浅谈范数正则化

作者：凯鲁嘎吉 - 博客园 http://www.cnblogs.com/kailugaji/

这篇博客介绍不同范数作为正则化项时的作用。首先介绍了常见的向量范数与矩阵范数，然后说明添加正则化项的原因，之后介绍向量的L_0，L_1，L_2范数及其作为正则化项的作用，对三者进行比较分析，并用贝叶斯观点解释传统线性模型与正则化项。随后，介绍矩阵的$L_{2,1}$范数及其推广形式$L_{p,q}$范数，以及矩阵的核范数及其推广形式Schatten范数。最后，用MATLAB程序编写了Laplace分布与Gauss分布的概率密度函数图。有关矩阵范数优化求解问题可参考：一类涉及矩阵范数的优化问题 - 凯鲁嘎吉 - 博客园

1．向量范数与矩阵范数

浅谈范数正则化

➢向量范数 $\boldsymbol{x} \in \mathbf{R}^n$

- 向量的0范数

  向量$\boldsymbol{x}$中非0元素的个数

$$\|\boldsymbol{x}\|_0 = \#(i \mid x_i \neq 0)$$

- 向量的1范数

$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i|$$

- 向量的2范数

$$\|\boldsymbol{x}\|_2 = (\boldsymbol{x}, \boldsymbol{x})^{\frac{1}{2}} = \sqrt{\sum_{i=1}^{n} x_i^2}$$

- 向量的p范数

$$\|\boldsymbol{x}\|_p = (\sum_{i=1}^{n} |x_i|^p)^{\frac{1}{p}}, p \in [1, \infty)$$

- 向量的无穷范数

$$\|\boldsymbol{x}\|_\infty = \max_{1 \le i \le n} |x_i|$$

➢矩阵范数 $\mathrm{A} \in \mathbf{R}^{m \times n}$

- 矩阵的列范数 $\|\mathrm{A}\|_1 = \max_{1 \le j \le n} \sum_{i=1}^{m} |a_{ij}|$
- 矩阵的行范数 $\|\mathrm{A}\|_\infty = \max_{1 \le i \le m} \sum_{j=1}^{n} |a_{ij}|$
- 矩阵的2范数 $\|\mathrm{A}\|_2 = \sqrt{\lambda_{\max}(\mathrm{A}^{\mathrm{T}}\mathrm{A})}$

  其中 $\lambda_{\max}(\mathrm{A}^{\mathrm{T}}\mathrm{A})$表示 $\mathrm{A}^{\mathrm{T}}\mathrm{A}$ 的最大特征值

- 矩阵的Frobenius范数

$$\|\mathrm{A}\|_F = \sqrt{\mathrm{tr}(\mathrm{A}^{\mathrm{T}}\mathrm{A})} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2}$$

- $L_{2,1}$范数 $\|\mathrm{A}\|_{2,1} = \sum_{i=1}^{m} \sqrt{\sum_{j=1}^{n} a_{ij}^2}$
- 核范数：矩阵的奇异值之和(用来约束低秩low-rank) $\mathrm{A} = U\Sigma V^T$

$$\|\mathrm{A}\|_* = \mathrm{tr}\left(\sqrt{\mathrm{A}^{\mathrm{T}}\mathrm{A}}\right) = \mathrm{tr}(\Sigma)$$

1

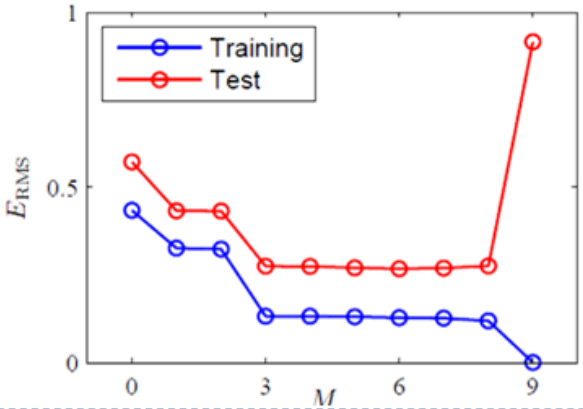
2．为什么要添加正则项?

