

MACHINE LEARNING IN BIOINFORMATICS

FEATURE SELECTION

Philipp Benner
philipp.benner@bam.de

S.3 - eScience
Federal Institute for Materials Research and Testing (BAM)

July 22, 2023

Feature selection problem

$$\hat{\theta} = \begin{cases} \arg \min_{\theta} & \|y - X\theta\|_2^2 \\ \text{subject to} & \|\theta\|_0 = m \end{cases} \quad \text{with } \binom{p}{m} \text{ possible subsets}$$

- Required are computationally efficient methods to approximate the feature selection problem
- *Offline methods*: Select features before estimating parameters
- *Online methods*: Features are selected during parameter estimation

■ Offline methods:

- ▶ Safe and Strong rules
- ▶ Sure independence screening (SIS)
- ▶ Estimation of mutual information

■ Online methods:

- ▶ (Orthogonal) matching pursuit
- ▶ Least angle regression (LARS) / Homotopy algorithm
- ▶ Penalty methods

LINEAR REGRESSION - RECAP

$$y = X\theta + \epsilon$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(p)} \\ x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(p)} \\ \vdots & \vdots & \ddots & \vdots \\ x_n^{(1)} & x_n^{(2)} & \dots & x_n^{(p)} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

response : $y \in \mathbb{R}^n$

covariates : $X \in \mathbb{R}^{n \times p}$

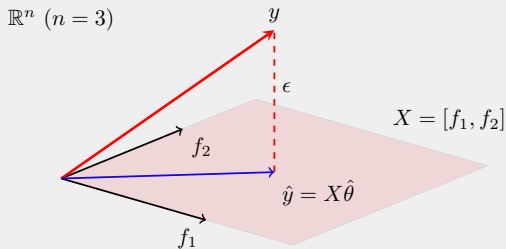
coefficients : $\theta \in \mathbb{R}^p$

residuals : $\epsilon \in \mathbb{R}^n$

LINEAR REGRESSION - RECAP

Geometric interpretation of ordinary least squares
[Hastie et al., 2009]:

$$\begin{aligned}\hat{\theta} &= \arg \min_{\theta} \|\epsilon\|_2^2 \\ &= \arg \min_{\theta} \|y - X\theta\|_2^2\end{aligned}$$



SURE INDEPENDENCE SCREENING (SIS)

SURE INDEPENDENCE SCREENING (SIS)

- Consider the case of ultrahigh-dimensional data, where the number of features p is much larger than the number of observations n
- Specifically, we assume that p is so large that we cannot compute an estimate of θ
- Assuming θ is sparse, we can first select a *promising* subset of q features M_q (called *feature screening*)
- The coefficients θ are estimated based on the subset M_q

SURE INDEPENDENCE SCREENING (SIS)

- Consider the solution of ridge regression:

$$\hat{\theta}(\lambda) = (X^T X + \lambda I)^{-1} X^T y$$

- For $\lambda \rightarrow 0$ we obtain the OLS solution
- For $\lambda \rightarrow \infty$ it follows that $\lambda \hat{\theta}(\lambda)$ converges to the componentwise regression estimator

$$\hat{\theta}_k(\lambda) = \tilde{X}^T y$$

where \tilde{X} is the data matrix X with normalized columns f_j such that $f_j^T f_j = 1$

- Traditionally, for very large p we would select λ large in order to decrease the variance of $\hat{\theta}$

SURE INDEPENDENCE SCREENING (SIS)

- $\tilde{X}^\top y = (f_1^\top y, \dots, f_p^\top y)$ can be interpreted as the correlation of features f_j with y
- Sure independence screening (SIS) [Fan and Lv, 2008] selects a subset of features

$$\Omega = \left\{ j \mid |f_j^\top y| > t \right\} \quad (1)$$

based on their correlation with y , where t is a threshold such that $|\Omega| = q < p$

- The OLS estimate $\hat{\theta}$ is computed using only the selected features Ω
- All remaining components of $\hat{\theta}$ are set to zero

SURE INDEPENDENCE SCREENING (SIS)

