解决的是分类问题，并且是只分为两类，马上想到用
  - 1.伯努利分布来估计事件发生的概率。
  - 2.伯努利分布属于指数族分布
  - 3.不妨接受假设：$\eta = \theta^T x$

$\eta = \log(\Phi/(1-\Phi))$ 参数Φ指的是y=1的概率，即事件发生的概率

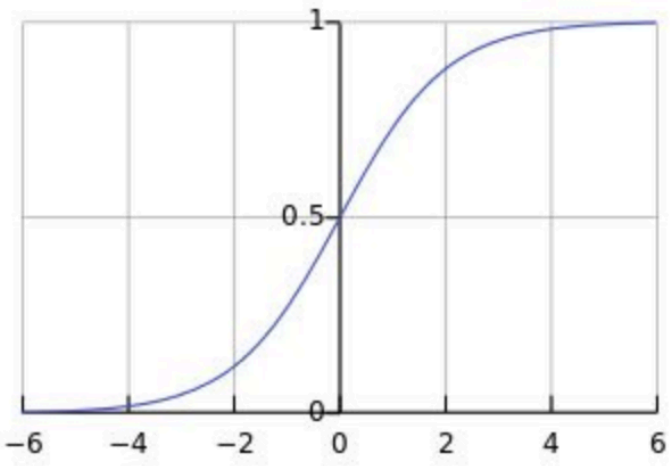
$\eta = \theta^T x$

联立两个公式：

$$\phi = \frac{1}{1+e^{-\theta^T x}}$$

这就是有名的sigmoid，也叫**Logistic函数**

- Sigmod函数求导：



$$
\begin{aligned}
f'(z) &= \left(\frac{1}{1+e^{-z}}\right)' \\
&= \frac{e^{-z}}{(1+e^{-z})^2} \\
&= \frac{1+e^{-z}-1}{(1+e^{-z})^2} \\
&= \frac{1}{(1+e^{-z})}\left(1-\frac{1}{(1+e^{-z})}\right) \\
&= f(z)(1-f(z))
\end{aligned}
$$

函数：$f(z)=\dfrac{1}{1+e^{-z}}$

极大似然的思想：

$$
\begin{aligned}
L(\theta) &= p(\vec{y} \mid X; \theta) \\
&= \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}; \theta) \\
&= \prod_{i=1}^{m} \left(h_\theta(x^{(i)})\right)^{y^{(i)}} \left(1 - h_\theta(x^{(i)})\right)^{1-y^{(i)}}
\end{aligned}
$$

$$
\begin{aligned}
\ell(\theta) &= \log L(\theta) \\
&= \sum_{i=1}^{m} y^{(i)} \log h(x^{(i)}) + (1 - y^{(i)}) \log(1 - h(x^{(i)}))
\end{aligned}
$$

- 梯度上升：

$$
\begin{aligned}
\frac{\partial}{\partial \theta_j}\ell(\theta) &= \left(y\frac{1}{g(\theta^T x)} - (1-y)\frac{1}{1-g(\theta^T x)}\right)\frac{\partial}{\partial \theta_j}g(\theta^T x) \\
&= \left(y\frac{1}{g(\theta^T x)} - (1-y)\frac{1}{1-g(\theta^T x)}\right)g(\theta^T x)(1-g(\theta^T x)\frac{\partial}{\partial \theta_j}\theta^T x \\
&= \left(y(1-g(\theta^T x)) - (1-y)g(\theta^T x)\right)x_j \\
&= \left(y - h_\theta(x)\right)x_j
\end{aligned}
$$

$$
\theta_j := \theta_j + \alpha\left(y^{(i)} - h_\theta(x^{(i)})\right)x_j^{(i)}
$$