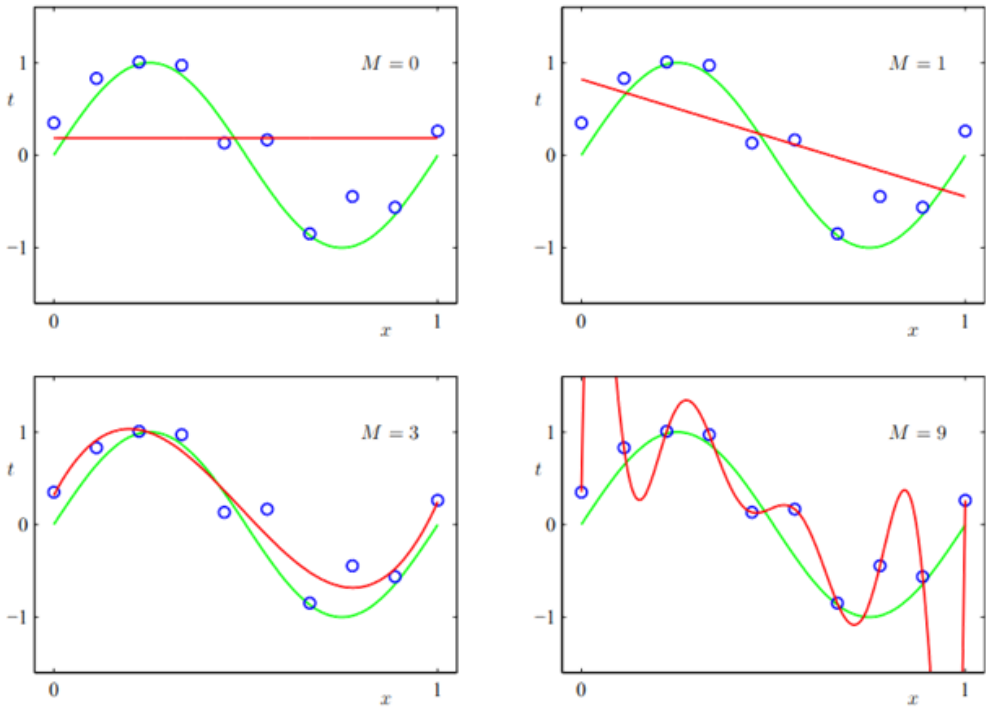
➢ 为什么要添加正则项?

- 考虑如下损失函数

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} (y^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)})^2$$

- 当样本特征很多，而样本数相对较少时，上式很容易陷入过拟合，为了缓解过拟合问题，故引入正则化项，在损失函数上加上某些规则(限制)，强制训练过程使权重转向相对“简单”的权重，缩小解空间，从而减少求出过拟合解的可能性，使模型更通用，可扩展。





- 如上图所示，多项式M的次数越高，自由度越大，对训练数据的过拟合能力也就越大。
- 典型的过拟合是指训练数据上的误差很低，但是测试数据上的误差很高(如左图所示)。可以采用交叉验证来确定最佳的参数M。

3. L_0范数

➢ L_0范数

- L_0范数表示向量$\boldsymbol{x}$中非零元素的个数。

$$\|\boldsymbol{x}\|_0 = \#(i \mid x_i \neq 0)$$

- 如果我们使用L_0范数，即希望$\boldsymbol{x}$的大部分元素都是0(即$\boldsymbol{x}$是稀疏的)，所以L_0范数可用于机器学习中做稀疏编码、特征选择与压缩感知。通过最小化L_0范数，来寻找最少最优的稀疏特征项。
- 但不幸的是，L_0范数在优化上不连续，且非凸非平滑，其最优化问题是一个NP难问题，而且理论上有证明，L_1范数是L_0范数的凸包络(convex envelope)/最优凸松弛(the tightest convex relaxation)，因此通常使用L_1范数来近似L_0范数。L_1范数比L_0范数求解容易。

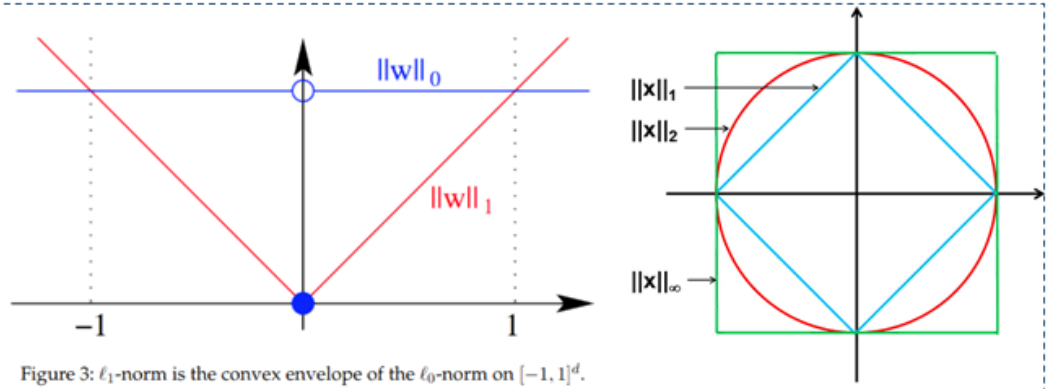


Figure 3: ℓ_1-norm is the convex envelope of the ℓ_0-norm on $[-1, 1]^d$.

