Sparse Solutions to Least-Squares Problems Using the LASSO

Objectives

- motivate search for sparse solutions
- introduce ℓ_1-norm regularization (LASSO)
- overview attributes of ℓ_1-regularization

Sparse classifiers / models give insight

稀疏模型就是模型中的绝大部分参数都是0（意味着 "有的部分特征是缺的"）

$(\underline{x}_i, d_i)_{i=1,\dots N}$ features, labels （N samples）

$\underline{x}_i^T \underline{w} \approx d_i$

$$\underline{A}\,\underline{w} = \begin{bmatrix} \underline{a}_1 & \underline{a}_2 & \cdots & \underline{a}_m \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{bmatrix} = \sum_{i=1}^{m} w_i \underline{a}_i$$

（m features；$N \times m$；对所有特征列的加权平均）

$\underline{a}_\ell$: ℓ^{th} feature component

Suppose $w_\ell \approx 0 \Rightarrow \underline{a}_\ell$ is unimportant

If a small number of w_ℓ are nonzero, only those few features matter! $\underline{w}$ is sparse

$$\|\underline{w}\|_0 = \sum_{i=1}^{m} \mathbb{1}_{\{w_i \neq 0\}}$$

（非零 w 的个数）

ℓ_0 "norm" (number of nonzero elements)

$\|a\underline{w}\|_0 \neq a\|\underline{w}\|_0$ （其实不满足数学上对于范数的定义）

Consider $\min_{\underline{w}} \|\underline{w}\|_0$ s.t. $\|\underline{A}\underline{w} - \underline{d}\|_2^2 < \varepsilon$

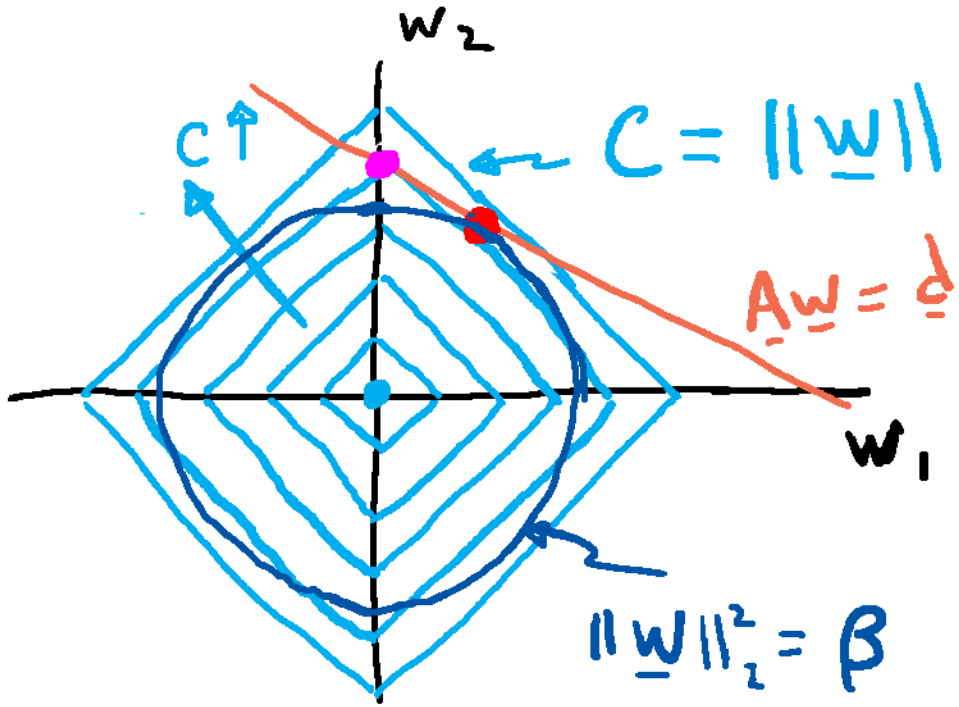
（The more sparse, the better）

non convex - intractable

Convex relaxation gives tractable problem

$\min_{\underline{w}} \|\underline{w}\|_1$ s.t. $\|\underline{A}\underline{w} - \underline{d}\|_2^2 < \varepsilon$ LASSO: Least Absolute Selection & Shrinkage Operator

