深蓝学院
shenlanxueyuan.com

深度学习：从理论到实践

第一章：深度学习基础（下）

深度学习基础（二）

- ✓ 回归与分类
- ✓ 梯度下降
- ✓ 信息论

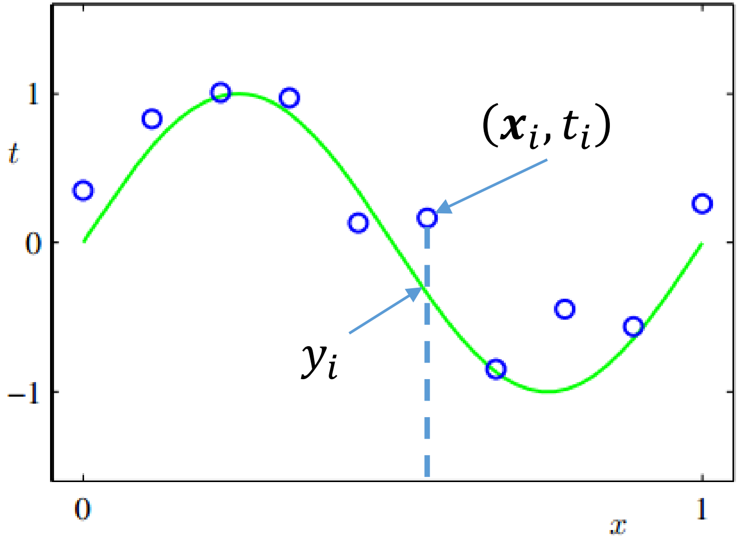
回归与分类

- ✓ 曲线拟合
- ✓ 线性回归
- ✓ logistic回归
- ✓ softmax回归

曲线拟合

- ✓ 回归基本问题：已知样本集D的n个样本$(\boldsymbol{x}_i, t_i)_{i=1}^{n}$，望获知自变量$\boldsymbol{x}$与变量$y$的映射关系$f$。
- ✓ 一个例子：

真实曲线$\sin(2\pi x)$，
n=10个训练样本。
根据样本点拟合曲线。



曲线拟合

- 多项式函数拟合

$$g(x,\mathbf{w}) = \sum_{j=0}^{M} w_j x^j = w_0 + w_1 x_1 + \cdots + w_M x^M$$

其中多项式的阶、即最高次数M待选择。

- 最小二乘拟合损失

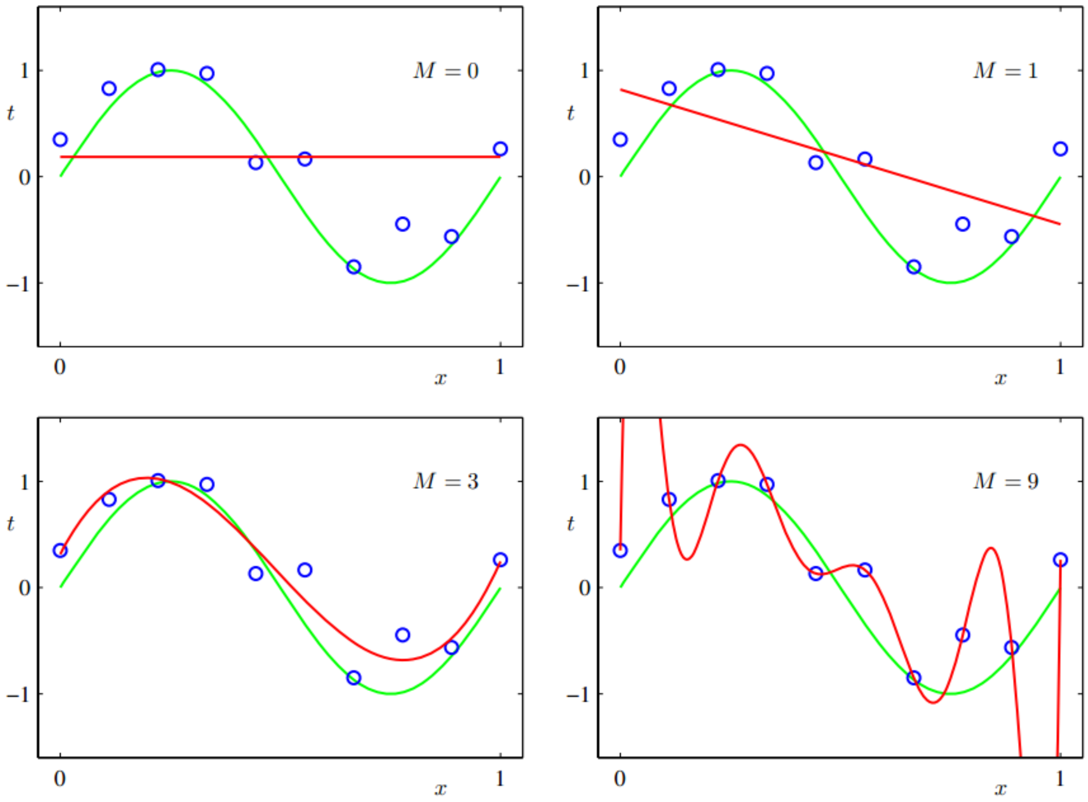
$$E(w) = \frac{1}{2}\sum_{n=1}^{N}\{\boxed{g(x_n,\mathbf{w})} - \boxed{t_n}\}^2$$