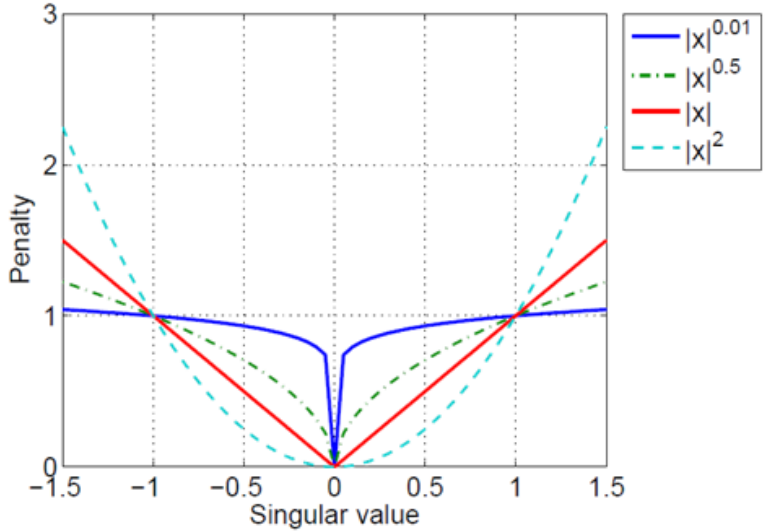


FIG. 2.1. *Penalty functions $|x|^p$ over one singular value x are schematically illustrated for various p. The absolute penalty function $|x|$ is the tightest convex lower bound of the rank ($p \to 0$) in the interval $[-1, 1]$.*

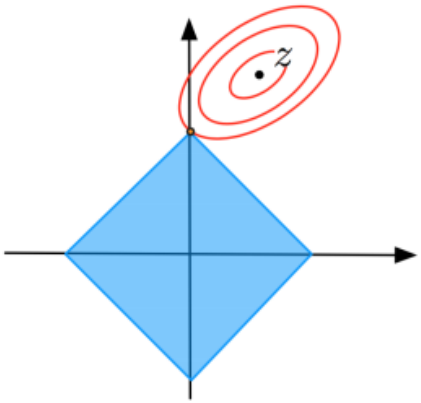
4. L_1范数

➢ L_1范数

- L_1范数表示向量$\boldsymbol{x}$中各个元素绝对值之和。

$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i|$$

- L_1范数，又叫做 taxicab-norm 或者 Manhattan-norm，或 LAD (Least Absolute Deviation，最小绝对偏差)，或Lasso Regression (Least Absolute Shrinkage and Selection Operator Regression，最小绝对收缩和选择算子回归)，计算的是向量$\boldsymbol{x}$与0点之间的曼哈顿距离。
- L_1范数趋向于产生少量特征，而其余特征全为0。其解通常是稀疏性的，倾向于选择数目较少的一些非常大的值或者数目较多的无关紧要的小值。
- 从贝叶斯先验的角度看，正则化等价于对模型参数引入先验分布。即当训练一个模型时，仅依靠当前的训练数据集是不够的，为了实现更好的泛化能力，往往需要加入先验项。L_1范数相当于加入了一个Laplace先验。



$$\min_{w} \frac{1}{2}\|z-w\|^2$$

$$\text{s.t. } \|w\|_1 \le R$$

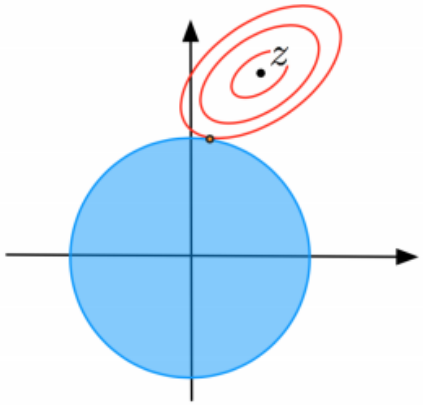
5. L_2范数

➢ L_2范数

- L_2范数表示向量$\boldsymbol{x}$中各个元素平方和再开方。

$$\|\boldsymbol{x}\|_2 = (\boldsymbol{x}, \boldsymbol{x})^{\frac{1}{2}} = \sqrt{\sum_{i=1}^{n} x_i^2}$$

