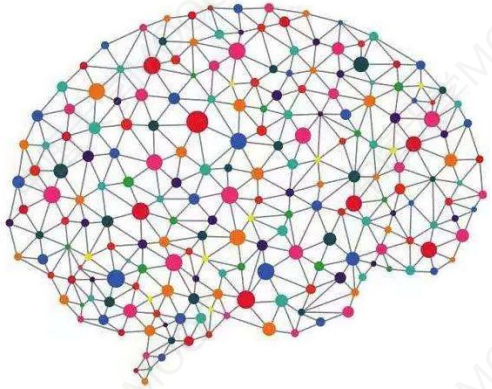
神经网络&深度学习
TENSORFLOW
Google

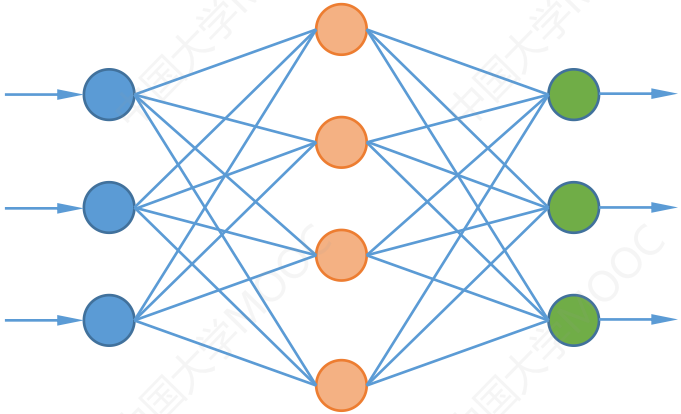


13 人工神经网络(2)

西安科技大学 牟琦
muqi@xust.edu.cn

13.1 小批量梯度下降法

多层神经网络——非线性分类问题



- 损失函数**不是凸函数**，很难计算解析解
- 通常采用**梯度下降法**，得到**数值解**

梯度下降法：求解函数极值问题

- **批量**梯度下降
- **随机**梯度下降
- **小批量**梯度下降

一元线性回归

$$Loss=\frac{1}{2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\frac{1}{2}\sum_{i=1}^{n}(y_i-(wx_i+b))^2$$

n：样本数

批量梯度下降（Batch Gradient Descent，BGD）

- **每次迭代**都使用**所有样本**来计算偏导数

$$w^{(k+1)} = w^{(k)} - \eta \frac{\partial Loss(w,b)}{\partial w} = w^{(k)} - \eta \sum_{i=1}^{n} x_i (w^{(k)} x_i + b - y_i)$$

$$b^{(k+1)} = b^{(k)} - \eta \frac{\partial Loss(w,b)}{\partial b} = b^{(k)} - \eta \sum_{i=1}^{n} (w^{(k)} x_i + b - y_i)$$

- 由**所有样本**确定梯度方向
- 每一步都是准确地向着极值点趋近，**迭代次数少**
- 收敛于**全局极小值**或**局部极小值**点
- 可以利用**向量运算**进行**并行计算**
- 计算量大，训练时间长，不适合大规模数据集

样本集：**20万**个样本，**10次**迭代
计算量：20万×10 = **200万**

随机梯度下降（Stochastic Gradient Descent, SGD）

- **每次迭代**只选择**一个样本**训练模型，使网络的输出尽可能逼近这个样本的标签值
- **一轮**：使用**所有样本**训练一遍
- 反复**训练多轮**，直到网络对所有样本的误差足够小
- **参数更新非常频繁**，**无法快速收敛**
- 不易于实现并行计算

