逻辑回归

陈飞宇

fchen@cqu.edu.cn

办公室：软件学院529

前情提要

线性回归模型：

$$\underset{\boldsymbol{w}_0,\boldsymbol{w}_1}{\operatorname{argmin}} \ \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i}}(\boldsymbol{y}_{\boldsymbol{i}} - (\boldsymbol{w}_{\mathbf{0}} + \boldsymbol{w}_{\mathbf{1}}\boldsymbol{x}_{\boldsymbol{i}}))^{\mathbf{2}}$$

建模，计算，应用。

概率解释（极大似然法）：

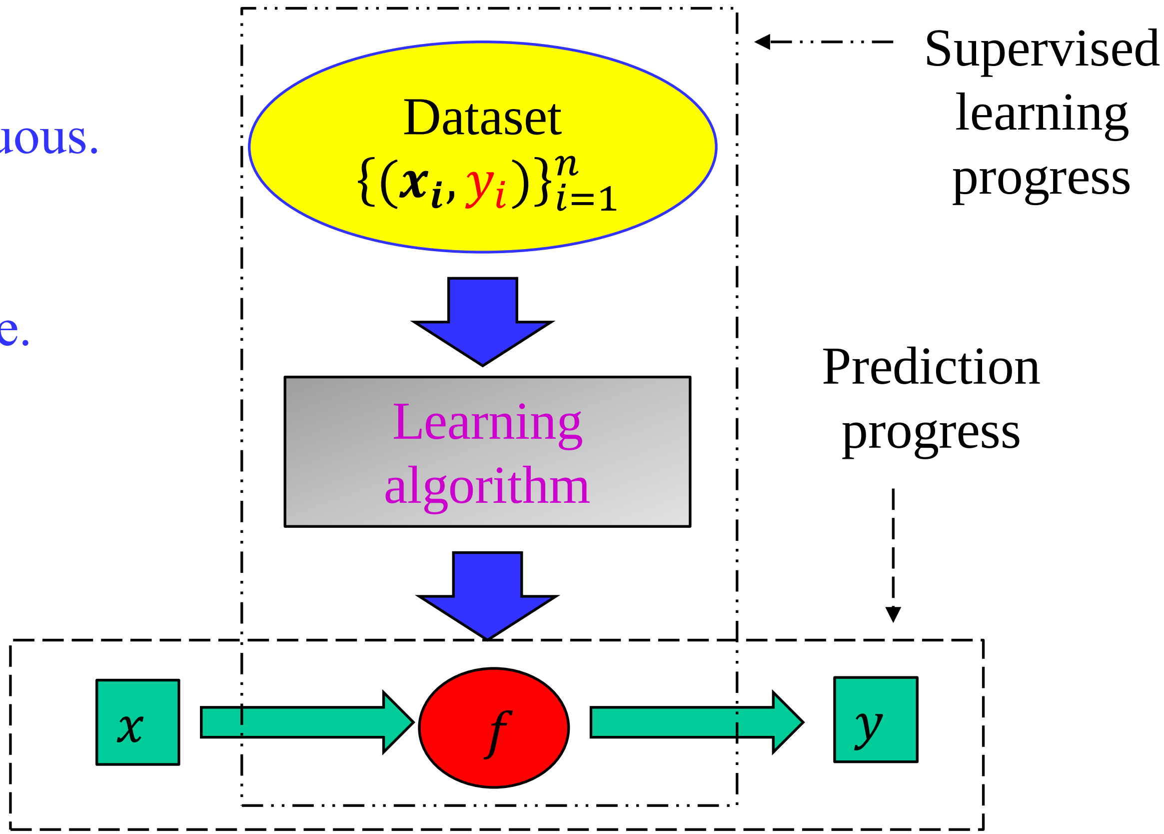
$$\underset{\widehat{\boldsymbol{w}}}{\operatorname{argmax}} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} \ln \boldsymbol{p}(\boldsymbol{y}_{\boldsymbol{i}} | \widehat{\boldsymbol{x}}_{\boldsymbol{i}}; \widehat{\boldsymbol{w}})$$

