Coursera Machine Learning Week10 大批数据

Coursera Machine Learning Lunar's note

`MachineLearning` `Coursera`

大批量数据学习

大训练集

在很多情况下，大数据量带来的好处显而易见，但是在某些情况下，更多的数据不仅没有好处，还会带来计算上的问题，所以在决定使用大批量数据前要注意斟酌。（可以利用前面提到的学习曲线）

随机梯度下降 Stochastic Gradient Descent

相比于之前用的一般的批量梯度下降，随机梯度下降先将数据随机排序，然后对于每个数据单独去拟合，而不是像之前那样对所有数据一起拟合。单独拟合的话，不需要等到求完所有数据都计算一遍就能得出下降量。
随机梯度下降的迭代次数非常多，下降路线较为曲折，但是每次迭代的速度很快，收敛后会在最优区域徘徊。
批量梯度下降迭代次数较小，下降路线较直，但是每次迭代需要较多时间，最后收敛后就不会移动。

- 小批梯度下降

梯度下降方法	**样本使用**
批量梯度下降	每次迭代使用m个数据
随机梯度下降	每次迭代使用1个数据
小批梯度下降	每次迭代使用b个数据$b \in (1, m)$

随机梯度下降判断收敛的方法：
对于每k次迭代（比如1000次），计算着k次迭代的平均代价函数，画出图像判断收敛。

在线学习

在线学习的要旨在于，每次获取一个样本，使用它更新算法，然后丢弃那个样本。
适合于长期有较大数据流量的网站，因为总是有新的用户会提供样本，而且在线算法会反映出新的用户的特征。
根本区别在于不使用固定的数据集。

映射约减 Map Reduce

将训练集分为几个子集，分给不同的计算机计算出各自的临时变量，然后将临时变量融合并进行迭代，