# 第七章 优化算法

## 1.有监督学习的损失函数

有监督学习涉及的损失函数，请举例并简述它们的优点？

f为预测值

对于二分类问题：

**0-1损失函数**：L(0-1)(f, y) = 1, fy≤0 当且仅当fy≤0时L取值为1，否则取值为0.

**Hinge损失函数**：L(f, y) = max{0, 1-fy}

**Logistic损失函数**：L(f, y) = log2(1+exp(-fy))

**交叉熵损失函数**：L(f, y) = -log2((1+fy)/2)

对于回归问题：

**平方损失函数**：L(f, y) = (f-y)²，为光滑函数，可以使用梯度下降进行优化，对异常值敏感

**绝对损失函数**：L(f, y) = |f - y|

**Huber损失函数**：当|f-y|≤σ时，L(f ,y) = (f - y)²，当|f-y| > σ时，L(f ,y) = 2σ|f - y| - σ²

## 2.机器学习中的优化问题

凸优化的基本概念

什么是凸函数：凸函数曲面上任意两点连接而成的线段，其上的任意一点都不会处于该函数曲面的下方

对于凸优化问题，所有的局部极小值都是全局极小值，因此这类问题一般认为是比较容易求解的问题。

一般来说非凸优化的问题是比较难求解的，但PCA是一个特例，可以借助SVD直接得到主成成分分析的全局最小值。

## 3.经典优化算法

经典的优化算法可以分为**直接法**和**迭代法**两大类。

直接法求解目标函数需要满足两个条件：（1）L是凸函数，那么θ是最优解的充分必要条件是L在θ处的梯度为0.（2）上式有闭式解。同时满足这两个条件的经典例子是Ridge Regression（L2正则化）。

迭代法就是迭代地修正对最优解的估计。一阶法称为梯度下降法，梯度就是目标函数我的一阶信息。二阶法称为牛顿法，当目标函数非凸时，二阶法有可能会收敛到鞍点。

## 4.梯度验证

P.153

## 5.随机梯度下降法

随机梯度下降法适用于数据源源不断到来的在线场景。

小批量梯度下降 (Mini Batch Gradient Descent)：

参数的选择：

（1）如何选取批量大小m：一般m取2的幂次时能充分利用矩阵运算操作。

（2）如何挑选m个训练数据：一般会在每次遍历训练数据之前，先对所有的数据进行随机排序，然后在每次迭代时按顺序挑选m个训练数据直至遍历完所有的数据。

（3）如何选取学习速率α：一开始采用较大的学习率，当误差曲线进入平台期后，减小学习率做更精确的调整

## 6.随机梯度下降法的加速

随机梯度下降法失效的原因：

每次只用到一部分的信息，对梯度的估计常常出现偏差，造成目标函数曲线收敛得很不稳定。而且容易陷入局部最优解。更严重的是进入到山谷和鞍点两种地形。

解决办法：

引入惯性保持和环境感知

**动量方法（Momentum）：**

参数更新方法：

$v_t = \gamma v_{t-1} + n g_t$

$\theta_{(t+1)} = \theta_t - v_t$

前进步伐-v_t由两部分组成，一是学习速率n乘以当前估计的梯度g_t；二是带衰减的前一次步伐v_t-1。这里，惯性就提现在对前一次步伐信息的重利用上。

**AdaGrad方法：**

采用历史梯度平方和来衡量不同参数的梯度的稀疏性，取值越小表明越系数，具体更新公式在P.162，学习速率使用了退火策略，即随着时间推移越来越小，

**Adam方法：**

将惯性和环境感知这两个优点集合。一方面记录梯度的一阶矩，即过往梯度与当前梯度的平均，这体现了惯性保持；另一方面记录梯度的二阶矩，即过往梯度平方与当前梯度平方的平均，这类似AdaGrad方法，体现了环境感知能力，为不同参数产生自适应的学习速率。一

阶和二阶矩采用类似于滑动窗口内求平均的思想进行融合，即当前梯度和近一段时间内梯度的平均值，时间久远的梯度对当前平均值的贡献呈指数衰减。参数更新公式参考P.162

还有其他几种优化方法：
RMSProp，AdaMx, Nadam， AdaDelta

## 7.L1正则化与稀疏性

稀疏性：就是模型的很多参数为0，这相当于对模型进行了一次特征选择，只留下一些比较重要的特征，提高模型的泛化能力，降低过拟合的可能。
L1正则化使得模型参数具有稀疏性的原理是什么?
角度1：解空间的形状
L2正则化项约束后的解空间是圆形，而L1正则化约束的解空间是多边形。显然多边形的解空间更容易在尖角处与等高线碰撞出稀疏解。

角度2：函数叠加
参考书P.167和Github的Kaggle笔记

角度3：贝叶斯先验
L1正则化相当于对模型参数w引入了拉普拉斯先验，L2正则化相当于引入了高斯先验，而拉普拉斯先验使参数为0的可能性更大。
参考书上P.167~P.168拉普拉斯分布和高斯分布的图像特点。