回归值　样本值

曲线拟合

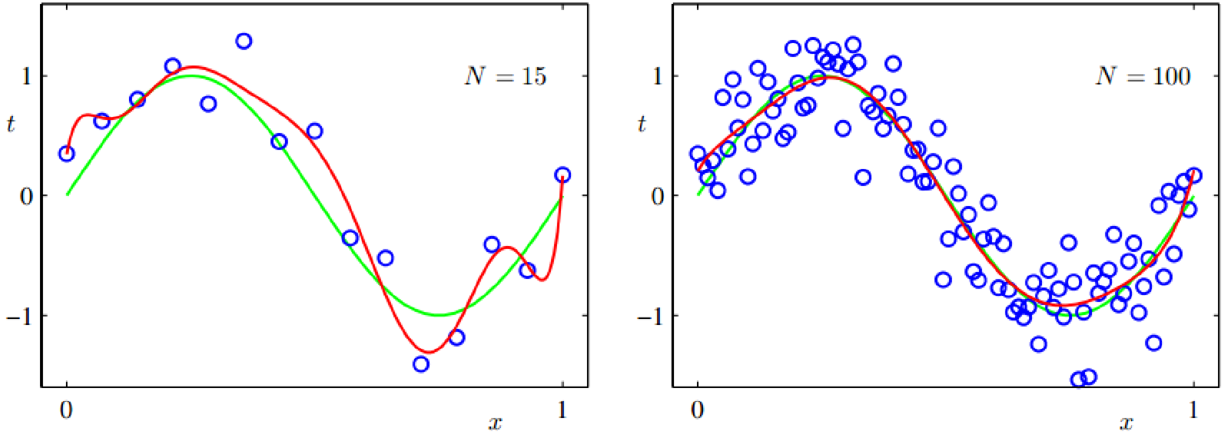
✓ 不同阶多项式拟合的结果

高次多项式精确拟合，预测性能却极差，产生过拟合。



回避过拟合的两类办法

- ✓ 获取更多的训练样本



- ✓ 获取关于模式的先验知识，例如$A\sin(\omega x + t)$

曲线拟合优化方法

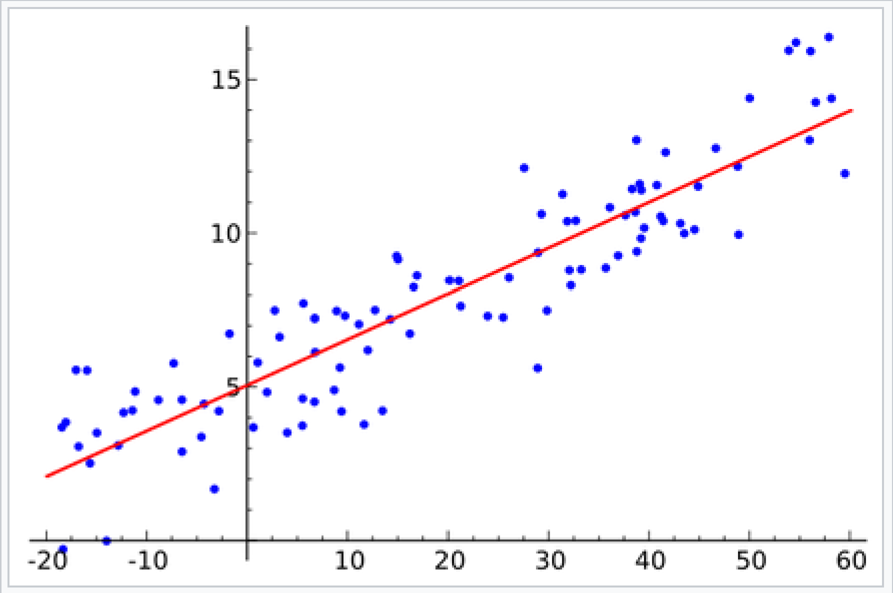
- ✓ 令$\boldsymbol{z}=[1,x,x^2,\ldots,x^M]^T$，拟合函数$g=\boldsymbol{z}^T\boldsymbol{w}$，损失函数$E(\boldsymbol{w})=\frac{1}{2}\sum_{n=1}^{N}\{\boldsymbol{z}_n^T\boldsymbol{w}-t_n\}^2$

- ✓ 可见，拟合函数转化为线性函数，即M维空间的线性回归。

特征变换

线性回归

✓ 线性回归：变量与自变量之间线性相关性，使用线性回归估计二者间的映射。



https://en.wikipedia.org/wiki/Linear_regression

线性回归

深蓝学院 shenlanxueyuan.com

- ✓ 线性回归问题

$$E(\boldsymbol{w}) = \min_{\boldsymbol{w}} \frac{1}{2}\sum_{n=1}^{N}\{\boldsymbol{a}_n^T\boldsymbol{w} - b_n\}^2 = \min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b}\|^2 = \min_{\boldsymbol{w}}\left(\frac{\boldsymbol{w}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{w}}{2} - \boldsymbol{w}^T\boldsymbol{A}^T\boldsymbol{b} + \frac{\boldsymbol{b}^T\boldsymbol{b}}{2}\right)$$

，其中$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{a}_1^T \\ \vdots \\ \boldsymbol{a}_N^T \end{pmatrix}$，$\boldsymbol{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_N \end{pmatrix}$。

- ✓ 最小二乘解：$\nabla_{\boldsymbol{w}}E(\boldsymbol{w}) = \boldsymbol{A}^T\boldsymbol{A}\boldsymbol{w} - \boldsymbol{A}^T\boldsymbol{b} = 0$

$$\boldsymbol{w}^* = \left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T\boldsymbol{b}$$

logistic回归

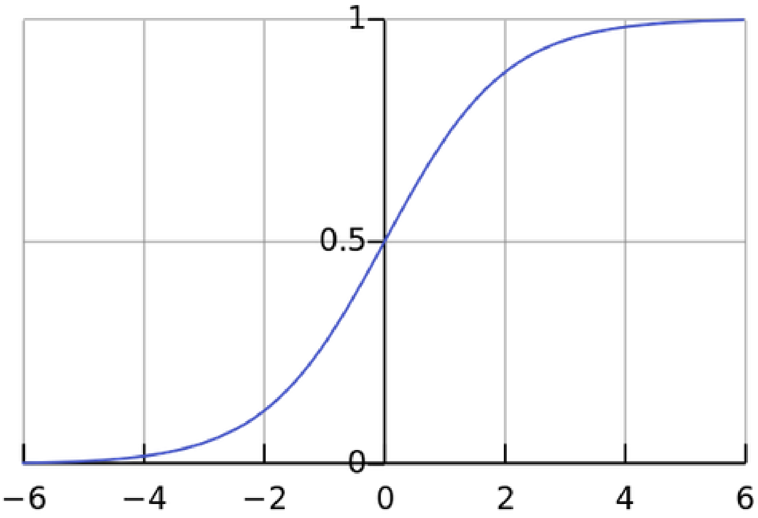
- 二分类问题，已知训练样本集D的n个样本$(\boldsymbol{x}_i, t_i)_{i=1}^{n}$，其中$t_i \in \{0,1\}$为类别标签，$\boldsymbol{x}_i \in \mathcal{R}^d$为特征向量。