构造代价函数角度

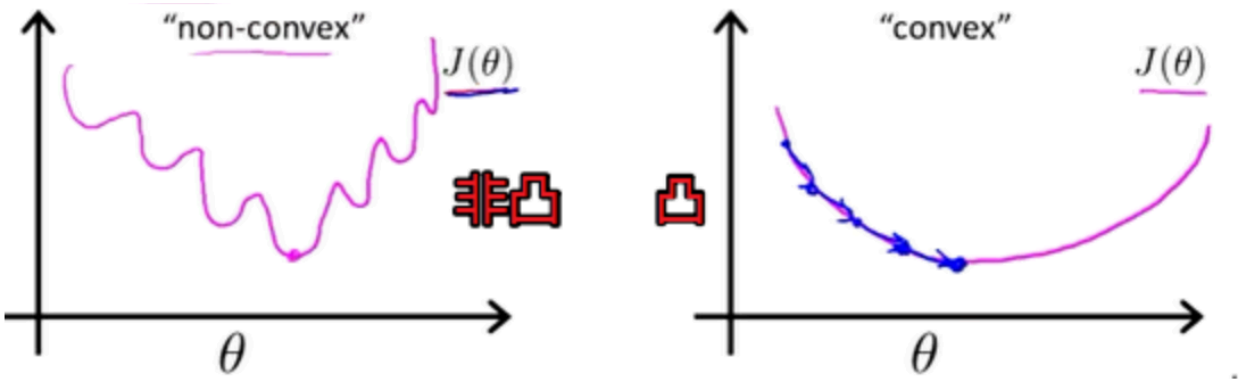
- 选择sigmoid函数做映射

代价函数： $J(\theta) = \frac{1}{m}\sum_{i=1}^{m} Cost(h_\theta(x^{(i)}), y^{(i)})$

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

如何确定Cost的显示的形式？

理论推导

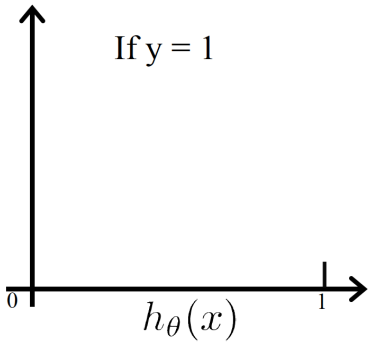


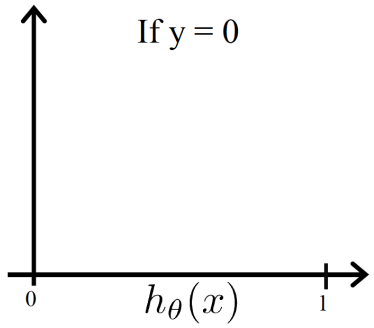
不好的损失函数　　好的损失函数

选择平方差损失函数，经过sigmoid映射之后，为非凸。不合适

构造代价函数：

$$Cost(h_\theta(x), y) = \begin{cases} -log(h_\theta(x)), & \text{if } y = 1 \\ -log(1 - h_\theta(x)), & \text{if } y = 0 \end{cases}$$





等价于：

$$Cost(h_\theta(x), y) = -ylog(h_\theta(x)) - (1 - y)log(1 - h_\theta(x))$$

- 多分类softmax

如果需要分为k个类别，概率函数的取值：

Sigmoid 是 softmax 一个特例

$$h_\theta(x^{(i)}) = \begin{bmatrix} p(y^{(i)}=1|x^{(i)};\theta) \\ p(y^{(i)}=2|x^{(i)};\theta) \\ \vdots \\ p(y^{(i)}=k|x^{(i)};\theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^{k} e^{\theta_j^T \cdot x^{(i)}}} \begin{bmatrix} e^{\theta_1^T \cdot x^{(i)}} \\ e^{\theta_2^T \cdot x^{(i)}} \\ \vdots \\ e^{\theta_k^T \cdot x^{(i)}} \end{bmatrix}$$