- L_2范数，又叫做 Ridge Regression (岭回归) 或 Tikhonov正则化，计算的是向量$\boldsymbol{x}$与0点之间的欧式距离。
- L_2范数趋向于选择更多的特征，而这些特征全部都很小，趋近于0，而不是等于0。
- 从学习理论的角度来说， L_2范数作为正则项可以防止过拟合，提升模型的泛化能力。L_2范数拟合过程中通常都倾向于让权值尽可能小，最后构造出一个让所有参数都比较小的模型。因为一般认为参数值小的模型比较简单，能够适应于不同的数据集。若目标函数参数很大，那么数据只要偏倚一点点，那么对结果的影响就很大；如果参数很小，即使数据变化范围比较大，对结果影响也不是很大。相对来说，参数较小的话，对模型的抗扰动能力强。
- 从优化或者数值计算的角度来说，L_2范数有助于处理条件数不好的情况下矩阵求逆困难的问题。其优化问题为凸问题，因此可以得到解析解(或封闭解)。 L_2范数比L_1范数求解容易。
- 从贝叶斯先验的角度看，L_2范数相当于加入了一个Gauss先验。



$$\min_{w} \ \frac{1}{2}\|z - w\|^2$$

$$\text{s.t.} \ \|w\|_2 \leq R$$

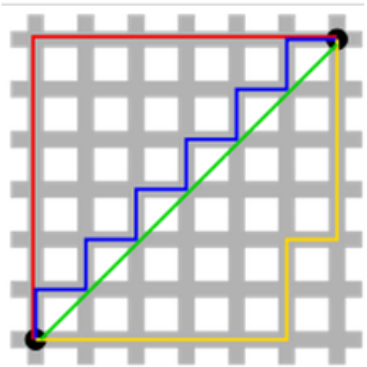
6. L_1范数与L_2范数作为正则项的区别

L_1范数正则化

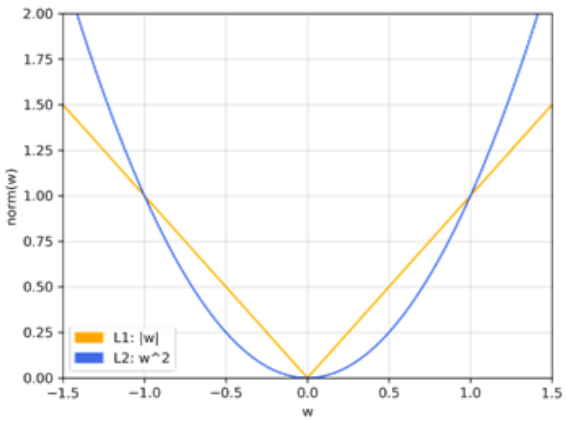
- Lasso回归
- 具有稀疏解，可以用做特征选择
- 选择更少的特征，其余为0，存储效率高
- 适合处理无相关关系的特征
- 无闭式解，不平滑
- 解的个数不唯一，曼哈顿路径不止一条
- 对异常值鲁棒，大异常值成本呈线性增加
- 模型简单可解释性强
- 无法学习复杂的数据模式，如样本数少于维数
- 计算代价大，在非稀疏条件下效率低
- 等价于Laplace先验

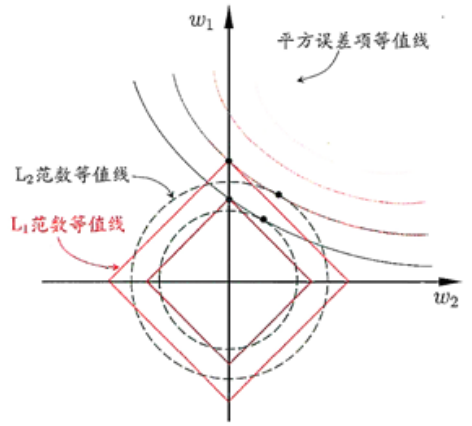
L_2范数正则化

- 岭回归
- 具有非稀疏解
- 选择更多的特征，这些特征都趋于0
- 适合处理相关性较强的特征
- 有闭式解，可导
- 解的个数唯一，欧氏距离仅有一条路径
- 对较大的异常值敏感，大异常值成本呈指数增加
- 模型可解释性差，预测最小化误差方面更准确
- 可以学习复杂的数据模式
- 计算效率高
- 等价于Gauss先验



可视化两个范数的解。L_2范数欧氏距离路径只有一条，而L_1范数曼哈顿路径不止一条。





L_1正则化比L_2正则化数更易获得稀疏解。

7. 用概率解释传统线性回归模型

➢ 传统线性回归模型与概率解释

- 考虑如下线性回归模型

$$\boldsymbol{y}^{(i)} = \boldsymbol{\theta}^T \boldsymbol{x}^{(i)} + \boldsymbol{\varepsilon}^{(i)}$$