随机梯度下降通常是指**小批量梯度下降**算法

- **小批量梯度下降**（Mini-Batch Gradient Descent, MBGD）
- **小批量随机梯度下降**（Mini-Batch SGD）
  - 把数据分为多个**小批量**，每次迭代使用**一个小批量**来训练模型

$$Loss=\frac{1}{2}\sum_{i=1}^{t}(y_i-\hat{y}_i)^2=\frac{1}{2}\sum_{i=1}^{t}(y_i-(w^{(k)}x_i+b^{(k)}))^2$$

t：每一批中的样本数量

$$w^{(k+1)}=w^{(k)}-\eta\frac{\partial Loss(w^{(k)},b^{(k)})}{\partial w^{(k)}}=w^{(k)}-\eta\sum_{i=1}^{t}x_i(w^{(k)}x_i+b^{(k)}-y_i)$$

$$b^{(k+1)}=b^{(k)}-\eta\frac{\partial Loss(w^{(k)},b^{(k)})}{\partial b^{(k)}}=b^{(k)}-\eta\sum_{i=1}^{t}(w^{(k)}x_i+b^{(k)}-y_i)$$

- **小批量梯度下降**（Mini-Batch Gradient Descent, MBGD）
- **小批量随机梯度下降**（Mini-Batch SGD）
  - 把数据分为多个**小批量**，每次迭代使用**一个小批量**来训练模型
  - 每个**小批量中的所有样本**共同决定了本次迭代中的梯度方向
  - **一轮**：使用**所有小批量**训练一遍
  - 需要**训练多轮**，使网络对**所有样本**的误差足够小
  - **每次迭代**的**训练样本数固定**，与整个训练集的样本数量无关
  - 可以实现**并行运算**
  - 训练**大规模**数据集

样本集：**2000**个样本，10批，每个批中的样本数量：**200**

200000个样本，1000批，每个批中的样本数量：**200**

抽样（Sampling）

从数据集中**随机抽取**出来一部分样本，它们的特征可以**在一定程度上代表完整数据集的特征**

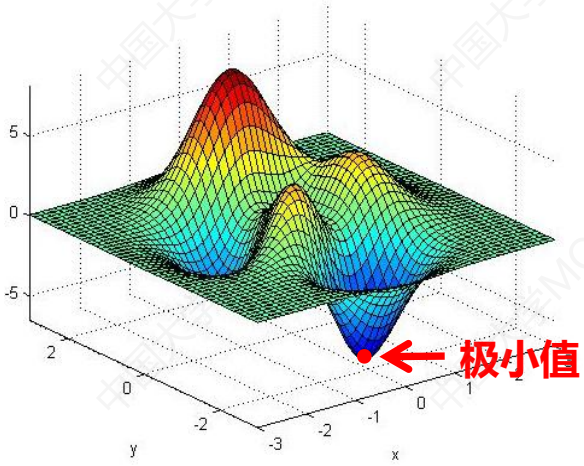
独立同分布：小批量样本能够代表整个样本集的特征

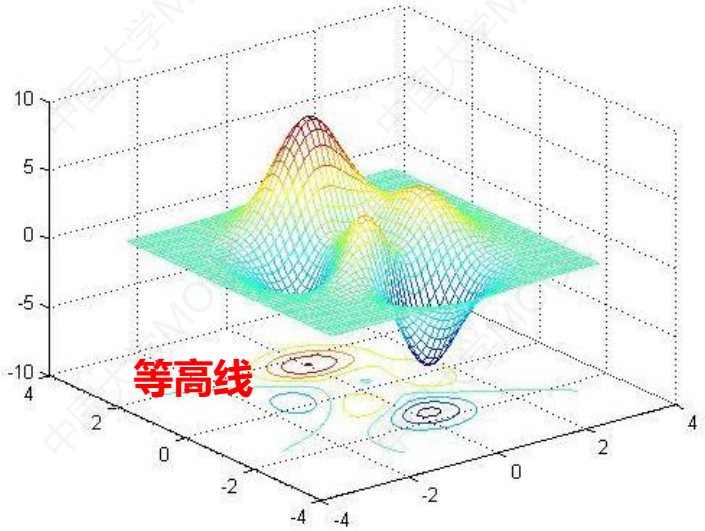
随机抽取：小批量样本的特征和整体样本的特征存在差别

小批量梯度下降

- 小批量样本计算出的**梯度**和使用全体样本计算出的**标准梯度**之间存在偏差
- 总体向最优化的方向前进
- 提高模型的**泛化能力**

损失函数： z=f(x,y)





损失函数： z=f(x,y)