- The same idea can be applied to more complex models [Fan and Song, 2010], such as logistic regression, where

$$\hat{\theta} = \arg \max_{\theta} \text{pr}_{\theta}(y | X)$$

- Select a subset of features

$$\Omega = \{ j \mid \text{score}(f_j, y) > t \} \quad (2)$$

- The score is given by the independent estimate

$$\text{score}(f_j, y) = \arg \max_{\theta_j} \text{pr}_{\theta_j}(y | f_j)$$

for all $j = 1, \dots, p$

MATCHING PURSUIT FOR LINEAR REGRESSION

Feature selection problem

$$\hat{\theta} = \begin{cases} \arg \min_{\theta} & \|y - X\theta\|_2^2 \\ \text{subject to} & \|\theta\|_0 = m \end{cases} \quad \text{with } \binom{p}{m} \text{ possible subsets}$$

MATCHING PURSUIT FOR LINEAR REGRESSION

Feature selection problem

$$\hat{\theta} = \begin{cases} \arg \min_{\theta} & \|y - X\theta\|_2^2 \\ \text{subject to} & \|\theta\|_0 = m \end{cases} \quad \text{with } \binom{p}{m} \text{ possible subsets}$$

Matching Pursuit

Greedy approximation to feature selection problem.

MATCHING PURSUIT FOR LINEAR REGRESSION

Feature selection problem

$$\hat{\theta} = \begin{cases} \arg \min_{\theta} & \|y - X\theta\|_2^2 \\ \text{subject to} & \|\theta\|_0 = m \end{cases} \quad \text{with } \binom{p}{m} \text{ possible subsets}$$

Matching Pursuit

Greedy approximation to feature selection problem.

If we must represent y with only one feature, which one should we take?

MATCHING PURSUIT FOR LINEAR REGRESSION

Feature selection problem

$$\hat{\theta} = \begin{cases} \arg \min_{\theta} & \|y - X\theta\|_2^2 \\ \text{subject to} & \|\theta\|_0 = m \end{cases} \quad \text{with } \binom{p}{m} \text{ possible subsets}$$

Matching Pursuit

Greedy approximation to feature selection problem.

If we must represent y with only one feature, which one should we take?

$$j_1 = \arg \min_j \|y - f_j \hat{\theta}_j\|_2^2, \quad \text{where} \quad \hat{\theta}_j = \arg \min_{\theta_j} \|y - f_j \theta_j\|_2^2$$

MATCHING PURSUIT FOR LINEAR REGRESSION

$$j_1 = \arg \min_j \left\| y - f_j \hat{\theta}_j \right\|_2^2, \quad \text{where} \quad \hat{\theta}_j = \arg \min_{\theta_j} \left\| y - f_j \theta_j \right\|_2^2$$

MATCHING PURSUIT FOR LINEAR REGRESSION

$$j_1 = \arg \min_j \left\| y - f_j \hat{\theta}_j \right\|_2^2, \quad \text{where} \quad \hat{\theta}_j = \arg \min_{\theta_j} \left\| y - f_j \theta_j \right\|_2^2$$

$$= \arg \max_j \frac{(f_j^\top y)^2}{f_j^\top f_j}$$

$$= \arg \max_j \left| f_j^\top y \right|$$

[assuming normalized data, i.e. $f_j^\top f_j = 1$]

\Rightarrow select feature j with maximal scalar projection of y onto f_j

MATCHING PURSUIT FOR LINEAR REGRESSION

$$\begin{aligned}\epsilon &= y - X\theta \\ &= \underbrace{y}_{r_0} - \underbrace{f_{j_1}\theta_{j_1}}_{r_1} - \underbrace{f_{j_2}\theta_{j_2} - \dots - f_{j_p}\theta_{j_p}}_{r_2}\end{aligned}$$

$$\begin{aligned}j_1 &= \arg \min_j \left\| y - f_j \hat{\theta}_j \right\|_2^2 &= \arg \min_j \left\| r_0 - f_j \hat{\theta}_j \right\|_2^2 \\ &= \arg \max_j \left| f_j^\top r_0 \right|\end{aligned}$$

