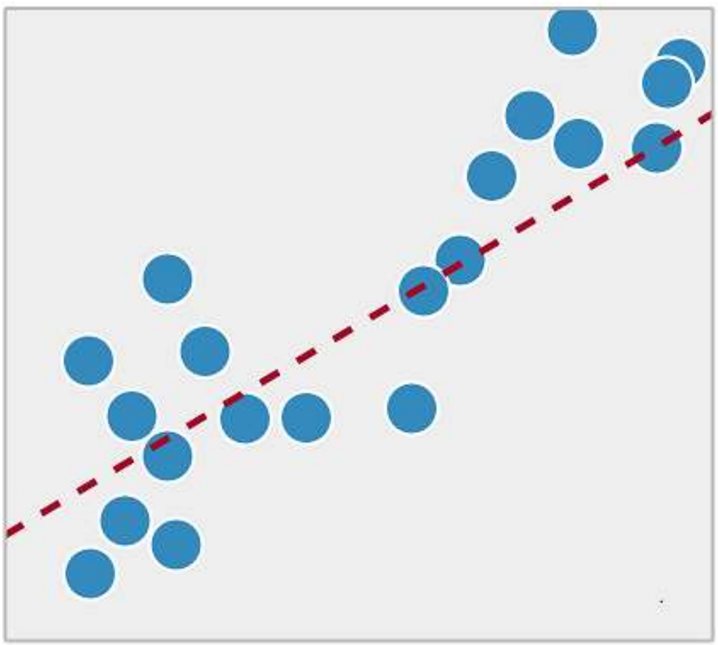
逻辑回归

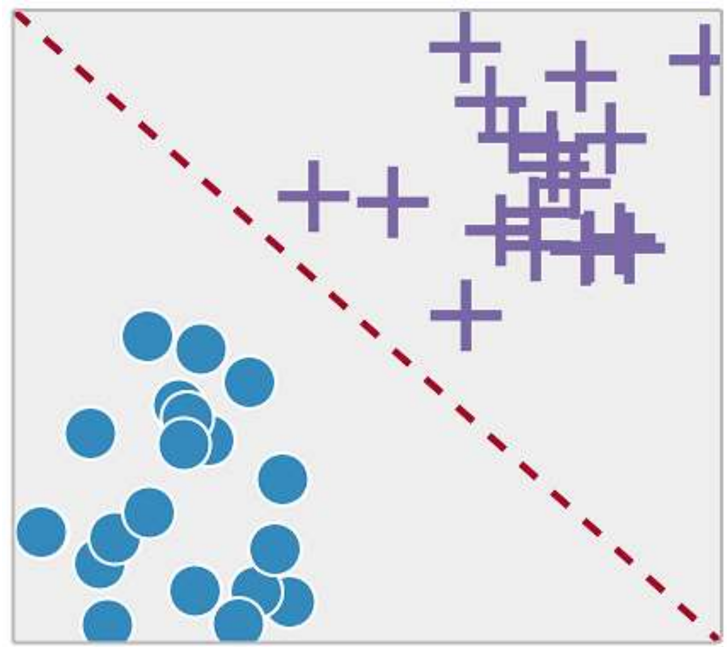
黄晟
huangsheng@cqu.edu.cn
办公室：信息大楼B701

回归与分类

Regression



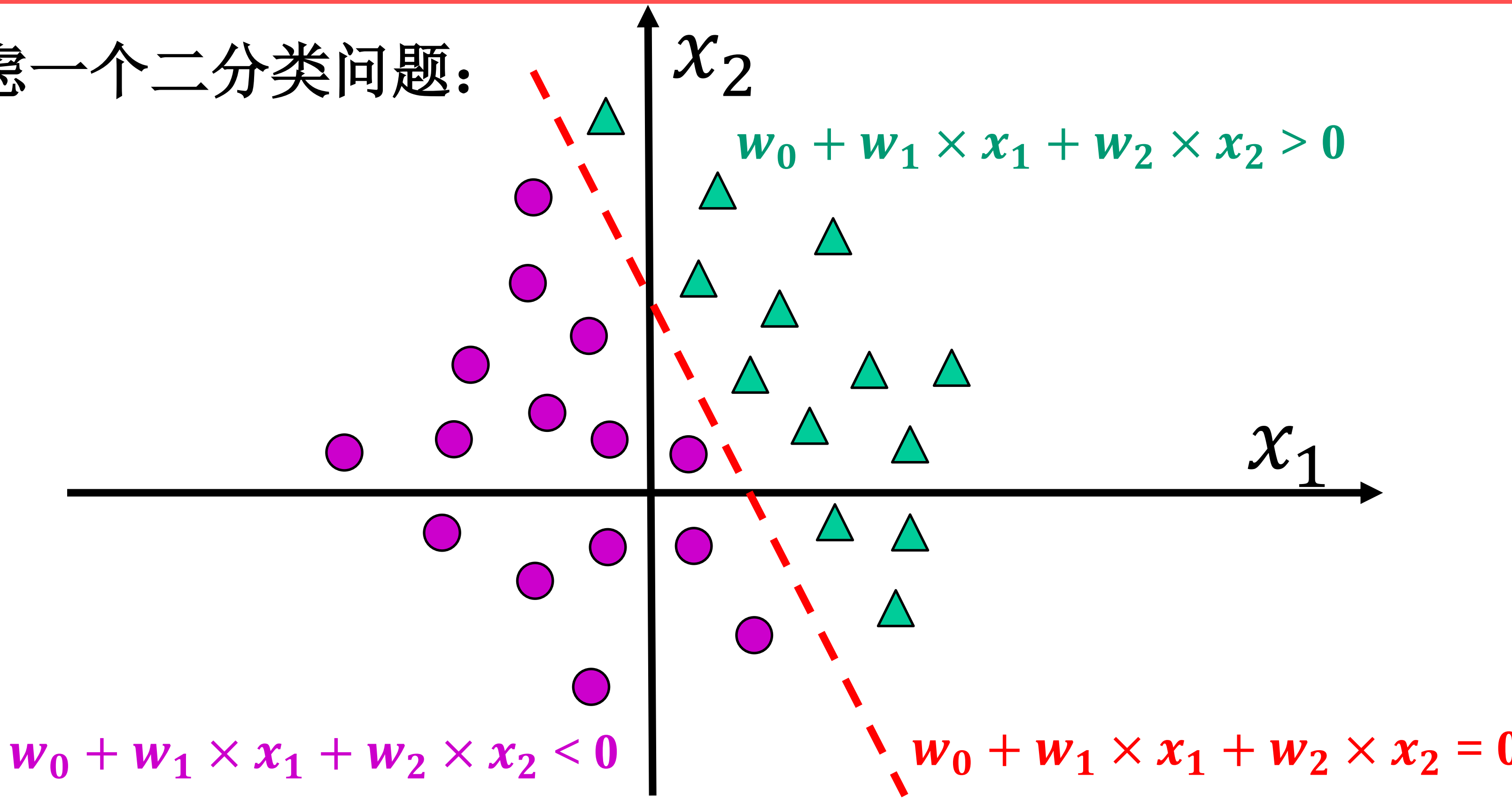
Classification



我们能否利用线性回归模型实现分类任务（二分类）？

分类

考虑一个二分类问题：



直观上说，可以让直线上方的点为正类 (Positive)，直线下方的点为负类 (Negative) 。

$$f(x) = w_0 + w_1 \times x_1 + w_2 \times x_2 = \hat{w}^T \hat{x}$$

回归与分类

显然有基本分类思想：

当$\boldsymbol{x}$为正类样本，$\boldsymbol{f(x) > 0}$，反之，则$\boldsymbol{x}$为负类样本，$\boldsymbol{f(x) < 0}$ 。

虽然现在我们有线性分类器$\boldsymbol{f(x)}$，但无法度量其好坏! Why?

分类器输出与标签无法对应！标签无法应用纠正错误的分类！

分类输出是连续，数值范围$[-\infty, \infty]$，

而样本标签$\boldsymbol{y}$是离散值，如$\{\mathbf{1}, \mathbf{0}\}$、$\{\mathbf{1}, -\mathbf{1}\}$.

So, How to handle this problem?

