机器学习中的范数规则化

L0范数

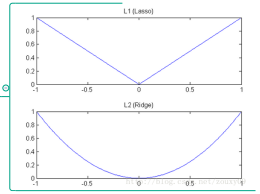
- 定义：向量中非0元素的个数 $\|x\|_0 := \sqrt[0]{\sum_{i=0}^{n} x_i^0}$
- 使用L0范数规则化参数矩阵W，即希望W的大部分元素都为0，即W是稀疏的
- L0范数很难优化求解（NP难问题）

L1范数

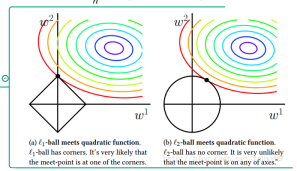
- 定义：向量中各个元素绝对值之和
  - “稀疏规则算子”（Lasso regularization）
  - $\|x\|_1 := \sum_{i=1}^{n} |x_i|$
- L1是L0范数的最优凸近似
  - 任何的规则化算子，如果他在Wi=0的地方不可微，并且可以分解为一个“求和”的形式，那么这个规则化算子就可以实现稀疏
  - $\min \|x\|_0$ s.t. $Ax = b$ ⟺（在一定条件下，以概率1意义下等价）$\min \|x\|_1$ s.t. $Ax = b$
- 参数稀疏的原因
  - 特征选择(Feature Selection)
    - 特征的大部分元素都与最终的输出没有关系或不提供任何信息，稀疏规则用于剔除特征向量中无用的信息
    - 稀疏规则化算子会学习去掉没有信息的特征，将对应的权重置为0
  - 可解释性(Interpretability)
    - 参数稀疏的模型更容易解释
    - 考虑特征维数较多的情况，适用性较好

L2范数

- 定义：||W||2，向量中各元素的平方和再开方
  - “岭回归”（Ridge Regression）
  - “权值衰减”（weight decay）
  - $\|x\|_2 := \sqrt{\sum_{i=1}^{n} x_i^2}$
- 改善过拟合问题
  - L2范数规则项||W||2最小，可以使得W的每个元素都很小，都接近于0。越小的参数说明模型越简单，越简单的模型则越不容易产生过拟合现象
- L2范数的好处
  - 学习理论角度
    - L2范数可以防止过拟合，提升模型的泛化能力
  - 优化计算角度
    - L2范数有助于处理 condition number不好的情况下矩阵求逆很困难的问题
- L1范数与L2范数的区别
  - 下降速度
    - L1是Lasso，L2是Ridge



  - $Lasso:\ \min_w \frac{1}{n}\|y - Xw\|^2,\ \ s.t. \|w\|_1 \le C$
  - $Ridge:\ \min_w \frac{1}{n}\|y - Xw\|^2,\ \ s.t. \|w\|_2 \le C$
  - 模型空间的限制



(a) ℓ_1-ball meets quadratic function. ℓ_1-ball has corners. It's very likely that the meet-point is at one of the corners.

(b) ℓ_2-ball meets quadratic function. ℓ_2-ball has no corner. It is very unlikely that the meet-point is on any of axes."

核范数

- 定义：核范数||W||*是指矩阵奇异值的和(Nuclear Norm)
- 约束Low-Rank（低秩）
  - 如果矩阵表达的是结构性信息，例如图像、用户-推荐信息等，那么这个矩阵各行之间存在这一定的相关性，则矩阵是低秩的
  - 低秩矩阵
    - 如果X是一个m行n列的数值矩阵，rank(X)是X的秩，假如rank (X)远小于m和n，则称X是低秩矩阵
    - 低秩矩阵包含大量冗余信息，这些信息可以对缺失数据进行恢复，也可以对数据进行特征提取
    - 低秩矩阵是非凸的，需要寻找凸近似求解
  - 约束低秩的实例
    - 矩阵填充(Matrix Completion)
      - 推荐系统
      - 低秩矩阵重构问题
        - 已知一个给定的m*n矩阵A，如果其中一些元素因为某种原因丢失了，能够根据其他行和列的元素，将这些元素恢复。如果没有其他的参考条件，要确定数据是困难的。如果已知A的秩rank(A)<<m且rank(A)<<n，则可通过矩阵各行(列)之间的线性相关将丢失的元素求出
    - 鲁棒PCA
      - 主成分分析
        - 有效的找出数据中最“主要”的元素和结构，去除噪音和冗余，将原有的复杂数据降维，揭示隐藏在复杂数据背后的简单结构
      - PCA
        - 目标就是使用另一组基去重新描述得到的数据空间，在这组新的基下，能尽量揭示原有的数据间的关系，最大程度的去除冗余和噪音的干扰
      - 鲁棒主成分分析（Robust PCA）
        - 将矩阵分解为两个矩阵相加，一个低秩矩阵一个稀疏矩阵
        - $\min_{A,E} rank(A) + \lambda \|E\|_0,\ s.t.\ X = A + E$
        - $\min_{A,E} \|A\|_* + \lambda \|E\|_1,\ s.t.\ X = A + E$
        - rank和L0范数在优化上存在非凸和非光滑特性，所以转换为上式子
    - 背景建模
      - 从固定摄相机拍摄的视频中分离背景和前景
    - 变换不变低秩纹理（TILT）
      - 低秩纹理映射算法(TransformInvariant Low-rank Textures，TILT)是一种用低秩性与噪声的稀疏性进行低秩纹理恢复的算法
      - 思想是通过几何变换x把D所代表的图像区域校正成正则的区域，如具有横平竖直、对称等特性，这些特性可以通过低秩性来进行刻画
  - 规则化参数的选择
    - 参数λ：hyper-parameters（超参）
      - λ越大，就表示规则项要比模型训练误差更重要

带参模型

- Loss = Square loss : 最小二乘
- Loss = Hinge loss : SVM
- Loss = Exp-Loss : Boosting
- Loss = Log-Loss : Logistic Regression

监督学习可视为最小化目标函数

$$w^* = \arg\min_w \sum_i L(y_i, f(x_i; w)) + \lambda \Omega(w)$$

- 第一项表示Loss，是L(yi,f(xi;w)) 衡量我们的模型（分类或者回归）对第i个样本的预测值f(xi;w)和真实的标签yi之前的误差
- 第二项表示规则化参数，对参数w的规则化函数Ω(w)去约束模型
- 在所有可能选择的模型中，我们应该选择能够很好地解释已知数据并且十分简单的模型
  - 奥卡姆剃刀(Occam's razor)原理
- 规则化项对应于模型的先验概率
  - 贝叶斯估计
- 在经验风险上加一个正则化项(regularizer)或惩罚项(penalty term)
  - 结构风险最小化策略

L2范数为什么是使||W||2最小？可以理解为：假设权重只有两项，并且权重之和为1，则一定是两项越接近||W||2越小，权重越接近，说明不同特征对输出的影响几乎同等重要，所以模型就更简单了

L1范数和L0范数可以实现稀疏，L1因具有比L0更好的优化求解特性而被广泛应用