其中$\boldsymbol{\varepsilon}$为随机误差，假设 $\boldsymbol{\varepsilon}^{(i)} \sim N(0, \sigma^2)$，即

$$p(\boldsymbol{\varepsilon}^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\boldsymbol{\varepsilon}^{(i)})^2}{2\sigma^2}}$$

$$\Rightarrow p(\boldsymbol{y}^{(i)} \mid \boldsymbol{x}^{(i)}; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\boldsymbol{y}^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)})^2}{2\sigma^2}}$$

这里$\boldsymbol{\theta}$是未知参数，$(\boldsymbol{x}^{(1)}, \boldsymbol{y}^{(1)}), \ldots, (\boldsymbol{x}^{(N)}, \boldsymbol{y}^{(N)})$是来自总体的样本，$L(\boldsymbol{\theta})$是样本的似然函数。采用极大似然估计来求解参数$\boldsymbol{\theta}$的估计量。(注意：这里用的是极大似然估计，并未考虑$\boldsymbol{\theta}$的先验分布情况)

$$\begin{aligned} l(\boldsymbol{\theta}) &= \ln L(\boldsymbol{\theta}) \\ &= \ln \prod_{i=1}^{N} p(\boldsymbol{y}^{(i)} \mid \boldsymbol{x}^{(i)}; \boldsymbol{\theta}) \\ &= \ln \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\boldsymbol{y}^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)})^2}{2\sigma^2}} \\ &= \sum_{i=1}^{N} \ln \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\boldsymbol{y}^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)})^2}{2\sigma^2}} \\ &= N \ln \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (\boldsymbol{y}^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)})^2 \end{aligned}$$

即

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} (\boldsymbol{y}^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)})^2$$

上述目标函数即为原始的最小二乘损失函数。

8. L_2范等价于Gauss先验

➢ L_2范数等价于Gauss先验

- 贝叶斯估计、极大似然估计与最大后验估计之间的关系

贝叶斯估计　极大似然估计　最大后验估计

$$p(\theta \mid X) = \frac{p(X \mid \theta)p(\theta)}{p(X)} = \frac{p(X \mid \theta)p(\theta)}{\int p(X,\theta)d\theta} \propto p(X \mid \theta)p(\theta)$$

Thomas Bayes

- 引入参数$\boldsymbol{\theta}$的先验分布，假设 $\boldsymbol{\theta}^{(j)} \sim N(0,\delta^2)$

$$l(\boldsymbol{\theta}) = \ln L(\boldsymbol{\theta}) = \ln\left(\prod_{i=1}^{N} p(\boldsymbol{y}^{(i)} \mid \boldsymbol{x}^{(i)};\boldsymbol{\theta}) \prod_{j=1}^{M} p(\boldsymbol{\theta}^{(j)})\right)$$

$$= \ln\left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\boldsymbol{y}^{(i)}-\boldsymbol{\theta}^T\boldsymbol{x}^{(i)})^2}{2\sigma^2}} \prod_{j=1}^{M} \frac{1}{\sqrt{2\pi}\delta} e^{-\frac{(\boldsymbol{\theta}^{(j)})^2}{2\delta^2}}\right)$$

$$= \sum_{i=1}^{N} \ln \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\boldsymbol{y}^{(i)}-\boldsymbol{\theta}^T\boldsymbol{x}^{(i)})^2}{2\sigma^2}} + \sum_{j=1}^{M} \ln \frac{1}{\sqrt{2\pi}\delta} e^{-\frac{(\boldsymbol{\theta}^{(j)})^2}{2\delta^2}}$$

$$= N \ln \frac{1}{\sqrt{2\pi}\sigma} + M \ln \frac{1}{\sqrt{2\pi}\delta} - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (\boldsymbol{y}^{(i)} - \boldsymbol{\theta}^T\boldsymbol{x}^{(i)})^2 - \frac{1}{2\delta^2} \sum_{j=1}^{M} (\boldsymbol{\theta}^{(j)})^2$$

等价于

$$\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \sum_{i=1}^{N} (\boldsymbol{y}^{(i)} - \boldsymbol{\theta}^T\boldsymbol{x}^{(i)})^2 + \eta \|\boldsymbol{\theta}\|_2^2$$