监督学习（ Supervised Learning ）

Learning with a teacher

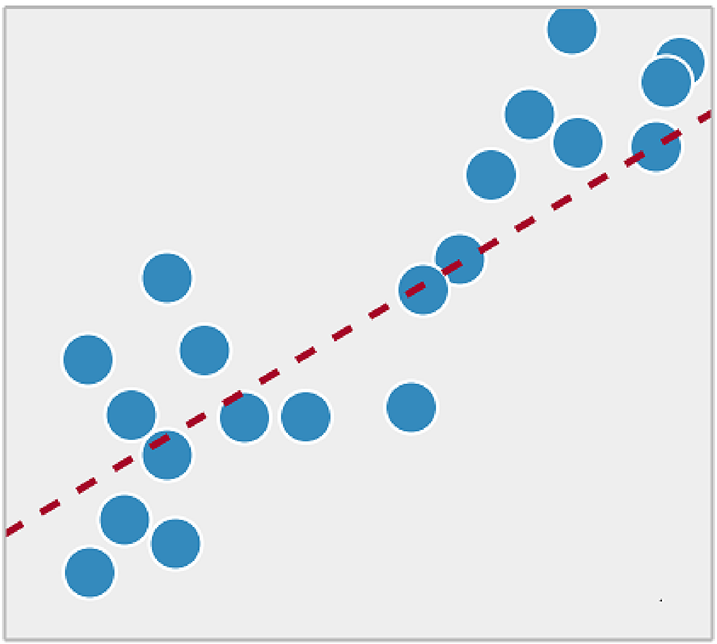
Regression:
y_i’s are continuous.

Classification:
y_i’s are discrete.

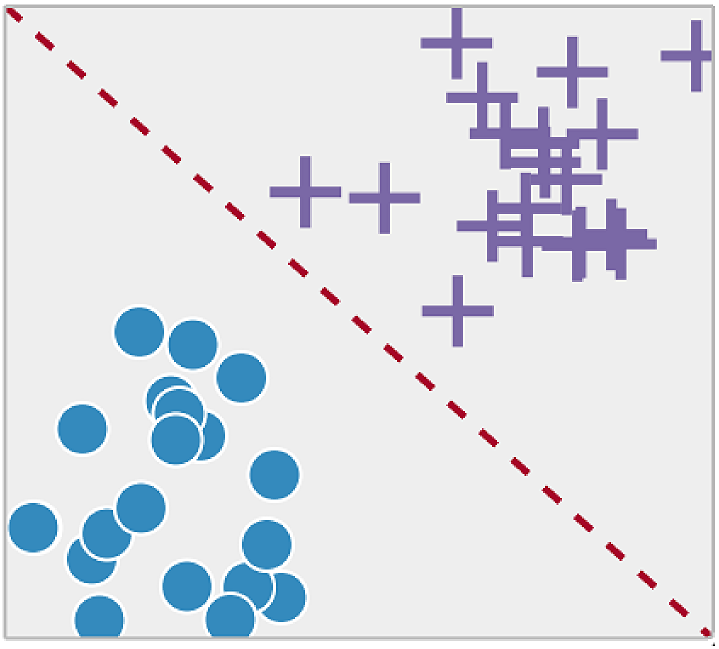


回归与分类

Regression



Classification



我们能否利用线性回归的思想解决分类任务（二分类）？

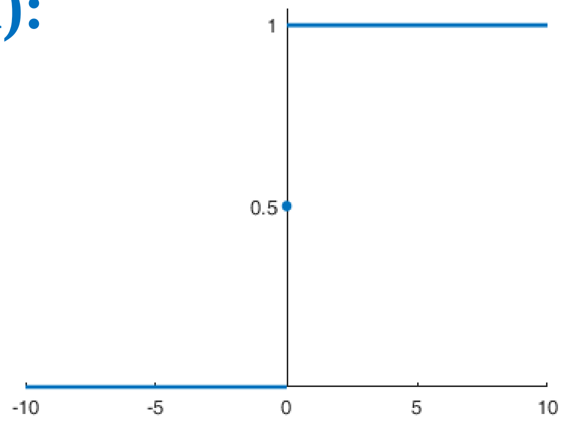
回归与分类

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

本质上说，需要把连续值转化为离散值(例如: $\{0, 1\}$)。

单位阶跃函数 (Unit-step function):