代价函数：

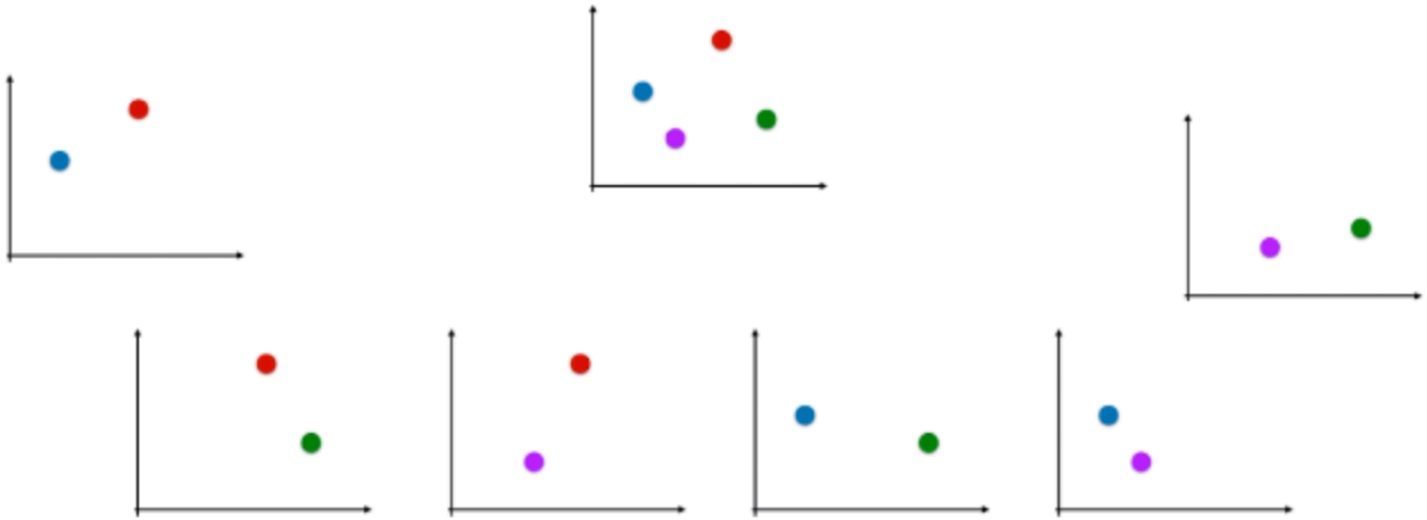
$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\{y^{(i)}=j\} \cdot log(p(y^{(i)}=j|x^{(i)};\theta))\right]$$

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\{y^{(i)}=j\} \cdot (\theta_j^T x^{(i)} - log(\sum_{l=1}^{k} e^{\theta_l^T \cdot x^{(i)}}))\right]$$

- **OvR 与 OVO**
  - 本质上是在二分类的问题上进行的策略改造。改造方法不是指针对逻辑回归算法，而是在机器学习领域有通用性
  - **OvR（One vs Rest）**
  - **OvO（One vs One）**

- **OvR**
  - **思想**：n 种类型的样本进行分类时，**分别**取一种样本作为一类，将剩余的所有类型的样本看做另一类，这样就形成了 **n 个**二分类问题，使用逻辑回归算法对 n 个数据集训练出 n 个模型，将待预测的样本传入这 n 个模型中，所得概率最高的那个模型对应的样本类型即认为是该预测样本的类型；

- **OvO**
  - **思想：** n 类样本中，每次挑出 2 种类型，两两结合，一共有 C_n^2 种二分类情况，使用 C_n^2 种模型预测样本类型，有 C_n^2 个预测结果，种类最多的那种样本类型，就认为是该样本最终的预测类型；



- LR优缺点和注意事项:
  - 优点
    - 以概率的形式输出结果, 可以用于排序
    - 可解释性强，可控度高
    - 训练快、效果不错
  - 注意事项：
    - 注意样本保持均衡
    - 合适的正则化

- sklearn是机器学习中一个常用的python第三方模块，网址：http://scikit-learn.org/stable/index.html

- 里面对一些常用的机器学习方法进行了封装，只需要简单的调用sklearn里的模块就可以实现大多数机器学习任务。

- 常用的分类器包括SVM、KNN、贝叶斯、线性回归、逻辑回归、决策树。

- Sklearn 进行鸢尾花分类（课堂实践）：

  - Iris（鸢尾花）数据集是多重变量分析的数据集。
    数据集包含150行数据，分为3类，每类50行数据。
    每行数据包括4个属性：Sepal Length（花萼长度）、Sepal Width（花萼宽度）、Petal Length（花瓣长度）、Petal Width（花瓣宽度）。可通过这4个属性预测鸢尾花属于3个种类的哪一类。

逻辑回归需要注意的参数：

1. 正则化选择参数：penalty

目的解决过拟合，一般penalty选择L2正则化就够了。
模型系数稀疏化，选择L1

2.优化算法：

solver : *str, {'newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'}, default: 'liblinear'.*

- For small datasets, 'liblinear' is a good choice, whereas 'sag' and 'saga' are faster for large ones.
- For multiclass problems, only 'newton-cg', 'sag', 'saga' and 'lbfgs' handle multinomial loss; 'liblinear' is limited to one-versus-rest schemes.
- 'newton-cg', 'lbfgs' and 'sag' only handle L2 penalty, whereas 'liblinear' and 'saga' handle L1 penalty.