- 拟合特征向量到类别标签的回归，常用logistic回归。

logistic函数

- ✓ 标准logistic函数

$$f(x) = \frac{1}{1 + exp(-x)}$$



- ✓ $f(-x) = 1 - f(x),$
- ✓ $f'(x) = f(x)\big(1 - f(x)\big)$

logistic回归

- logistic回归的回归函数为

$$g(\boldsymbol{x}) = \frac{1}{1 + exp(-\boldsymbol{w}^T\boldsymbol{x})},$$

其中$\boldsymbol{w}$为回归参数。

- logistic回归，样本$(\boldsymbol{x}_i, t_i)$的概率密度

$$P(\boldsymbol{x}_i, t_i; \boldsymbol{w}) = \begin{cases} g(\boldsymbol{x}_i), t_i = 1 \\ 1 - g(\boldsymbol{x}_i), t_i = 0 \end{cases} = g(\boldsymbol{x}_i)^{t_i} \cdot \left(1 - g(\boldsymbol{x}_i)\right)^{1-t_i}$$

logistic回归

- ✓ 极大似然估计$\boldsymbol{w}$
- ✓ 似然函数$p(D|\boldsymbol{w}) = \prod_{i=1}^{n} g(\boldsymbol{x}_i)^{t_i}\left(1 - g(\boldsymbol{x}_i)\right)^{1-t_i}$，取负对数似然作为代价函数有

$$\mathrm{E}(\mathbf{w}) = -\left[\sum_{i=1}^{n} t_i \ln g(\boldsymbol{x}_i) + (1 - t_i) \ln\left(1 - g(\boldsymbol{x}_i)\right)\right]$$

logistic回归:优化

- ✓ 该优化问题采用Newton-Raphson迭代优化，

$$\boldsymbol{w}^{new} = \boldsymbol{w}^{old} - \boldsymbol{H}^{-1}\nabla E(\boldsymbol{w}),$$

其中$\boldsymbol{H}$为$E(w)$关于w的二阶导数矩阵。

- ✓ $\nabla E(\boldsymbol{w}) = \sum_{i=1}^{n}(y_i - t_i)\,\boldsymbol{x}_i = \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{t})$
- ✓ $\boldsymbol{H} = \nabla\nabla E(\boldsymbol{w}) = \sum_{i=1}^{n} y_i(1 - y_i)\,\boldsymbol{x}_i\boldsymbol{x}_i^T = \boldsymbol{X}^T\boldsymbol{R}\boldsymbol{X}$

logistic回归：举例

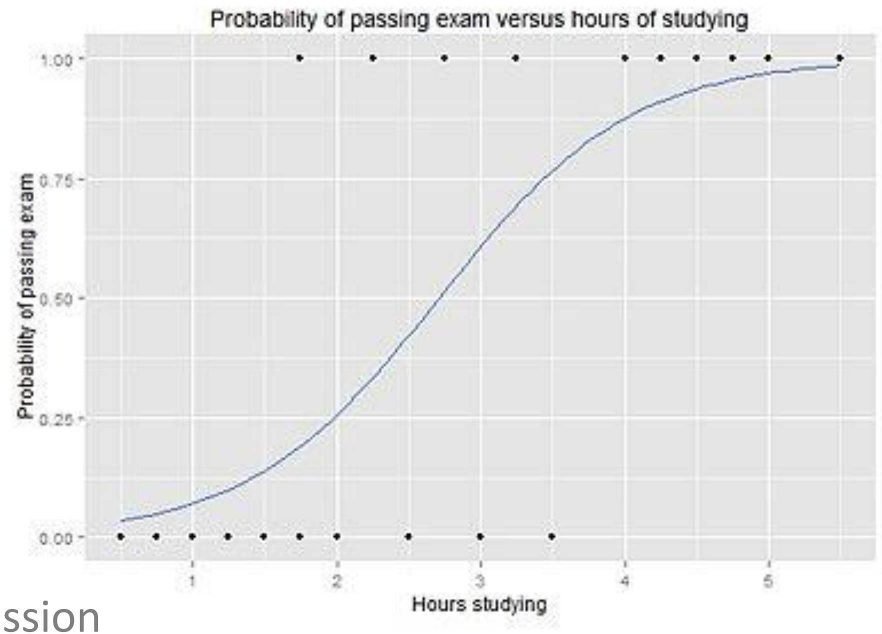
深蓝学院
shenlanxueyuan.com

✓ 一门考试，20位考生花费0~6小时备考。现在希望获悉备考时长与是否通过考试的关系。

0.50	0.75	1.00	1.25	1.50
0	0	0	0	0
1.75	1.75	2.00	2.25	2.50
0	1	0	1	0
2.75	3.00	3.25	3.50	4.00
1	0	1	0	1
4.25	4.50	4.75	5.00	5.50
1	1	1	1	1