$$y = \begin{cases} 0, & z < 0 \\ 0.5 & z = 0 \\ 1 & z > 0 \end{cases}$$

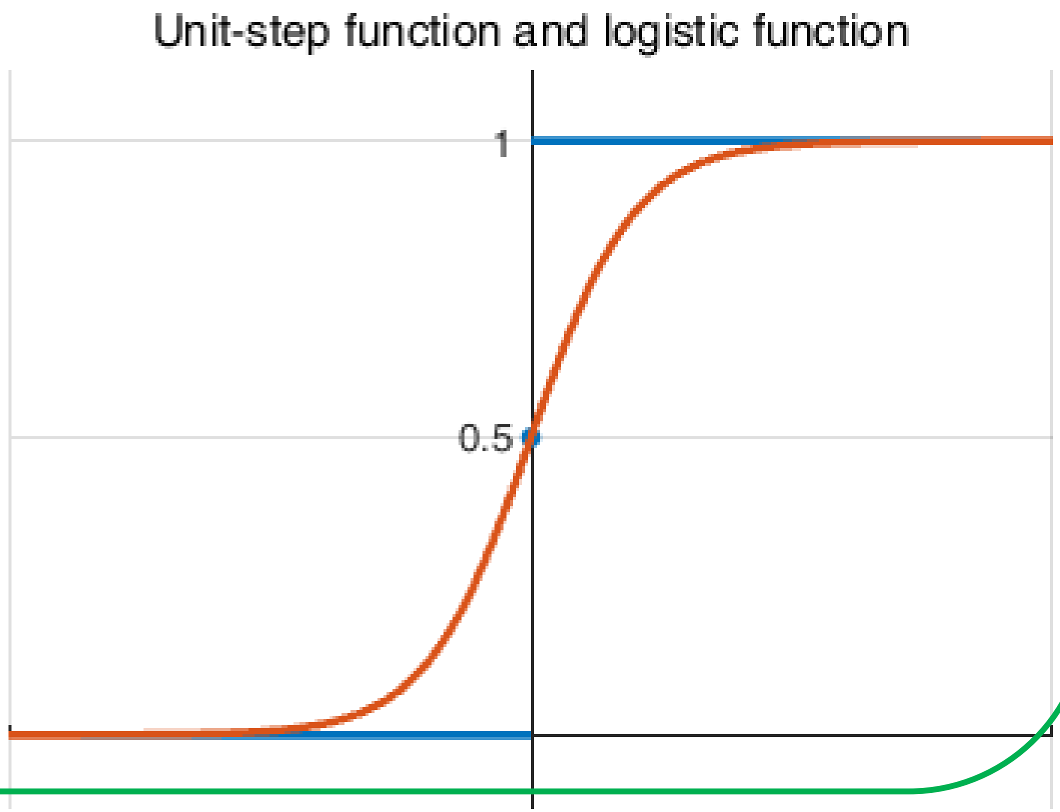


对数几率函数（Logistic function）

由于单位阶跃函数不是一个连续函数，我们通常选择一些性质好的函数作为替代函数。

Logistic function:

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



对数回归（ Logistic regression ）

结合线性回归的思想与对数几率函数的特点，我们得到对数回归模型（Logistic regression/ logit regression）：

$$y = g\left(w^{\mathrm{T}}x + b\right) = \frac{1}{1 + e^{-(w^{\mathrm{T}}x+b)}}$$

对数几率函数 g 为任意阶可导函数，它的一阶导数为：

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

对数回归（Logistic regression）

将 $\boldsymbol{y}$ 视为样本 $\boldsymbol{x}$ 作为正例的可能性，则 $\mathbf{1}-\boldsymbol{y}$ 是其为反例的可能性。

$$\boldsymbol{y}=\frac{\mathbf{1}}{\mathbf{1}+\boldsymbol{e}^{-(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{b})}} \Rightarrow \ln\frac{\boldsymbol{y}}{\mathbf{1}-\boldsymbol{y}}=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{b}$$