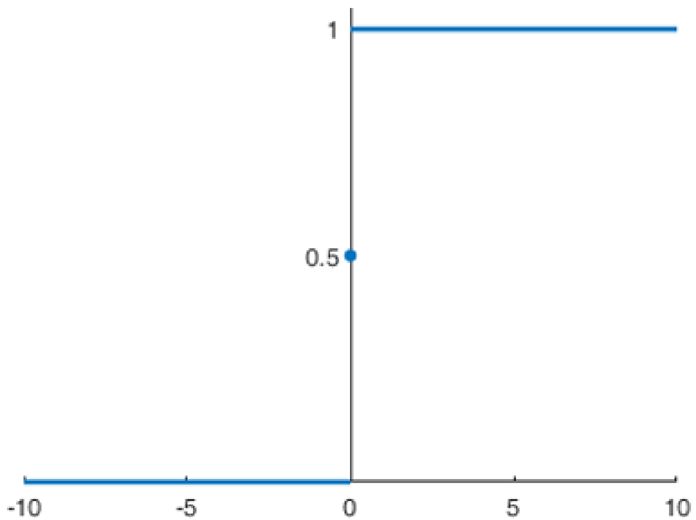
利用一个联系函数（link function）解决呗！(回顾广义线性回归模型)

回归与分类

最直观方式，直接选择单位阶跃函数作为联系函数把连续值转化为离散值(假设正负类标签为$\{1,0\}$)。

单位阶跃函数 (Unit-step function):

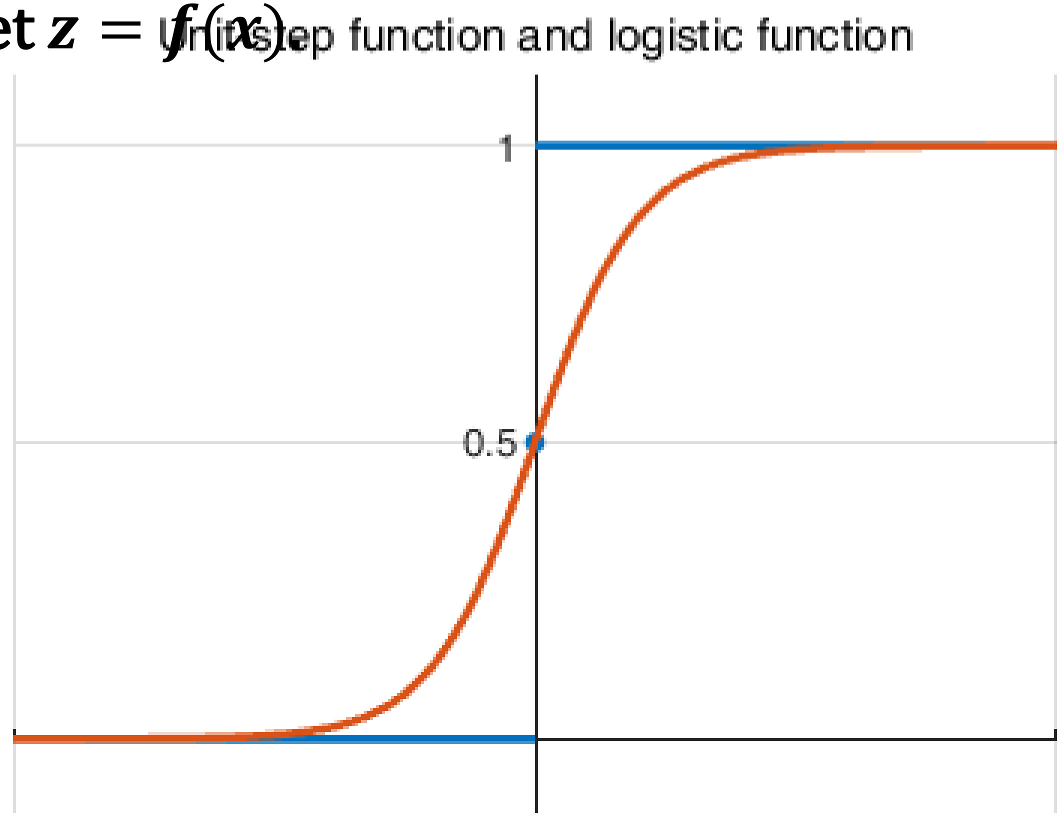
$$y = g(f(x)) = \begin{cases} 0, & f(x) < 0 \\ 0.5 & f(x) = 0 \\ 1 & f(x) > 0 \end{cases}$$



然而该函数在0点是不连续的，不存在对应反函数g^{-}，故无法作为联系函数。

对数几率函数（Logistic function）

Hopefully，我们寻找到一个和单位跃阶函数性质相似的连续函数。 Let $z = f(x)$.