上述目标函数即为岭回归损失函数。

由此得出：

参数$\boldsymbol{\theta}$的L_2范数正则项等价于引入$\boldsymbol{\theta}$的Gauss先验。

9. L_1范数等价于Laplace先验

➢ L_1范数等价于Laplace先验

- Laplace概率密度函数

$$p(x \mid \lambda, \mu) = \frac{1}{2\lambda} e^{-\frac{|x-\mu|}{\lambda}}$$

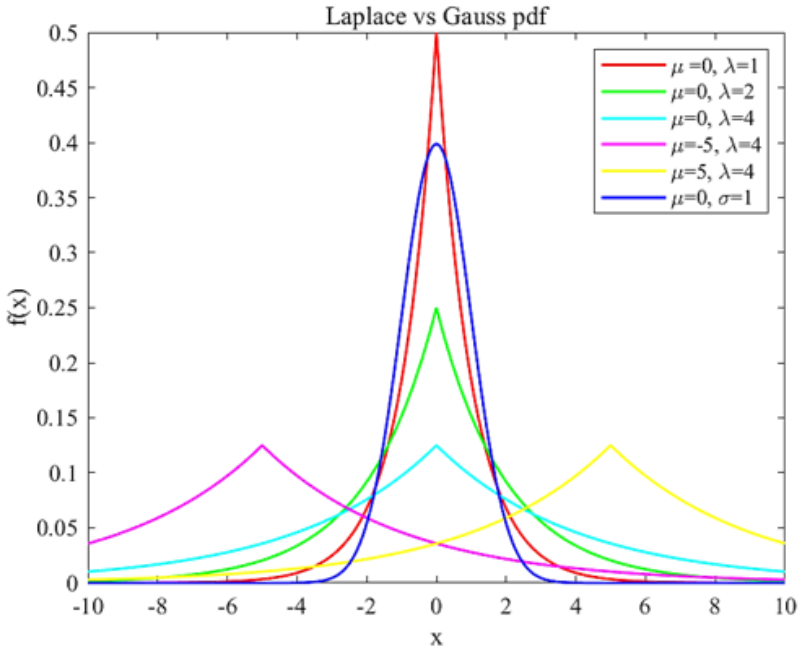
- 引入参数$\boldsymbol{\theta}$的先验分布，假设 $\boldsymbol{\theta}^{(j)} \sim La(\lambda, 0)$

$$l(\boldsymbol{\theta}) = \ln L(\boldsymbol{\theta}) = \ln\left(\prod_{i=1}^{N} p(\boldsymbol{y}^{(i)} \mid \boldsymbol{x}^{(i)}; \boldsymbol{\theta}) \prod_{j=1}^{M} p(\boldsymbol{\theta}^{(j)})\right)$$

$$= \ln\left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\boldsymbol{y}^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)})^2}{2\sigma^2}} \prod_{j=1}^{M} \frac{1}{2\lambda} e^{-\frac{|\boldsymbol{\theta}^{(j)}|}{\lambda}}\right)$$

$$= \sum_{i=1}^{N} \ln \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\boldsymbol{y}^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)})^2}{2\sigma^2}} + \sum_{j=1}^{M} \ln \frac{1}{2\lambda} e^{-\frac{|\boldsymbol{\theta}^{(j)}|}{\lambda}}$$

$$= N \ln \frac{1}{\sqrt{2\pi}\sigma} + M \ln \frac{1}{2\lambda} - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (\boldsymbol{y}^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)})^2 - \frac{1}{\lambda} \sum_{j=1}^{M} |\boldsymbol{\theta}^{(j)}|$$



蓝线表示Gauss分布，其他为Laplace分布。

等价于 $\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \sum_{i=1}^{N} (\boldsymbol{y}^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)})^2 + \xi \|\boldsymbol{\theta}\|_1$

上述目标函数即为Lasso损失函数。

由此得出：

参数$\boldsymbol{\theta}$的L_1范数正则项等价于引入$\boldsymbol{\theta}$的Laplace先验。

10. 矩阵的$L_{2，1}$范数及$L_{p，q}$范数

➢ 矩阵的$L_{2,1}$范数

$$\|X\|_{2,1}=\sum_{i=1}^{m}\sqrt{\sum_{j=1}^{n}x_{ij}^2}=\sum_{i=1}^{m}\|\boldsymbol{x}_i\|_2 \qquad X\in\mathbf{R}^{m\times n}$$

- $L_{2,1}$范数确保行稀疏。用于特征选择、稀疏编码。
- $L_{2,1}$范数是凸的且其极小化问题是很容易求解的。对离群点鲁棒且在计算上十分高效。用$L_{2,1}$范数来进行特征选择时会对所有数据点产生联合稀疏性。
- 注意：矩阵X是m*n的，因此确保行稀疏。如果矩阵为n*m的，则上述范数定义就是确保列稀疏。这取决于如何去定义$L_{2,1}$范数。(看了一些文章，这两种定义都有，注意区分。这里默认为行稀疏，包括下文的推广定义，默认X是m*n的。)

➢ 矩阵的$L_{p,q}$范数

- 矩阵的$L_{2,1}$范数的推广：

$$\|X\|_{p,q}=\left(\sum_{i=1}^{m}\left(\sum_{j=1}^{n}|x_{ij}|^p\right)^{q/p}\right)^{1/q}=\left(\sum_{i=1}^{m}\|\boldsymbol{x}_i\|_p^q\right)^{1/q}$$