MATCHING PURSUIT FOR LINEAR REGRESSION

$$\begin{aligned}\epsilon &= y - X\theta \\ &= \underbrace{y}_{r_0} - \underbrace{f_{j_1}\theta_{j_1}}_{r_1} - \underbrace{f_{j_2}\theta_{j_2} - \dots - f_{j_p}\theta_{j_p}}_{r_2}\end{aligned}$$

$$j_1 = \arg \min_j \left\| y - f_j \hat{\theta}_j \right\|_2^2 = \arg \min_j \left\| r_0 - f_j \hat{\theta}_j \right\|_2^2$$

$$= \arg \max_j \left| f_j^\top r_0 \right|$$

$$j_2 = \arg \min_j \left\| y - f_{j_1} \hat{\theta}_{j_1} - f_j \hat{\theta}_j \right\|_2^2 = \arg \min_j \left\| r_1 - f_j \hat{\theta}_j \right\|_2^2$$

$$= \arg \max_j \left| f_j^\top r_1 \right|$$

MATCHING PURSUIT FOR LINEAR REGRESSION

Matching pursuit (MP) [Tropp et al., 2007]

The MP feature selection rule is given by

$$j_k = \arg \max_j \left| \mathbf{f}_j^\top \mathbf{r}_{k-1} \right| \quad k = 1, \dots, m$$

where \mathbf{r}_k are the residuals at step k :

$$\begin{aligned} \epsilon &= \mathbf{y} - \mathbf{X}\theta \\ &= \underbrace{\mathbf{y}}_{\mathbf{r}_0} - \underbrace{\mathbf{f}_{j_1} \theta_{j_1}}_{\mathbf{r}_1} - \mathbf{f}_{j_2} \theta_{j_2} - \dots - \mathbf{f}_{j_p} \theta_{j_p} \\ &\quad \underbrace{\hspace{10em}}_{\mathbf{r}_2} \end{aligned}$$

Orthogonal Matching Pursuit

Orthogonal Matching Pursuit: Re-estimate parameters after every iteration.

After every iteration t , update all θ_{Ω_t} entries, where $\Omega_t = \{j_1, j_2, \dots, j_t\}$, i.e. compute

$$\theta_{\Omega_t} = \arg \min_{\theta} \|y_{\Omega_t} - X_{\Omega_t} \theta\|_2^2 .$$

This update changes the residuals

$$r_t = y - f_{j_1} \theta_{j_1} - f_{j_2} \theta_{j_2} - \dots - f_{j_t} \theta_{j_t}$$

used in the next iteration of the algorithm.

MATCHING PURSUIT FOR LOGISTIC REGRESSION

LOGISTIC REGRESSION

$$\begin{bmatrix} \text{pr}_{\theta}(y_1 = 1) \\ \text{pr}_{\theta}(y_2 = 1) \\ \vdots \\ \text{pr}_{\theta}(y_n = 1) \end{bmatrix} = \sigma \left(\begin{bmatrix} x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(p)} \\ x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(p)} \\ \vdots & \vdots & \ddots & \vdots \\ x_n^{(1)} & x_n^{(2)} & \dots & x_n^{(p)} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{bmatrix} \right)$$

class labels :

$$\mathbf{y} \in \{0, 1\}^n$$

covariates :

$$\mathbf{X} \in \mathbb{R}^{n \times p}$$

coefficients :

$$\theta \in \mathbb{R}^p$$

Parameter estimation for logistic regression:

$$\hat{\theta} = \arg \max_{\theta} \text{pr}_{\theta}(\mathbf{y}) \approx \arg \min_{\theta} \|\mathbf{y} - \sigma(\mathbf{X}\theta)\|_2^2 \quad [\text{but not convex}]$$

$$= \arg \max_{\theta} \sum_{i=1}^n \log \text{pr}_{\theta}(y_i)$$

$$= \arg \max_{\theta} \sum_{i=1}^n \{y_i \log \sigma(x_i \theta) + (1 - y_i) \log(-x_i \theta)\}$$

$$= \arg \max_{\theta} \sum_{i=1}^n \log \sigma(\tilde{y}_i x_i \theta),$$

where $\tilde{y}_i = 2y_i - 1 \in \{-1, 1\}$

Pseudo-residuals

$$r_k = y - \sigma(f_{j_1} \theta_{j_1} + f_{j_2} \theta_{j_2} + \cdots + f_{j_k} \theta_{j_k})$$
$$X^\top r_p = \nabla \log \text{pr}_\theta(y)$$

