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

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

features, labels

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

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

$\underline{a}_\ell$: 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

$$\|\underline{w}\|_0 = \sum_{i=1}^{M} \mathbb{1}_{\{w_i \neq 0\}}$$ (number of nonzero elements)

ℓ_0 "norm"

$\|a\underline{w}\|_0 \neq a\|\underline{w}\|_0$

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

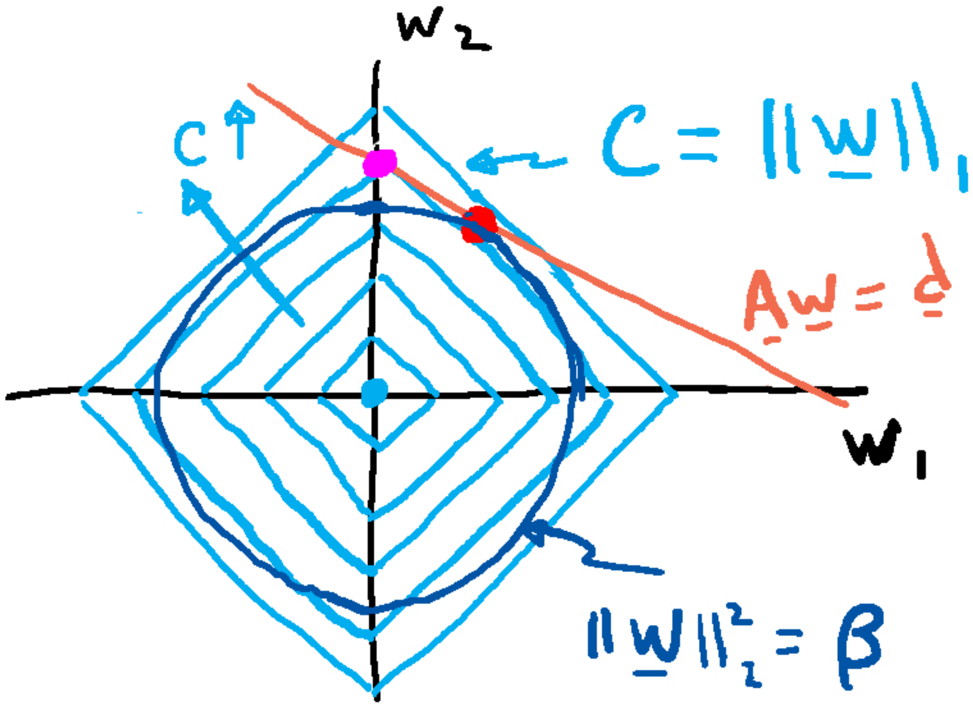
non convex - intractable

Convex relaxation gives tractable problem

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

convex

LASSO: Least Absolute Selection & Shrinkage Operator



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

1^{st} quad $\quad w_1 + w_2 = c$

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

"corners" on $\|\underline{w}\|_1 \Rightarrow$ sparse sol'ns

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

LASSO is a regularized least-squares problem

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

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

Note: $\min_{\underline{w}} \|\underline{w}\|_1 + \frac{1}{\lambda} \|\underline{A}\,\underline{w} - \underline{d}\|_2^2$

LASSO	Ridge Regression
$\underline{w}_L = \arg\min_{\underline{w}} \|\underline{A}\,\underline{w} - \underline{d}\|_2^2 + \lambda \|\underline{w}\|_1$	$\underline{w}_R = \arg\min_{\underline{w}} \|\underline{A}\,\underline{w} - \underline{d}\|_2^2 + \lambda \|\underline{w}\|_2^2$
sparse $\underline{w}_L$	non sparse $\underline{w}_R$
can have small model error $\underline{w}_{opt} - \underline{w}_L$	great prediction error $\|\underline{A}\,\underline{w}_{opt} - \underline{A}\,\underline{w}_R\|_2^2$
iterative solution	can solve in closed form

LASSO may be used for model/feature selection 5

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

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

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

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

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

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