logistic回归：举例

- ✓ 解释变量仅仅为1维的学习时间，回归参数为2维向量。
- ✓ 通过考试的概率为
- ✓ $\dfrac{1}{1+\exp\left(-\left(1.5046\cdot\text{时长}-4.0777\right)\right)}$



https://en.wikipedia.org/wiki/Logistic_regression

softmax回归

- ✓ softmax回归用于多类分类问题。
- ✓ K类问题中，不采用标签定义$1{:}\,K$，而使用标签向量：

$$
\boldsymbol{t} = \begin{array}{cc} \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} & \begin{array}{c} 1 \\ \\ k \\ \\ K \end{array} \end{array}
$$

$0-1$的K维向量，若属于k类，则向量的k分量为1，其他分量均为0。

softmax回归

- softmax回归的回归函数为

$$p(\mathcal{C}_k|\boldsymbol{x}) = \frac{\exp\left(\boldsymbol{w}_k^T\boldsymbol{x}\right)}{\sum_j \exp\left(\boldsymbol{w}_j^T\boldsymbol{x}\right)}$$

其中$\boldsymbol{w}_k$为第k类的回归参数。

- softmax回归，某个$\boldsymbol{x}_i$样本的概率：

$$\prod_{k=1}^{K} p(\mathcal{C}_k|\boldsymbol{x}_i)^{t_{ik}}$$

其中，$\boldsymbol{t}_i = (t_{i1}, \dots, t_{ik}, \dots, t_{iK})^T$为$\boldsymbol{x}$的标签向量

softmax回归

- ✓ 极大似然估计回归参数$\boldsymbol{w}_1, \dots, \boldsymbol{w}_K$

- ✓ 似然函数$p(D|\boldsymbol{w}_1, \dots, \boldsymbol{w}_K) = \prod_i \prod_{k=1}^{K} p(\mathcal{C}_k|\boldsymbol{x}_i)^{t_{ik}}$，取负对数似然

$$E(\boldsymbol{w}_1, \dots, \boldsymbol{w}_K) = -\sum_i \sum_{k=1}^{K} t_{ik} \ln p(\mathcal{C}_k|\boldsymbol{x}_i)$$

梯度下降(Gradient Descent)