Pseudo-residuals

$$r_k = y - \sigma(\mathbf{f}_{j_1} \theta_{j_1} + \mathbf{f}_{j_2} \theta_{j_2} + \cdots + \mathbf{f}_{j_k} \theta_{j_k})$$
$$\mathbf{X}^\top \mathbf{r}_p = \nabla \log \text{pr}_\theta(\mathbf{y})$$

$$j_1 = \arg \min_j \left\| \mathbf{y} - \sigma(\mathbf{f}_j \hat{\theta}_j) \right\|_2^2$$
$$\approx \arg \max_j \left| \mathbf{f}_j^\top \mathbf{r}_0 \right|$$

Pseudo-residuals

$$r_k = y - \sigma(f_{j_1} \theta_{j_1} + f_{j_2} \theta_{j_2} + \cdots + f_{j_k} \theta_{j_k})$$
$$X^\top r_p = \nabla \log \text{pr}_\theta(y)$$

$$j_1 = \arg \min_j \left\| y - \sigma(f_j \hat{\theta}_j) \right\|_2^2$$

$$\approx \arg \max_j \left| f_j^\top r_0 \right|$$

$$j_2 = \arg \min_j \left\| y - \sigma(f_{j_1} \hat{\theta}_{j_1} - f_j \hat{\theta}_j) \right\|_2^2$$

$$\approx \arg \max_j \left| f_j^\top r_1 \right|$$

MATCHING PURSUIT FOR LOGISTIC REGRESSION

Matching pursuit feature selection rule [Lozano et al., 2011]

Assuming normalized data, i.e. $f_j^\top f_j = 1$, the OMP rule is given by

$$j_k = \arg \max_j \left| f_j^\top r_{k-1} \right|$$

where r_k are the k th pseudo-residuals

$$r_k = y - \sigma(f_{j_1} \theta_{j_1} + f_{j_2} \theta_{j_2} + \cdots + f_{j_k} \theta_{j_k})$$
$$X^\top r_p = \nabla \log \text{pr}_\theta(y)$$

MATCHING PURSUIT FOR LOGISTIC REGRESSION

Matching pursuit feature selection rule [Lozano et al., 2011]

Assuming normalized data, i.e. $f_j^\top f_j = 1$, the OMP rule is given by

$$j_k = \arg \max_j |f_j^\top r_{k-1}|$$

where r_k are the k th pseudo-residuals

$$r_k = y - \sigma(f_{j_1} \theta_{j_1} + f_{j_2} \theta_{j_2} + \cdots + f_{j_k} \theta_{j_k})$$
$$X^\top r_p = \nabla \log \text{pr}_\theta(y)$$

OMP Performance

Greedy strategy causes poor performance of Orthogonal Matching Pursuit in practice

LEAST ANGLE REGRESSION (LARS)

LEAST ANGLE REGRESSION (LARS)

- Consider ℓ_1 -penalized linear regression (LASSO) where

$$\hat{\theta}(\lambda) = \arg \min_{\theta} \|y - X\theta\|_2^2 + \lambda \|\theta\|_1$$

- There exists a regularization strength $\lambda = \lambda_{\max}$ for which all estimated coefficients are zero
- Least Angle Regression (LARS) [Efron et al., 2004] is a method to efficiently compute $\hat{\theta}(\lambda)$ for all $0 \leq \lambda \leq \lambda_{\max}$
- LARS computes breakpoints λ_k at which individual coefficients $\hat{\theta}_j(\lambda_k) \in \mathbb{R}$ change its value from
 - ▶ zero to non-zero, or from
 - ▶ non-zero to zero
- Between breakpoints the values of coefficients can be linearly interpolated

LEAST ANGLE REGRESSION (LARS)