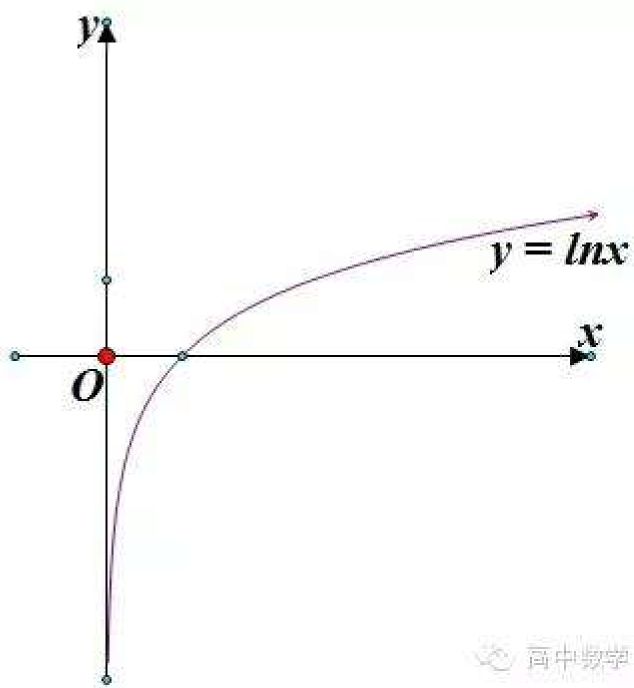
两者的比值：

$$\frac{\boldsymbol{y}}{\mathbf{1}-\boldsymbol{y}}$$

称为“几率”(Odds)， 反映了 $\boldsymbol{x}$ 作为正例的相对可能性。

对数回归（ Logistic regression ）

$$\ln \frac{y}{1-y} = w^{\mathrm{T}} x + b$$



说明：

如果 $w^{\mathrm{T}}x + b > 0$, 则 $\frac{y}{1-y} > 1$。意味着 x 作为正例的可能性大于其为反例的可能性。

反之，如果 $w^{\mathrm{T}}x + b < 0$, 则 $\frac{y}{1-y} < 1$。意味着 x 作为正例的可能性小于其为反例的可能性。

逻辑回归的本质是用线性回归的预测结果去逼近真实标记的对数几率。

对数回归（Logistic regression）

下面我们将 $\boldsymbol{y}$ 写成后验概率估计 $\boldsymbol{p}(\boldsymbol{y}=\mathbf{1}|\boldsymbol{x})$，则 $\mathbf{1}-\boldsymbol{y}$ 可写成后验概率估计 $\boldsymbol{p}(\boldsymbol{y}=\mathbf{0}|\boldsymbol{x})$，我们有

$$\ln\frac{\boldsymbol{y}}{\mathbf{1}-\boldsymbol{y}}=\ln\frac{\boldsymbol{p}(\boldsymbol{y}=\mathbf{1}|\boldsymbol{x})}{\boldsymbol{p}(\boldsymbol{y}=\mathbf{0}|\boldsymbol{x})}=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{b}$$

显然有，

$$\boldsymbol{p}(\boldsymbol{y}=\mathbf{1}|\boldsymbol{x})=\frac{\boldsymbol{e}^{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{b}}}{\mathbf{1}+\boldsymbol{e}^{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{b}}}$$

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

极大似然法（ Maximum Likelihood ）

考虑用极大似然法求解逻辑回归模型：

$$\underset{\widehat{\boldsymbol{w}}}{\operatorname{argmax}} \sum_{i=1}^{n} \ln \boldsymbol{p}(\boldsymbol{y}_i | \widehat{\boldsymbol{x}}_i; \widehat{\boldsymbol{w}})$$

对于二分类问题，输出变量 $\boldsymbol{y}_i \in \{\mathbf{0}, \mathbf{1}\}$。

$$\boldsymbol{p}(\boldsymbol{y}_i | \widehat{\boldsymbol{x}}_i; \widehat{\boldsymbol{w}}) = (\boldsymbol{p}_1(\widehat{\boldsymbol{x}}_i; \widehat{\boldsymbol{w}}))^{\boldsymbol{y}_i} (\boldsymbol{p}_0(\widehat{\boldsymbol{x}}_i; \widehat{\boldsymbol{w}}))^{(1-\boldsymbol{y}_i)}$$