convex → This solution is tractable



$C = \|\underline{w}\|_1 = \sum_{i=1}^{M} |w_i|$: $|w_1| + |w_2| = C$

$\underline{A}\underline{w} = \underline{d}$

1st quad $w_1 + w_2 = C$

$\min \|\underline{w}\|_1$ s.t. $\underline{A}\underline{w} = \underline{d}$

$w_1 = (0, w_2)$

"corners" on $\|\underline{w}\|_1$ ⇒ sparse solns

$\min \|\underline{w}\|_2^2$ s.t. $\underline{A}\underline{w} = \underline{d}$ circular $\|w\|_2^2$ ⇒ non sparse solutions

$\|w\|_2^2$ 情況下 $Aw=d$ 的直線和此點交於 (w_1, w_2) (w_1 和 w_2 都不=0)

LASSO is a regularized least-squares problem 4

$$\min_{\underline{w}} \|\underline{w}\|_1 \text{ s.t. } \|\underline{A}\underline{w}-\underline{d}\|_2^2 < \varepsilon$$ is equivalent to

$$\min_{\underline{w}} \|\underline{A}\underline{w}-\underline{d}\|_2^2 + \lambda\|\underline{w}\|_1$$ for some λ, ε

Note: $\min_{\underline{w}} \|w\|_1 + \frac{1}{\lambda}\|\underline{A}\underline{w}-\underline{d}\|_2^2$ → If we factor out the λ, this is the same with the above line

LASSO

$$\underline{w}_L = \arg\min_{\underline{w}} \|\underline{A}\underline{w}-\underline{d}\|_2^2 + \lambda\|\underline{w}\|_1$$

Sparse $\underline{w}_{Lasso}$

Can have small model error $\underline{w}_{opt} - \underline{w}_L$

iterative solution

Ridge Regression

$$\underline{w}_R = \arg\min_{\underline{w}} \|\underline{A}\underline{w}-\underline{d}\|_2^2 + \lambda\|\underline{w}\|_2^2$$

non sparse $\underline{w}_{Ridge}$

great prediction error $\|\underline{A}\underline{w}_{opt} - \underline{A}\underline{w}_R\|_2^2$ ($\underline{w}_{opt}$: optimal; $\underline{w}_R$: ridge regression solution)

can solve in closed form

LASSO may be used for model/feature selection

$$\underline{w}_L = \arg\min_{\underline{w}} \|\underline{A}\underline{w} - \underline{d}\|_2^2 + \lambda\|\underline{w}\|_1$$

Model error

Sparsity penalty

(迫使不重要的特征的权重 w 为0，我们选择出那些权重不为0的特征)

$$S_L = \{i : [\underline{w}_L]_i \neq 0\}$$ selected features

被选中的特征集合，那些 w 不为0的索引 set

$$\underline{A}\underline{w}_L = \sum_{i=1}^{n} \underline{a}_i [\underline{w}_L]_i = \sum_{i \in S_L} \underline{a}_i [\underline{w}_L]_i$$

Debiasing

$$\underline{A}_L = \{\underline{a}_i : i \in S_L\}$$

只保留 LASSO 选中的那些特征列

$$\hat{\underline{w}}_L = \arg\min_{\underline{w}} \|\underline{A}_L \underline{w} - \underline{d}\|_2^2 = (\underline{A}_L^T \underline{A}_L)^{-1} \underline{A}_L^T \underline{d}$$

avoids shrinkage due to $\|\underline{w}\|_1$

① 只用 LASSO 选 feature

② 模型重训 → 组成一个更小的 A 做标准 LS