- Remember that the OLS solution $\hat{\theta}(0)$ for $\lambda = 0$ requires that

$$\nabla_{\theta} \|y - X\theta\|_2^2 = 2X^T(y - X\theta) = 0$$

- For $\lambda > 0$ the solution requires

$$X^T(y - X\theta) \in \frac{\lambda}{2} \partial \|\theta\|_1$$

where $\partial \|\theta\|_1$ is the subgradient with respect to θ

- We define

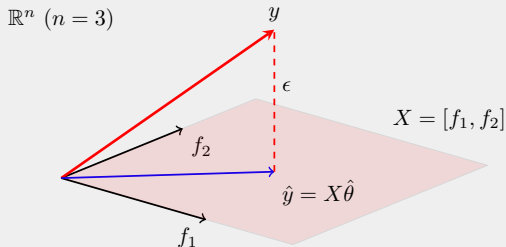
$$c(\theta) = X^T(y - X\theta)$$

which is interpreted as the correlation of features $X = [f_1, f_2, \dots, f_p]$ with the residuals $\epsilon = y - X\theta$

LEAST ANGLE REGRESSION (LARS)

■ The correlation $\hat{c}(\lambda) = c(\hat{\theta}(\lambda))$ varies with λ as follows:

- ▶ $\hat{c}(\lambda) = c_{\max}$ for $\lambda = \lambda_{\max}$
- ▶ $\hat{c}(\lambda) = 0$ for $\lambda = 0$



LEAST ANGLE REGRESSION (LARS)

- LARS maintains a set of active features $\Omega \subset \{1, \dots, p\}$ all equally correlated with the residuals $y - X\hat{\theta}(\lambda)$ for the current estimate $\hat{\theta}(\lambda)$
- Let $X_\Omega = (f_j)_{j \in \Omega}$ denote the covariate matrix and $\theta_\Omega = (\theta_j)_{j \in \Omega}$ the coefficients restricted to the features in the active set Ω
- In each iteration, the coefficients θ are updated

$$\theta \leftarrow \theta + \gamma^* \mathbf{v},$$

where γ^* is the amount by which the correlation $c_\Omega(\theta)$ is reduced and $\mathbf{v} \in \mathbb{R}^p$ defines the direction and relative size of the update

LEAST ANGLE REGRESSION (LARS)

- The vector v is selected so that for features in Ω the difference in correlation $c_{\Omega}(\theta) - c_{\Omega}(\theta + \gamma v)$ shrinks uniformly towards zero with rate γ , i.e.

$$c_{\Omega}(\theta) - c_{\Omega}(\theta + \gamma v) = \gamma \operatorname{sign} c_{\Omega}(\theta), \quad \text{while} \\ c_{\Omega^c}(\theta) - c_{\Omega^c}(\theta + \gamma v) = 0.$$

- Both conditions can be combined into

$$c(\theta) - c(\theta + \gamma v) = \gamma \operatorname{sign} c(\theta),$$

since $\operatorname{sign} c_{\Omega^c}(\theta) = 0$

- It follows that

$$v_{\Omega} = [X_{\Omega}^{\top} X_{\Omega}]^{-1} \operatorname{sign} c_{\Omega}(\theta)$$

and $v_{\Omega^c} = 0$

LEAST ANGLE REGRESSION (LARS)

- LARS stop shrinking the correlations whenever:
 - ▶ Case 1: A non-active feature becomes equally correlated with the residuals
 - ▶ Case 2: A coefficient of an active feature becomes zero¹
- Case 1: More formally, γ is increased until some feature $j' \in \Omega^c$ outside the active group satisfies

$$\begin{aligned} |c_{j'}(\theta + \gamma \mathbf{v})| &= |c_j(\theta + \gamma \mathbf{v})| \\ &= \lambda - \gamma, \end{aligned}$$

where $j \in \Omega$, and $\lambda = |c_j(\theta)|$ is the absolute correlation of the active features

¹This case was not part of the initial LARS algorithm but was later on added in order to ensure equivalence with the LASSO (see also Homotopy algorithm [Osborne et al., 2000])

LEAST ANGLE REGRESSION (LARS)

- The solution is given by

$$\gamma^+ = \min_{j \in \Omega^c}^+ \left\{ \frac{\lambda - c_j(\theta)}{1 - f_j^\top X v}, \frac{\lambda + c_j(\theta)}{1 + f_j^\top X v} \right\},$$

where \min^+ is the minimum over positive elements and note that $f_j^\top X v = f_j^\top X_\Omega v_\Omega$