其中 $\boldsymbol{p}_1(\widehat{\boldsymbol{x}}_i; \widehat{\boldsymbol{w}}) = \boldsymbol{p}(\boldsymbol{y} = \mathbf{1} | \widehat{\boldsymbol{x}}_i; \widehat{\boldsymbol{w}})$, $\quad \boldsymbol{p}_0(\widehat{\boldsymbol{x}}_i; \widehat{\boldsymbol{w}}) = \boldsymbol{p}(\boldsymbol{y} = \mathbf{0} | \widehat{\boldsymbol{x}}_i; \widehat{\boldsymbol{w}})$

$$\ln \boldsymbol{p}(\boldsymbol{y}_i | \widehat{\boldsymbol{x}}_i; \widehat{\boldsymbol{w}}) = \boldsymbol{y}_i \ln \boldsymbol{p}_1(\widehat{\boldsymbol{x}}_i; \widehat{\boldsymbol{w}}) + (\mathbf{1} - \boldsymbol{y}_i) \ln \boldsymbol{p}_0(\widehat{\boldsymbol{x}}_i; \widehat{\boldsymbol{w}})$$

极大似然法（ Maximum Likelihood ）

$$\underset{\widehat{\boldsymbol{w}}}{\operatorname{argmax}} \sum_{\boldsymbol{i}=1}^{\boldsymbol{n}} (\boldsymbol{y}_{\boldsymbol{i}} \ln \boldsymbol{p}(\boldsymbol{y} = \mathbf{1}|\widehat{\boldsymbol{x}}_{\boldsymbol{i}}; \widehat{\boldsymbol{w}}) + (\mathbf{1} - \boldsymbol{y}_{\boldsymbol{i}}) \ln \boldsymbol{p}(\boldsymbol{y} = \mathbf{0}|\widehat{\boldsymbol{x}}_{\boldsymbol{i}}; \widehat{\boldsymbol{w}}))$$

说明：

当 $\boldsymbol{y}_{\boldsymbol{i}} = \mathbf{1}$，即 $\boldsymbol{x}_{\boldsymbol{i}}$ 是正样本时，我们希望由参数 $\widehat{\boldsymbol{w}}$ 确定的直线方程可以使得 $\boldsymbol{p}(\boldsymbol{y} = \mathbf{1}|\widehat{\boldsymbol{x}}_{\boldsymbol{i}}; \widehat{\boldsymbol{w}})$ 最大化，此时 $\widehat{\boldsymbol{w}}^{\mathrm{T}}\widehat{\boldsymbol{x}}_{\boldsymbol{i}}$ 的值越大越好。

当 $\boldsymbol{y}_{\boldsymbol{i}} = \mathbf{0}$，即 $\boldsymbol{x}_{\boldsymbol{i}}$ 是负样本时，我们希望由参数 $\widehat{\boldsymbol{w}}$ 确定的直线方程可以使得 $\boldsymbol{p}(\boldsymbol{y} = \mathbf{0}|\widehat{\boldsymbol{x}}_{\boldsymbol{i}}; \widehat{\boldsymbol{w}})$ 最大化，此时 $\widehat{\boldsymbol{w}}^{\mathrm{T}}\widehat{\boldsymbol{x}}_{\boldsymbol{i}}$ 的值越小越好。

极大似然法（ Maximum Likelihood ）

分别把 $\boldsymbol{p}_i$ 的表达式代入 $\ln \boldsymbol{p}_i$，得到：

$$\ln \boldsymbol{p}_1 = \ln\left(\frac{e^{\widehat{\boldsymbol{w}}^{\mathrm{T}}\widehat{\boldsymbol{x}_i}}}{1+e^{\widehat{\boldsymbol{w}}^{\mathrm{T}}\widehat{\boldsymbol{x}_i}}}\right) = \widehat{\boldsymbol{w}}^{\mathrm{T}}\widehat{\boldsymbol{x}}_i - \ln\left(\mathbf{1} + e^{\widehat{\boldsymbol{w}}^{\mathrm{T}}\widehat{\boldsymbol{x}}_i}\right)$$

$$\ln \boldsymbol{p}_0 = \ln\left(\frac{\mathbf{1}}{1+e^{\widehat{\boldsymbol{w}}^{\mathrm{T}}\widehat{\boldsymbol{x}_i}}}\right) = -\ln\left(\mathbf{1} + e^{\widehat{\boldsymbol{w}}^{\mathrm{T}}\widehat{\boldsymbol{x}}_i}\right)$$