深蓝学院
shenlanxueyuan.com

- ✓ 基本理论
- ✓ 初始化与步长
- ✓ 深度学习应用

梯度下降基本理论

深蓝学院
shenlanxueyuan.com

- 最小化问题

$$\min_{\boldsymbol{x}} f(\boldsymbol{x})$$

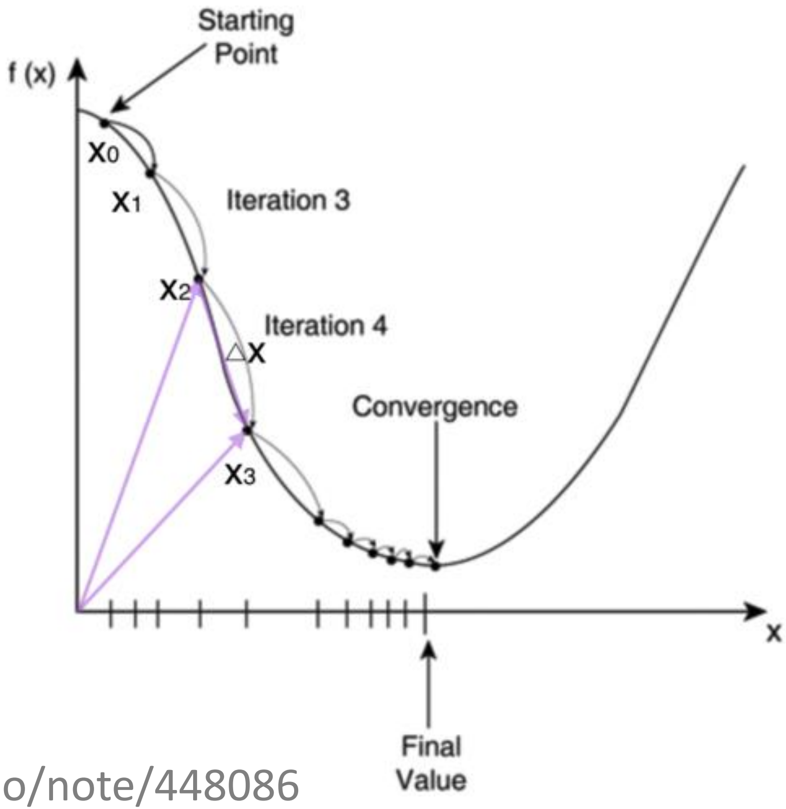
- 负梯度方向为函数值下降最快的方向

$$\boldsymbol{x}^{new} = \boldsymbol{x}^{old} - \lambda \nabla_{\boldsymbol{x}} f(\boldsymbol{x})$$

其中，$\nabla_{\boldsymbol{x}} f(\boldsymbol{x})$为函数在$\boldsymbol{x}$处的梯度，$\lambda$为步长、亦称学习率。