- Case 2: The algorithm also removes a feature j from the active set when for some γ

$$\theta_j + \gamma v_j = 0$$

so that $\gamma^- = \min_{j \in \Omega} \{-\theta_j / v_j\}$

- The subsequent breakpoint is given by $\gamma^* = \min\{\gamma^+, \gamma^-\}$

SAFE AND STRONG RULES

Penalized regression

$$\omega(\theta) = -\log \text{pr}_{\theta}(\mathbf{y})$$

(logistic regression), or

$$\omega(\theta) = \|\mathbf{y} - \mathbf{X}\theta\|_2^2$$

(linear regression)

$$\hat{\theta} = \begin{cases} \arg \min_{\theta} & \omega(\theta) \\ \text{subject to} & \|\theta\|_1 = \Lambda \end{cases}$$

Basic idea: Select Λ such that $\|\theta\|_0 = m$

Numerical solution of penalized regression

Identify saddle points of Lagrangian

$$\mathcal{L}(\theta, \lambda) = \omega(\theta) + \lambda(\|\theta\|_1 - \Lambda)$$

Numerical solution of penalized regression

Identify saddle points of Lagrangian

$$\mathcal{L}(\theta, \lambda) = \omega(\theta) + \lambda(\|\theta\|_1 - \Lambda)$$

In practice the constraint $\|\theta\|_1 = \Lambda$ is ignored, but λ is chosen such that classification performance is optimal:

Penalized regression in practice

$$\hat{\theta}(\lambda) = \arg \min_{\theta} \omega(\theta) + \lambda \|\theta\|_1$$

SAFE RULE FOR LINEAR REGRESSION

SAFE rule: What features can we neglect for a fixed λ ?

SAFE RULE FOR LINEAR REGRESSION

SAFE rule: What features can we neglect for a fixed λ ?

SAFE rule [Ghaoui et al., 2010, Kim et al., 2007] for ℓ_1 -penalized linear regression

j th component of $\hat{\theta}$ must be zero if

$$|f_j^\top y| < \lambda - \|f_j\|_2 \|y\|_2 \frac{\lambda_{\max} - \lambda}{\lambda_{\max}}$$
$$\lambda_{\max} = \max_j |f_j^\top y|$$

SAFE RULE FOR LINEAR REGRESSION

SAFE rule: What features can we neglect for a fixed λ ?

SAFE rule [Ghaoui et al., 2010, Kim et al., 2007] for ℓ_1 -penalized linear regression

j th component of $\hat{\theta}$ must be zero if

$$|f_j^\top y| < \lambda - \|f_j\|_2 \|y\|_2 \frac{\lambda_{\max} - \lambda}{\lambda_{\max}}$$
$$\lambda_{\max} = \max_j |f_j^\top y|$$

$$|f_j^\top (y - \underbrace{X\theta}_{\theta=0})| < \lambda - \|f_j\|_2 \|y\|_2 \frac{\lambda_{\max} - \lambda}{\lambda_{\max}}$$

STRONG RULE FOR LINEAR REGRESSION

SAFE rule for linear regression: j th component of $\hat{\theta}$ must be zero if

$$|f_j^\top y| < \lambda - \|f_j\|_2 \|y\|_2 \frac{\lambda_{\max} - \lambda}{\lambda_{\max}}$$
$$\lambda_{\max} = \max_j |f_j^\top y|$$

Strong rule for ℓ_1 -penalized linear regression
[Tibshirani et al., 2012]

Discard j th component if

$$|f_j^\top y| < \lambda - (\lambda_{\max} - \lambda) = 2\lambda - \lambda_{\max}$$
$$\lambda_{\max} = \max_j |f_j^\top y|$$

STRONG RULE FOR LINEAR REGRESSION

Strong rule for ℓ_1 -penalized linear regression
[Tibshirani et al., 2012]

Discard j th component if

$$|\mathbf{f}_j^\top \mathbf{y}| < \lambda - (\lambda_{\max} - \lambda) = 2\lambda - \lambda_{\max}$$
$$\lambda_{\max} = \max_j |\mathbf{f}_j^\top \mathbf{y}|$$

Remark

Strong rule may drop features that should not be discarded \Rightarrow
KKT conditions must be checked, i.e.