Logistic function:

$$y = \frac{1}{1 + e^{-z}}$$

Logit (log odds):

$$z = \ln \frac{y}{1 - y}$$

分类损失函数（Loss function）

Logistic函数性质优良， 为任意阶可导函数，其的一阶导数为：

$$g' = \frac{1}{(1+e^{-z})^2} e^{-z} = \frac{1}{1+e^{-z}} \frac{e^{-z}}{1+e^{-z}} = g(1-g)$$

遵从线性回归的均方损失(均方误差)思想，引入logistic函数作为联系函数，可以定义以下分类损失函数：

$$L(w) = \frac{1}{n}\sum_{i=1}^{n}\|y_i - g(f(x))\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}\left\|y_i - \frac{1}{1+e^{-(\hat{w}^T\hat{x}_i)}}\right\|_2^2$$

分类损失函数（Loss Function）

根据上述推理，可推出“我们自己定义”的逻辑回归模型:

$$\min_{w} L(w) := \frac{1}{n}\sum_{i=1}^{n}\left\| y_i - \cancel{\frac{1}{1+e^{-(\hat{w}^T\hat{x}_i)}}} \right\|_2^2 \qquad (1)$$

就这样结束了？

故事还得继续！因为这个模型根本不work！

① 该模型是个非凸问题，存在多个局部最优解，求解全局最优十分困难，甚至没有全局最优解。

② 类别标签本身就是符号，在数值上没有任何意义！

另寻他解（Find another solution）

观察$f(x)$与logistic函数$g(f(x))$：

① $f(x)$的值有一定的物理意义。

② 单调性一致

③ $g(f(x))$输出范围为$[0,1]$，而概率取值范围也为$[0,1]$。

根据上述特性，不妨大胆假设 $g(f(x))$为样本x属于正类样概率，那么$1-g(f(x))$为样本x属于负类的概率。

① 模型(1)直接让$g(f(x))$显性关联标签 $\tilde{y}=g(f(x))$。

② 新模型通过概率的形式让$g(f(x))$隐性关联标签 $p(y=1|x)=g(f(x))$。

概率角度

- 利用极大似然的思想构建目标函数：
  - 对于负类样本应最大化概率：

$$\boldsymbol{p_0} = \boldsymbol{p}(\boldsymbol{y} = \mathbf{0}|\boldsymbol{x}) = \mathbf{1} - \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x})) = \frac{e^{-w^{\mathrm{T}}x}}{1+e^{-w^{\mathrm{T}}x}} = \frac{1}{1+e^{w^{\mathrm{T}}x}}$$

  - 对于正类样本应最大化概率：

$$\boldsymbol{p_1} = \boldsymbol{p}(\boldsymbol{y} = \mathbf{1}|\boldsymbol{x}) = \boldsymbol{g}\big(\boldsymbol{f}(\boldsymbol{x})\big) = \frac{\mathbf{1}}{\mathbf{1} + e^{-w^{\mathrm{T}}x}} = \frac{e^{w^{\mathrm{T}}x}}{\mathbf{1} + e^{w^{\mathrm{T}}x}}$$

  - 已知标签需最大化概率通式：

$$\boldsymbol{p}(\boldsymbol{y_i}|\boldsymbol{x_i};\boldsymbol{w}) = \boldsymbol{p_0}(\boldsymbol{x_i};\boldsymbol{w})^{(1-y_i)}\boldsymbol{p_1}(\boldsymbol{x_i};\boldsymbol{w})^{y_i}, \boldsymbol{y_i} \in \{\mathbf{1}, \mathbf{0}\}$$

目标函数

- 独立同分布假设下可得关于训练样本关于标签的联合概率函数：

$$P(\{y_i\}_{\mathrm{i}=1}^{n}|\{x_i\}_{\mathrm{i}=1}^{n};w) = \prod_{i=1}^{n} \boldsymbol{p}(\boldsymbol{y_i}|\boldsymbol{x_i};\boldsymbol{w})$$

- 考虑联合概率函数的对数似然函数作为目标函数：

$$\Theta(w) = \ln P(\{y_i\}_{\mathrm{i}=1}^{n}|\{x_i\}_{\mathrm{i}=1}^{n};w) = \sum_{i=1}^{n} \ln \boldsymbol{p}(\boldsymbol{y_i}|\boldsymbol{x_i};\boldsymbol{w})$$