逻辑回归需要注意的参数：

3. 分类方式：multi_class
ovr和multinomial两个值可以选择，默认是 ovr

4. class_weight 和 sample_weight
用于解决样本不平衡问题的。

特征工程

- 有这么一句话在业界广泛流传：数据和特征决定了机器学习的上限，而模型和算法只是逼近这个上限而已。

- 特征工程的目的是最大限度地从原始数据中提取特征以供算法和模型使用

- 数据通常是复杂冗余，富有变化的，有必要从原始数据发现有用的特性。人工选取出来的特征依赖人力和专业知识。

特征工程

- 特征的标准化
  - **z-score**标准化
    - （x-mean)/std
  - **max-min**标准化
    - (x-min)/(max-min)
  - 离群值
    - 特征值限制到 某个上限或者下限
    - 3σ法则

特征工程

- one-hot编码

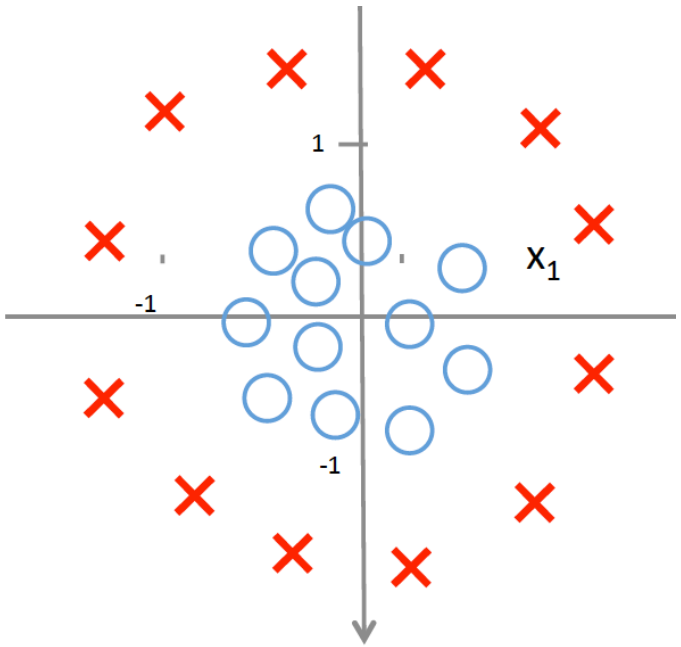
- 性别：["male"，"female"]
- 地区：[”北京"，”上海"，”杭州"]
- 浏览器：["Firefox"，"Chrome"，"Safari"，"Internet Explorer"]

- 可以用以下向量表示：
- [0,1],[1,0]
- [1,0,0],[0,1,0],[0,0,1]
- [1,0,0,0],[0,1,0,0], [0,0,1,0],[0,0,0,1]

特征工程

- 分箱法
  - 有时候，将数值型属性转换成类别呈现更有意义，同时能使算法减少噪声的干扰
  - 年龄分布划分成1-10,11-18,19-25,26-40等级。
  - 收入分布划分等级。

- 特征构造



$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_1^2 x_2 + \theta_5 x_1^2 x_2^2 + \theta_6 x_1^3 x_2 + \ldots)$$

构造非线性特征

- 缺失值处理

  - 1.缺失值作为一个单独特征
  - 2.填充
  - 3.舍弃这部分样本。

- 作业https://www.kaggle.com/c/predicting-red-hat-business-value

- 1. 清洗数据
- 2. 数据预处理
- 3.切分训练测试集
- 4.选择算法，训练模型
- 5.交叉验证调参
- 6.终止模型