梯度下降基本理论

✓ 迭代更新

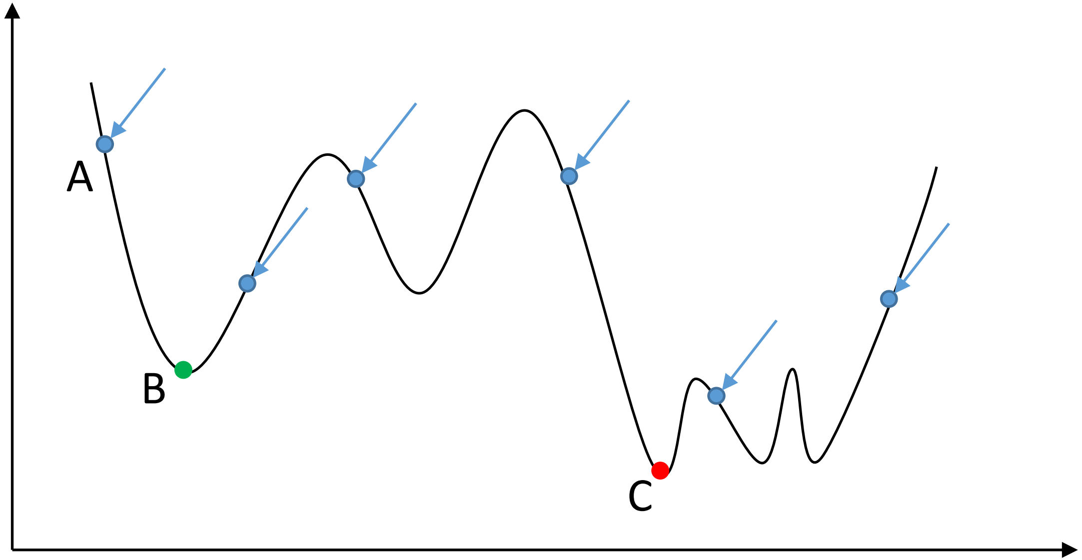


https://www.zybuluo.com/hanbingtao/note/448086

初始化

深蓝学院
shenlanxueyuan.com

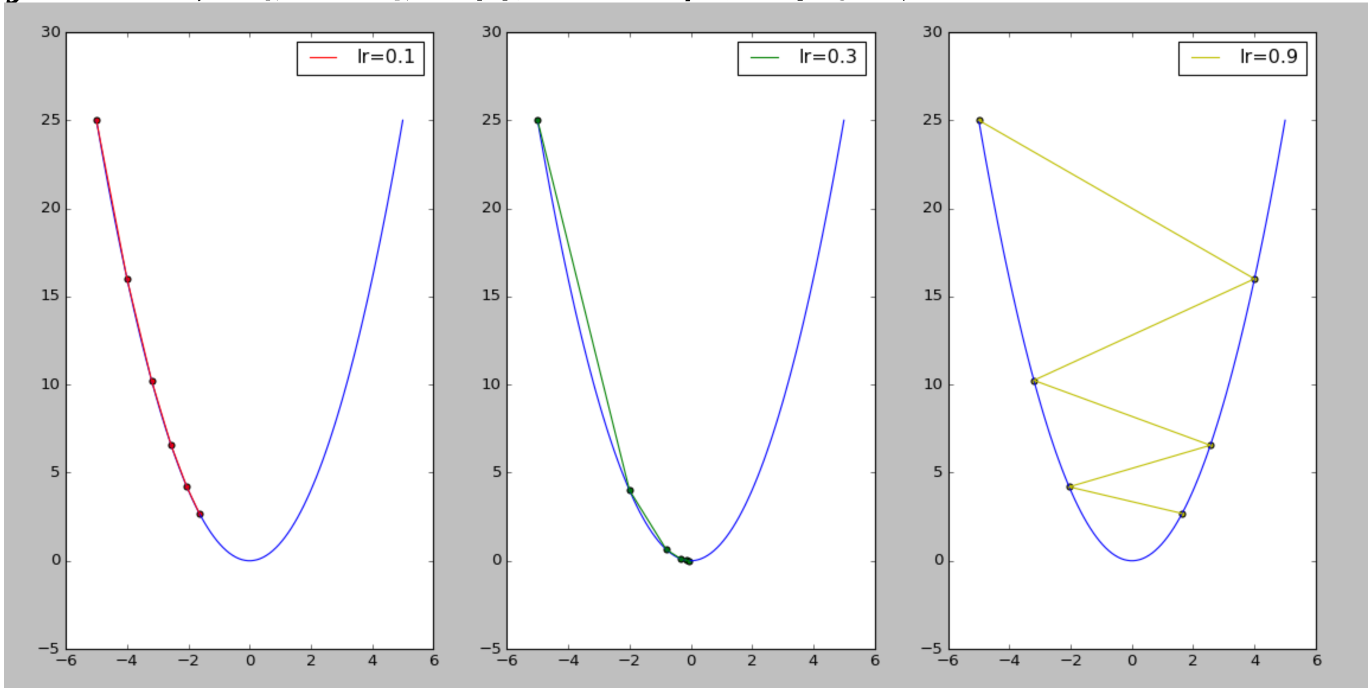
✓ 函数存在多个局部极小值点。迭代前需要选取恰当的初始点。



步长选取

深蓝学院 shenlanxueyuan.com

✓ 以$y=x^2$为例。初始点−5，迭代5次。



http://www.jianshu.com/p/186df2db8898

深度学习应用

- ✓ $\boldsymbol{\theta}$参数估计问题：假定样本集D含有n个样本$(\boldsymbol{x}_1, y_1), \ldots (\boldsymbol{x}_n, y_n)$。假定损失函数为$J(\boldsymbol{\theta}; D, f)$，其中$f(\boldsymbol{x}; \boldsymbol{\theta})$为待估计参数的映射。

- ✓线性回归$J(\boldsymbol{\theta}; D, f) = \frac{1}{2n}\sum_{i=1}^{n}(f(\boldsymbol{x}_i; \boldsymbol{\theta}) - y_i)^2$
- ✓logistic回归$J(\boldsymbol{\theta}; D, f) = -\frac{1}{n}\left[\sum_{i=1}^{n} y_i \ln f(\boldsymbol{x}_i; \boldsymbol{\theta}) + (1 - y_i)\ln\big(1 - f(\boldsymbol{x}_i; \boldsymbol{\theta})\big)\right]$

批量梯度下降(Batch GD)

- ✓ 梯度下降：$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \lambda \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}; D, f)$
- ✓ 线性回归：$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \frac{\lambda}{n} \sum_{i=1}^{n} (f_{\boldsymbol{\theta}}'(\boldsymbol{x}_i; \boldsymbol{\theta}) - y_i)$
- ✓ 每次迭代，使用所有的训练样本，迭代速度受到样本量的影响。尤其样本特别多时，迭代速度较慢

随机梯度下降(Stochastic GD)

- ✓ 为回避一次使用所有样本，随机梯度下降每次迭代随机选取单个样本而决定迭代方向。

- ✓ 线性回归：$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \lambda(f'_{\boldsymbol{\theta}}(\boldsymbol{x}_i; \boldsymbol{\theta}) - y_i)$