11．矩阵的核范数及Schatten范数

➢ 矩阵秩与核范数

- 矩阵补全：当矩阵的奇异值具有稀疏性(即矩阵是低秩的)且采样数目满足一定条件时，大多数矩阵可以通过求解如下优化问题来精确地恢复矩阵中所有元素。其中矩阵秩表示非零奇异值的个数。

$$\min\ rank(\boldsymbol{X})$$
$$s.t.\ X_{ij} = M_{ij}, (i,j) \in \Omega$$

- 然而，该目标函数不是凸函数，求解其最优化问题是一个NP难问题，而且理论上有证明，核范数是矩阵秩的凸包络(convex envelope)/最优凸松弛(the tightest convex relaxation)，因此通常使用核范数来近似矩阵秩。核范数是凸函数，比矩阵秩求解容易。因此上述优化问题转化为最小化核范数优化问题：

$$\min\ \|\boldsymbol{X}\|_*$$
$$s.t.\ X_{ij} = M_{ij}, (i,j) \in \Omega$$

- 核范数表示矩阵奇异值之和，考虑矩阵的奇异值分解 $X = U\Sigma V^T$，即 $\|X\|_* = tr\left(\sqrt{X^T X}\right) = tr(\Sigma)$
- 核范数(nuclear norm)也称迹范数(trace norm)，用来约束低秩，去除冗余信息，即矩阵中线性相关的行或列。

➢ 矩阵的Schatten范数

- 矩阵的核范数的推广：

$$\|X\|_p = \left(\sum_{i=1}^{\min\{m,n\}} \sigma_i^p \right)^{1/p}$$

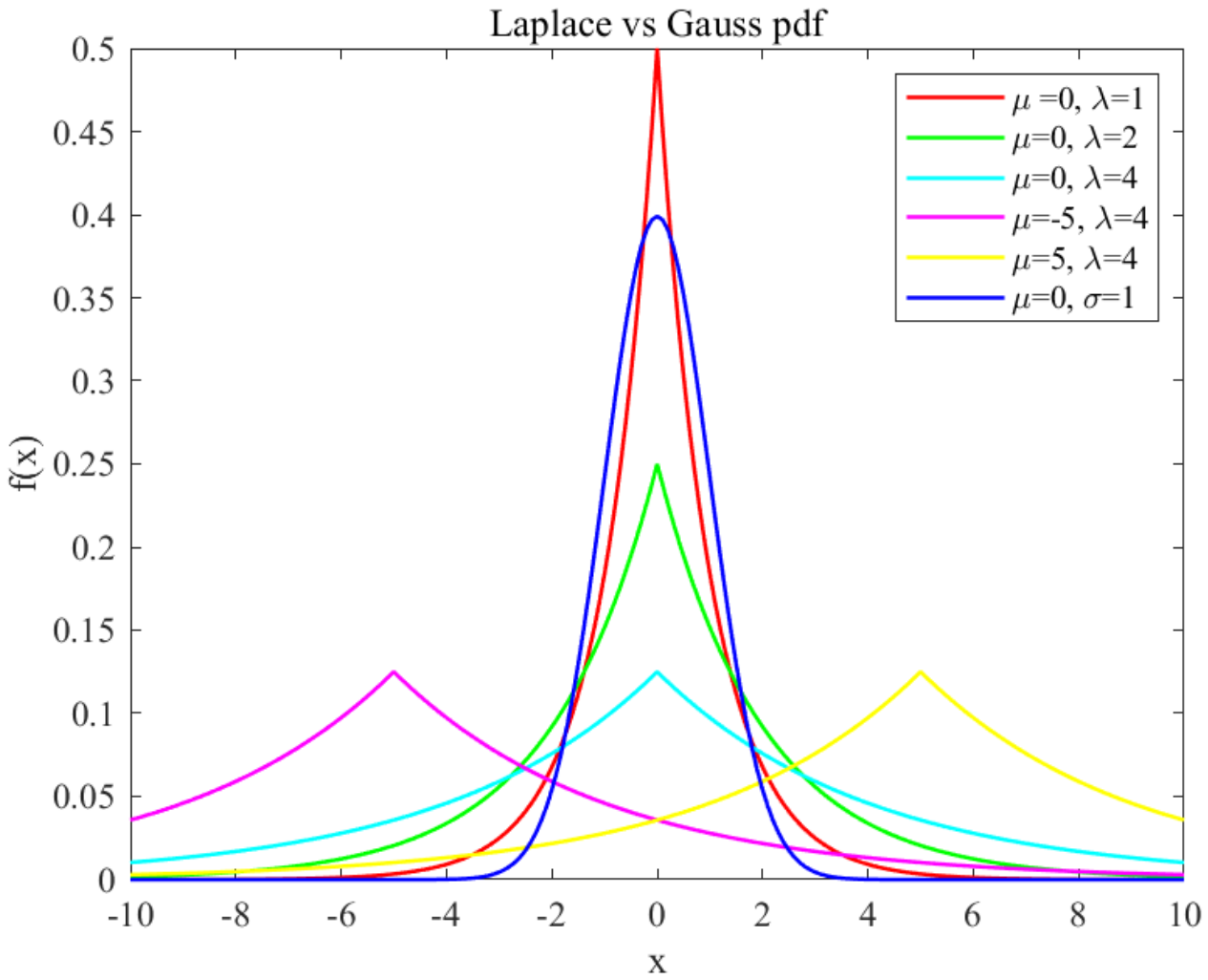
- 定义在矩阵的奇异值上，可用于解决各类低秩问题，例如压缩感知 (compressed sensing)、低秩矩阵/张量恢复(low-rank matrix/tensor completion)。