$$\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\hat{\theta}) \in \lambda \partial_{\theta=\hat{\theta}} \|\theta\|_1$$

STRONG SEQUENTIAL RULE FOR LINEAR REGRESSION

Strong rule for ℓ_1 -penalized linear regression [Tibshirani et al., 2012]

Discard j th component if

$$|\mathbf{f}_j^\top \mathbf{y}| < \lambda - (\lambda_{\max} - \lambda) = 2\lambda - \lambda_{\max}$$
$$\lambda_{\max} = \max_j |\mathbf{f}_j^\top \mathbf{y}|$$

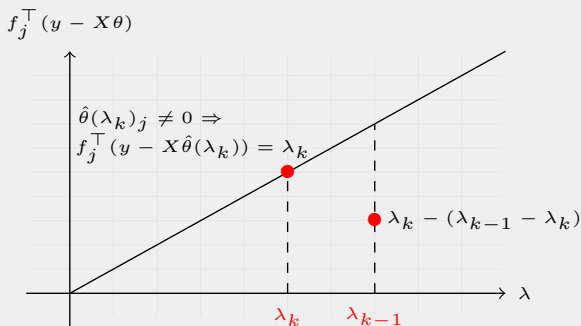
Strong sequential rule for ℓ_1 -penalized linear regression [Tibshirani et al., 2012]

Discard j th feature if

$$|\mathbf{f}_j^\top \{\mathbf{y} - \sigma(X\hat{\theta}(\lambda_{k-1}))\}| < 2\lambda_k - \lambda_{k-1}$$




STRONG SEQUENTIAL RULE FOR LINEAR REGRESSION

Compute $\hat{\theta}(\lambda_k)$ for all $\lambda_1 > \dots > \lambda_k > \dots > \lambda_K$







$$\text{Assumption : } |f_j^\top(y - X\hat{\theta}(\lambda_{k-1} - \epsilon)) - f_j^\top(y - X\hat{\theta}(\lambda_{k-1}))| \leq \epsilon$$
$$\Rightarrow |f_j^\top(y - X\hat{\theta}(\lambda_{k-1}))| < 2\lambda_k - \lambda_{k-1}$$

REFERENCES I

-  BOYD, S. AND VANDENBERGHE, L. (2004).
CONVEX OPTIMIZATION.
Cambridge university press.
-  DEFAZIO, A., BACH, F., AND LACOSTE-JULIEN, S. (2014).
SAGA: A FAST INCREMENTAL GRADIENT METHOD WITH SUPPORT FOR NON-STRONGLY CONVEX COMPOSITE OBJECTIVES.
Advances in neural information processing systems, 27.
-  EFRON, B., HASTIE, T., JOHNSTONE, I., TIBSHIRANI, R., ET AL. (2004).
LEAST ANGLE REGRESSION.
The Annals of statistics, 32(2):407–499.
-  FAN, J. AND LV, J. (2008).
SURE INDEPENDENCE SCREENING FOR ULTRAHIGH DIMENSIONAL FEATURE SPACE.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70(5):849–911.

REFERENCES II

-  FAN, J. AND SONG, R. (2010).
SURE INDEPENDENCE SCREENING IN GENERALIZED LINEAR MODELS WITH NP-DIMENSIONALITY.
The Annals of Statistics, 38(6):3567–3604.
-  GHAOUI, L. E., VIALON, V., AND RABBANI, T. (2010).
SAFE FEATURE ELIMINATION FOR THE LASSO AND SPARSE SUPERVISED LEARNING PROBLEMS.
arXiv preprint arXiv:1009.4219.
-  HASTIE, T., TIBSHIRANI, R., AND FRIEDMAN, J. (2009).
THE ELEMENTS OF STATISTICAL LEARNING: DATA MINING, INFERENCE, AND PREDICTION.
Springer Science & Business Media.
-  KIM, S.-J., KOH, K., LUSTIG, M., BOYD, S., AND GORINEVSKY, D. (2007).
AN INTERIOR-POINT METHOD FOR LARGE-SCALE ℓ_1 -REGULARIZED LOGISTIC REGRESSION.
In *Journal of Machine learning research*. Citeseer.