- 原子项化简：

$$\ln \boldsymbol{p}(\boldsymbol{y_i}|\boldsymbol{x_i};\boldsymbol{w}) = (\mathbf{1} - y_i) \times \ln \boldsymbol{p_0}(\boldsymbol{x_i};\boldsymbol{w}) + y_i \times \ln \boldsymbol{p_1}(\boldsymbol{x_i};\boldsymbol{w})$$

$$= \boldsymbol{y_i}\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x_i} - \ln\left(\mathbf{1} + \boldsymbol{e}^{\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x_i}}\right)$$

极大似然法（ Maximum Likelihood ）

用极大似然法求解逻辑回归模型：

$$\underset{\widehat{\boldsymbol{w}}}{\operatorname{argmax}}\, \Theta(w) := \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} \boldsymbol{y}_{\boldsymbol{i}} \boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{x}_{\boldsymbol{i}} - \ln\left(\mathbf{1} + \boldsymbol{e}^{\boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{x}_{\boldsymbol{i}}}\right)$$

该模型等价于：

$$\underset{\widehat{\boldsymbol{w}}}{\operatorname{argmin}} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} \ln\left(\mathbf{1} + \boldsymbol{e}^{\boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{x}_{\boldsymbol{i}}}\right) - \boldsymbol{y}_{\boldsymbol{i}} \boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{x}_{\boldsymbol{i}}$$

对数几率模型（Logistic Regression）

求解无约束优化问题：

$$\underset{\widehat{\boldsymbol{w}}}{\operatorname{argmin}} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} \ln\left(\mathbf{1}+\boldsymbol{e}^{\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}_{\boldsymbol{i}}}\right)-\boldsymbol{y}_{\boldsymbol{i}}\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}_{\boldsymbol{i}}$$

数值方法 I： 牛顿法（Newton’s method）

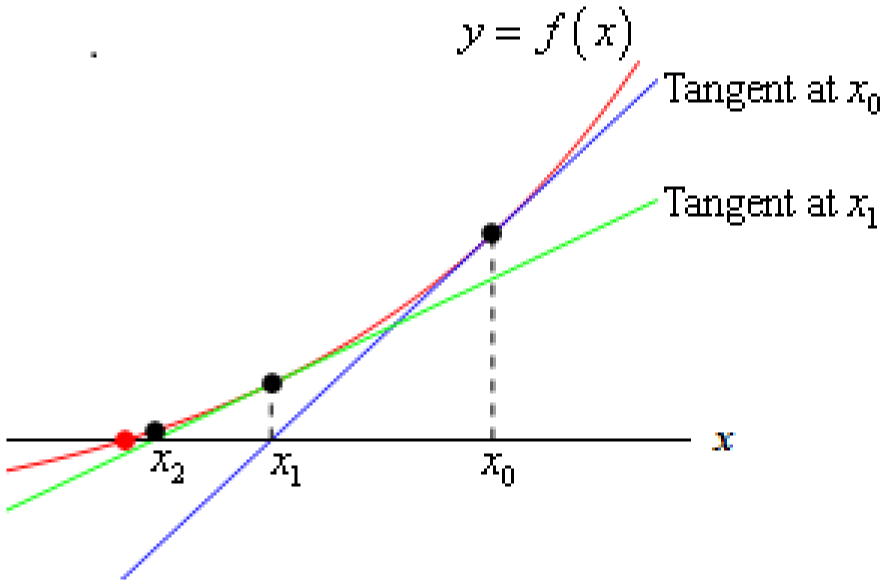
数值方法 II：梯度下降法（Gradient Decent Method）

牛顿法（Newton's Method）

方程求根问题：

$$f(x) = 0$$

$$f'(x_n) = \frac{f(x_n) - f(x_{n+1})}{x_n - x_{n+1}}$$



$$x_{n+1} = x_n - \frac{f(x_n) - f(x_{n+1})}{f'(x_n)} = x_n - \frac{f(x_n)}{f'(x_n)}$$

牛顿法（Newton’s Method）

最小化问题：

$$\boldsymbol{x}^* = \operatorname*{argmin}_{\boldsymbol{x}} \ell(\boldsymbol{x}) \quad \Longleftrightarrow \quad \ell'(\boldsymbol{x}) = \boldsymbol{0}$$