12. MATLAB程序：Laplace分布与Gauss分布的概率密度函数图

```
%% Demo of Laplace Density Function
% x : variable
% lambda : size para
%miu: location para
clear
```

```
clc
x = -10:0.1:10;
y_1=Laplace_distribution(x, 0, 1);
y_2=Laplace_distribution(x, 0, 2);
y_3=Laplace_distribution(x, 0, 4);
y_4=Laplace_distribution(x, -5, 4);
y_5=Laplace_distribution(x, 5, 4);
y_6=normpdf(x,0,1);
plot(x, y_1, 'r-', x, y_2, 'g-', x, y_3, 'c-', x, y_4, 'm-', x, y_5, 'y-', x, y_6, 'b-', 'LineWidth',1.2);
legend('\mu =0, \lambda=1','\mu=0, \lambda=2','\mu=0, \lambda=4','\mu=-5, \lambda=4','\mu=5, \lambda=4', '\mu=0, \sigma=1'); %图例的设置
xlabel('x');
ylabel('f(x)');
title('Laplace vs Gauss pdf');
set(gca, 'FontName', 'Times New Roman', 'FontSize',11);
saveas(gcf,sprintf('demo_Laplace_Gauss.jpg'),'bmp'); %保存图片

%% Laplace Density Function
function y=Laplace_distribution(x, miu, lambda)
    y = 1 / (2*lambda) * exp( -abs(x-miu)/lambda);
end
```



13. 参考文献

[1] 证明核范数是矩阵秩的凸包络

EJ Candès, Recht B . Exact Matrix Completion via Convex Optimization[J]. Foundations of Computational Mathematics, 2009, 9(6):717.

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.312.1183&rep=rep1&type=pdf

[2] 关于说明L_1范数是L_0范数的凸包络的文献及教案

Donoho D L ， Huo X . Uncertainty Principles and Ideal Atomic Decomposition[J]. IEEE Transactions on Information Theory, 2001, 47(7):2845-2862.

http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=00BC0C50CDECB265657379792F917FFE?doi=10.1.1.161.9300&rep=rep1&type=pdf

Learning with Combinatorial Structure Note for Lecture 12

http://people.csail.mit.edu/stefje/fall15/notes_lecture12.pdf

L1-norm Methods for Convex-Cardinality Problems

https://web.stanford.edu/class/ee364b/lectures/l1_slides.pdf

[3] 有关过拟合的教案及图片来源

2017 Lecture 2: Overfitting. Regularization

https://www.cs.mcgill.ca/~dprecup/courses/ML/Lectures/ml-lecture02.pdf

[4] 一些可供参考的资料

The difference between L1 and L2 regularization

https://explained.ai/regularization/L1vsL2.html

Why L1 norm for sparse models

https://stats.stackexchange.com/questions/45643/why-l1-norm-for-sparse-models

Why L1 regularization can “zero out the weights” and therefore leads to sparse models? [duplicate]

https://stats.stackexchange.com/questions/375374/why-l1-regularization-can-zero-out-the-weights-and-therefore-leads-to-sparse-m

What are L1, L2 and Elastic Net Regularization in neural networks?

https://www.machinecurve.com/index.php/2020/01/21/what-are-l1-l2-and-elastic-net-regularization-in-neural-networks/

Introduction. Sharpness Enhancement and Denoising of Image Using L1-Norm Minimization Technique in Adaptive Bilateral Filter.

https://www.ijsr.net/archive/v3i11/T0NUMTQxMzUy.pdf