特别地，我们可以得到$\mathbf{ln}\,\boldsymbol{p}(\boldsymbol{y}_i|\widehat{\boldsymbol{x}}_i;\widehat{\boldsymbol{w}})$的表达式

$$\mathbf{ln}\,\boldsymbol{p} = \boldsymbol{y}_i \ln \boldsymbol{p}_1 + (\mathbf{1} - \boldsymbol{y}_i)\ln \boldsymbol{p}_0 = \boldsymbol{y}_i\widehat{\boldsymbol{w}}^{T}\widehat{\boldsymbol{x}}_i - \mathbf{ln}\left(\mathbf{1} + e^{\widehat{\boldsymbol{w}}^{T}\widehat{\boldsymbol{x}}_i}\right)$$

对数几率模型（Logistic Regression）

求解无约束优化问题：

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

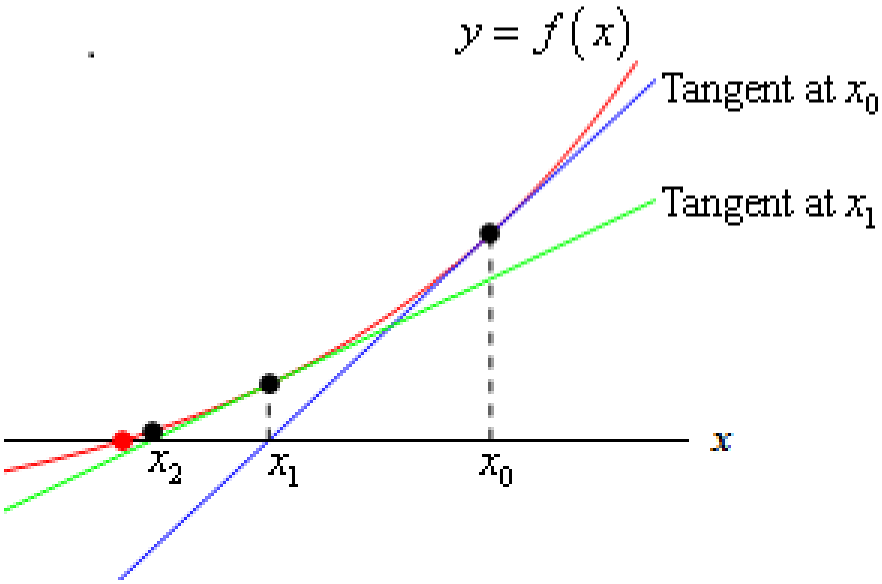
数值方法 I： 牛顿法（Newton’s method）

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：

$$f(x) = 0$$



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

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

牛顿法（Newton’s Method）

最小化问题：

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

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

迭代公式：

$$\boxed{x_{n+1} = x_n - \gamma_n \nabla f(x_n)}$$

其中 γ_n 是第n步下降时选取的步长，也称学习率。

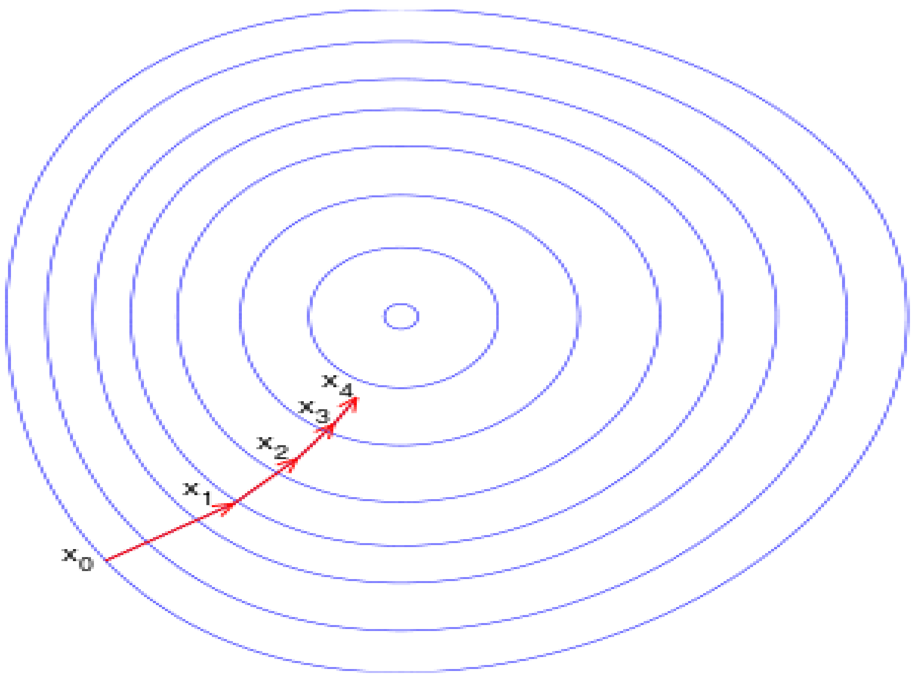
线搜索（Barzilai-Borwein Step）：

$$\gamma_n = \frac{(x_n - x_{n-1})^{\mathrm{T}}(\nabla f(x_n) - \nabla f(x_{n-1}))}{\|\nabla f(x_n) - \nabla f(x_{n-1})\|^2}$$

梯度下降法

基本思想：

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



算法总结

牛顿法

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

梯度下降法

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

- 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。
- BB步实际上借助了二阶导数信息(x_{n-1}和x_n的梯度的差)。

