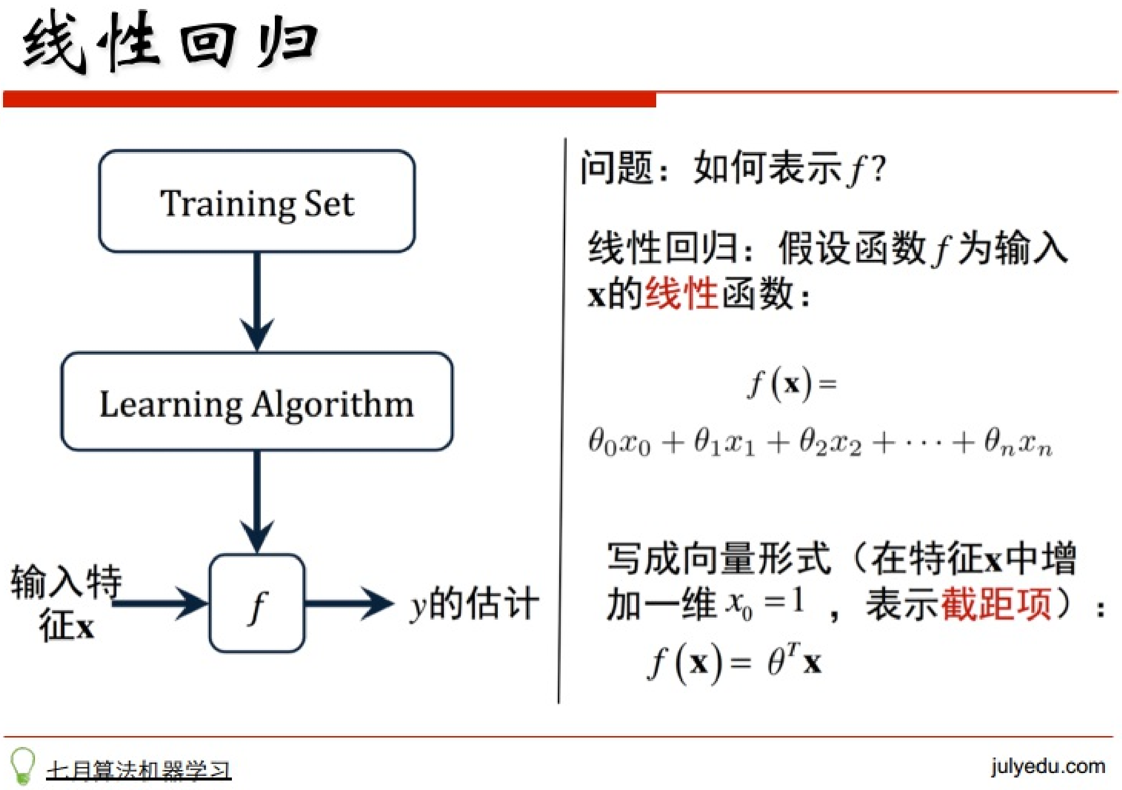
四、回归分析

一、线性回归

线性回归的作用是去预测连续值



线性回归的损失函数

- 怎么去衡量“最好”？
  - 我们把x到y的映射函数f记作 θ 的函数 $h_\theta(x)$
  - 定义损失函数为：

$$J(\theta_0, \theta_1, \ldots, \theta_n) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

具体训练

- 梯度下降
  - 假如现在有n个特征/变量 x_j (j=1…n)

$$\text{repeat until convergence } \{ \quad \theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \quad \}$$

$$\theta_0 := \theta_0 - \alpha \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_0^{(i)}$$

$$\theta_1 := \theta_1 - \alpha \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_1^{(i)}$$

$$\theta_2 := \theta_2 - \alpha \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_2^{(i)}$$

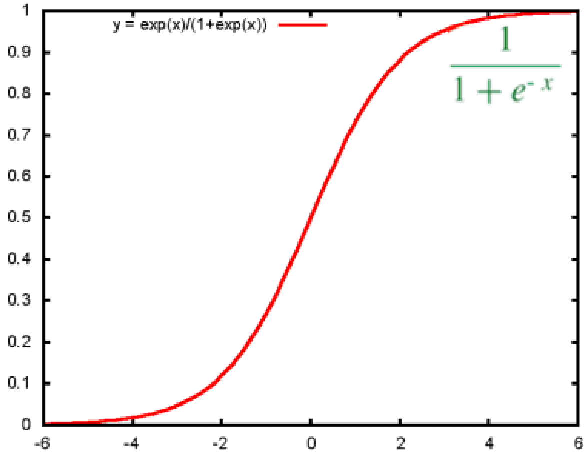
$$\cdots$$

二、逻辑回归

逻辑回归，可以用来进行排序，因为逻辑回归的结果是一个具体的概率值，可以根据概率的大小进行排序。

逻辑回归的具体定义

□ 归功于sigmoid



逻辑回归的损失函数：

□ 损失函数

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

梯度下降

□ 损失函数

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

$$\Big\Downarrow \text{ 别忘了正则化项}$$

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

LR的应用：

LR应用经验

- LR < SVM/GBDT/RandomForest ?
  - 并不是
    - LR能以概率的形式输出结果，而非只是0,1判定
    - LR的可解释性强，可控度高(你要给老板讲的嘛…)
    - 训练快，feature engineering之后效果赞
    - 因为结果是概率，可以做ranking model
    - 添加feature太简单…
  - 应用
    - CTR预估/推荐系统的learning to rank/各种分类场景
    - 某搜索引擎厂的广告CTR预估基线版是LR
    - 某电商搜索排序基线版是LR(广告也是哦)
    - 某电商的购物搭配推荐用了大量LR
    - 某现在一天广告赚1000w+的新闻app排序基线是LR

三、