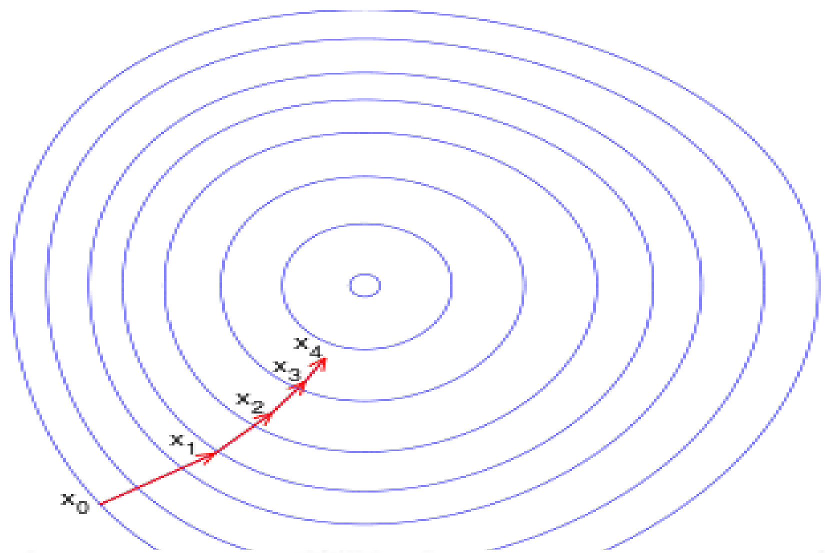
令 $\boldsymbol{f}(\boldsymbol{x}) = \ell'(\boldsymbol{x})$，则 $\boldsymbol{f}'(\boldsymbol{x}) = \ell''(\boldsymbol{x})$

$$\boxed{\boldsymbol{x}_{n+1} = \boldsymbol{x}_n - \frac{\boldsymbol{f}(\boldsymbol{x}_n)}{\boldsymbol{f}'(\boldsymbol{x}_n)} = \boldsymbol{x}_n - \frac{\ell'(\boldsymbol{x}_n)}{\ell''(\boldsymbol{x}_n)}}$$

梯度下降法

基本思想：

$$f(x_{n+1}) - f(x_n) = (x_{n+1}-x_n)^{\mathrm{T}}\nabla f(x_n) = -\gamma\nabla^2 f(x_n) < 0$$



$$f(x) = \frac{f(x_0)}{0!} + \frac{f'(x_0)}{1!}(x-x_0) + \frac{f''(x_0)}{2!}(x-x_0)^2 + \dots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n + R_n(x)$$

梯度下降法

迭代公式：

$$\boxed{\boldsymbol{x}_{n+1} = \boldsymbol{x}_n - \gamma_n \nabla f(\boldsymbol{x}_n)}$$

其中 γ_n 是第n步下降时选取的步长。

线搜索（Barzilai-Borwein Step)：

$$\gamma_n = \frac{(\boldsymbol{x}_n - \boldsymbol{x}_{n-1})^{\mathrm{T}}(\nabla f(\boldsymbol{x}_n) - \nabla f(\boldsymbol{x}_{n-1}))}{\|\nabla f(\boldsymbol{x}_n) - \nabla f(\boldsymbol{x}_{n-1})\|^2}$$

算法总结

牛顿法

$$x_{n+1} = x_n - \frac{\ell'(x_n)}{\ell''(x_n)}$$

梯度下降法

$$x_{n+1} = x_n - \gamma_n \ell'(x_n)$$

- 牛顿法和梯度下降法是求解最优化问题的常见的两种算法。
- 前者使用割线逐渐逼近最优解，后者使得目标函数逐渐下降。
- 牛顿法的收敛速度快，但是需要二阶导数信息。
- 梯度下降法计算速度快，但是需要人工确认步长参数。

极大似然法（ Maximum Likelihood ）

$$L = \ln p(y_i|x_i; w) = y_i \ln p_1 + (1 - y_i) \ln p_0$$

利用对数几率函数的性质 $p_1{}' = p_1(1 - p_1)$，可以得到目标函数的导数信息。

$$L' = y_i \frac{1}{p_1} p_1(1 - p_1)x_i + (1 - y_i) \frac{1}{1-p_1} (-p_1(1 - p_1))x_i = x_i(y_i - p_1)$$

$$L'' = (x_i(y_i - p_1))' = -x_i p_1(1 - p_1) x_i{}^{\mathrm{T}}$$

假设采用牛顿法：${}_{i+1}w = {}_{i}w - \frac{L'}{L''}$ 迭代更新直至目标函数收敛。