极大似然法（ Maximum Likelihood ）

$$\underset{\widehat{\boldsymbol{w}}}{\operatorname{argmin}}\ -\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}}(\boldsymbol{y}_{\boldsymbol{i}} \ln \boldsymbol{p}_{\mathbf{1}} + (\mathbf{1}-\boldsymbol{y}_{\boldsymbol{i}}) \ln \boldsymbol{p}_{\mathbf{0}})$$

利用对数几率函数的性质 $\boldsymbol{p}_{\mathbf{1}}{}' = \boldsymbol{p}_{\mathbf{1}}(\mathbf{1}-\boldsymbol{p}_{\mathbf{1}})$，可以得到目标函数的导数。

$$(\ln \boldsymbol{p}(\boldsymbol{y}_{\boldsymbol{i}}|\widehat{\boldsymbol{x}}_{\boldsymbol{i}};\widehat{\boldsymbol{w}}))' = \boldsymbol{y}_{\boldsymbol{i}}\frac{\mathbf{1}}{\boldsymbol{p}_{\mathbf{1}}}\boldsymbol{p}_{\mathbf{1}}(\mathbf{1}-\boldsymbol{p}_{\mathbf{1}})\widehat{\boldsymbol{x}}_{\boldsymbol{i}} + (\mathbf{1}-\boldsymbol{y}_{\boldsymbol{i}})\frac{\mathbf{1}}{\mathbf{1}-\boldsymbol{p}_{\mathbf{1}}}(-\boldsymbol{p}_{\mathbf{1}}(\mathbf{1}-\boldsymbol{p}_{\mathbf{1}}))\widehat{\boldsymbol{x}}_{\boldsymbol{i}}$$

$$= \widehat{\boldsymbol{x}}_{\boldsymbol{i}}(\boldsymbol{y}_{\boldsymbol{i}}-\boldsymbol{p}_{\mathbf{1}})$$

$$(\ln \boldsymbol{p}(\boldsymbol{y}_{\boldsymbol{i}}|\widehat{\boldsymbol{x}}_{\boldsymbol{i}};\widehat{\boldsymbol{w}}))'' = (\widehat{\boldsymbol{x}}_{\boldsymbol{i}}(\boldsymbol{y}_{\boldsymbol{i}}-\boldsymbol{p}_{\mathbf{1}}))' = -\widehat{\boldsymbol{x}}_{\boldsymbol{i}}\boldsymbol{p}_{\mathbf{1}}(\mathbf{1}-\boldsymbol{p}_{\mathbf{1}})\widehat{\boldsymbol{x}}_{\boldsymbol{i}}{}^{\mathbf{T}}$$