- ✓ 单样本用于单次迭代，计算代价相对BGD较小。同时单个样本的迭代方向，也许能应对多极值点，也可能导致迭代路径的曲折

最小批量梯度下降(MiniBatch GD)

- ✓ 折中批量和随机梯度下降，每次迭代随机选取m个样本

- ✓ 线性回归：$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \frac{\lambda}{m}\sum_{i\in\mathcal{J}_m}(f'_{\boldsymbol{\theta}}(\boldsymbol{x}_i;\boldsymbol{\theta}) - y_i)$

- ✓ 常用于深度学习的参数估计

信息论

- ✓ 信息熵
- ✓ 相对熵
- ✓ 互信息
- ✓ 交叉熵及深度学习的应用

信息熵

- ✓ 给定概率密度函数$p(x)$，定义该函数的信息熵

$$H(p) = H[\boldsymbol{x}] = -\int p(\boldsymbol{x})\, ln\, p(\boldsymbol{x})\, d\boldsymbol{x}$$

- ✓ 信息熵描述了分布的混乱程度。均匀分布是使得信息熵最大的概率分布。单点的冲击响应函数对应的信息熵最小。

相对熵

深蓝学院
shenlanxueyuan.com

- ✓ 给定两个概率密度函数$p(\boldsymbol{x})$和$q(\boldsymbol{x})$，描述二者之间的差异（距离），定义相对熵

$$KL(p||q) = -\int p(\boldsymbol{x})\, ln\, q(\boldsymbol{x})\, d\boldsymbol{x} - \left(-\int p(\boldsymbol{x})\, ln\, p(\boldsymbol{x})\, d\boldsymbol{x}\right)$$

$$= -\int p(\boldsymbol{x})\, ln \left\{\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\right\} d\boldsymbol{x}$$

- ✓ 对任意概率分布，$KL(p||q) \geq 0$，等号当且仅当$p = q$。

互信息

- 对于两个随机变量$\boldsymbol{x}, \boldsymbol{y}$，定义二者之间的互信息

$$I[\boldsymbol{x}, \boldsymbol{y}] = KL\big(p(\boldsymbol{x}, \boldsymbol{y}) || p(\boldsymbol{x})p(\boldsymbol{y})\big)$$
$$= -\iint p(\boldsymbol{x}, \boldsymbol{y}) \ln\left(\frac{p(\boldsymbol{x})p(\boldsymbol{y})}{p(\boldsymbol{x}, \boldsymbol{y})}\right) d\boldsymbol{x} d\boldsymbol{y}$$

- 若$\boldsymbol{x}, \boldsymbol{y}$相互独立，则互信息为0，二者相互无关
- $\mathrm{I}[\boldsymbol{x}, \boldsymbol{y}] = H[\boldsymbol{x}] - H[\boldsymbol{x}|\boldsymbol{y}] = H[\boldsymbol{y}] - H[\boldsymbol{y}|\boldsymbol{x}]$。

交叉熵

深蓝学院 shenlanxueyuan.com

- ✓ 给定两个概率密度函数$p(\boldsymbol{x})$和$q(\boldsymbol{x})$，定义$p(\boldsymbol{x})$关于$q(\boldsymbol{x})$的交叉熵

$$H(p,q) = E_p(-\ln q) = -\int p(\boldsymbol{x}) \ln q(\boldsymbol{x})\, d\boldsymbol{x}$$

$$= H(p) + KL(p||q)$$

- ✓ 交叉熵作为logistic、softmax回归的代价函数，常应用神经网络的输出层。

交叉熵

- ✓ 交叉熵$H(p,q) = H(p) + KL(p||q)$

- ✓ logistic回归的对数似然函数

$$\mathrm{E}(\mathbf{w}) = -\left[\sum_{i=1}^{n} t_i \ln g(\boldsymbol{x}_i) + (1 - t_i) \ln\big(1 - g(\boldsymbol{x}_i)\big)\right]$$

- ✓ logistic、softmax回归，常应用神经网络的分类层。

参考文献

- ✓ Pattern Recognition and Machine Learning，Christopher M. Bishop，Springer，2007
- ✓ Pattern Classification， Richard O. Duda，Peter E. Hart and David G. Stork，Wiley-Interscience，2000

在线问答

Q&A

课程地址

《点击此处添加演讲主题演讲主题演讲主题演讲主题》

感谢各位聆听！

Thanks for Listening