REFERENCES III

-  LOZANO, A., SWIRSZCZ, G., AND ABE, N. (2011).
GROUP ORTHOGONAL MATCHING PURSUIT FOR LOGISTIC REGRESSION.
In Proceedings of the fourteenth international conference on artificial intelligence and statistics, pages 452–460. JMLR Workshop and Conference Proceedings.
-  OSBORNE, M. R., PRESNELL, B., AND TURLACH, B. A. (2000).
A NEW APPROACH TO VARIABLE SELECTION IN LEAST SQUARES PROBLEMS.
IMA journal of numerical analysis, 20(3):389–403.
-  TIBSHIRANI, R., BIEN, J., FRIEDMAN, J., HASTIE, T., SIMON, N., TAYLOR, J., AND TIBSHIRANI, R. J. (2012).
STRONG RULES FOR DISCARDING PREDICTORS IN LASSO-TYPE PROBLEMS.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 74(2):245–266.

REFERENCES IV



TROPP, J., GILBERT, A. C., ET AL. (2007).

**SIGNAL RECOVERY FROM PARTIAL INFORMATION VIA ORTHOGONAL
MATCHING PURSUIT.**

IEEE Trans. Inform. Theory, 53(12):4655–4666.

DERIVATION OF THE SAFE RULE FOR LINEAR REGRESSION

$$\hat{\theta} = \arg \min_{\theta} \|y - X\theta\|_2^2 + \lambda \|\theta\|_1$$

Define

$$\beta = y - X\theta$$

Equivalent optimization problem

$$\hat{\theta} = \begin{cases} \arg \min_{\theta} & \beta^\top \beta + \lambda \|\theta\|_1 \\ \text{subject to} & \beta = y - X\theta \end{cases}$$

DERIVATION OF THE SAFE RULE FOR LINEAR REGRESSION

Lagrangian

$$\mathcal{L}(\theta, \beta, \nu) = \beta^\top \beta + \lambda \|\theta\|_1 + \nu^\top (\mathbf{y} - \mathbf{X}\theta - \beta)$$

Dual function

$$\inf_{\theta, \beta} \mathcal{L}(\theta, \beta, \nu) = \begin{cases} G(\nu) & \text{if } |f_j^\top \nu| \leq \lambda, j = 1, \dots, p \\ -\infty & \text{otherwise} \end{cases}$$

where $G(\nu) = -\frac{1}{4}\nu^\top \nu + \nu^\top \mathbf{y}$. Lagrange dual

$$\hat{\theta}^* = \begin{cases} \arg \max_{\nu} & G(\nu) \\ \text{subject to} & |f_j^\top \nu| \leq \lambda, j = 1, \dots, p \end{cases}$$

DERIVATION OF THE SAFE RULE FOR LINEAR REGRESSION

Side note: Since the primal problem satisfies Slater's condition, we know that the duality gap $\gamma = \hat{\theta} - \hat{\theta}^*$ is zero, i.e.

$$\hat{\theta} = \hat{\theta}^*$$

For a dual feasible point ν_0 , we solve for each $j = 1, \dots, p$

$$\begin{aligned}\xi_j(\nu_0) &= \begin{cases} \arg \max_{\nu} & |f_j^\top \nu| \\ \text{subject to} & G(\nu) \geq G(\nu_0) \end{cases} \\ &= |f_j^\top y| + \sqrt{(y^\top y - 2G(\nu_0))f_j^\top f_j}\end{aligned}$$

If $\xi_j(\nu_0) < \lambda$ we know that $\hat{\theta}_j = 0$. A simple dual feasible point is $\nu_0 = y\lambda/\lambda_{\max}$. The SAFE rule is obtained from

$$\xi_j(y\lambda/\lambda_{\max}) < \lambda$$

LOGISTIC REGRESSION CLASSIFIER

SAGA algorithm [Defazio et al., 2014]: select $j \in \{1, \dots, n\}$ at random

$$\vartheta_{j,t+1} = \theta_t$$

$$\vartheta_{i,t+1} = \vartheta_{i,t+1} \text{ for all } i \neq j$$

$$\theta_{t+1}^* = \theta_t - \gamma \left[\nabla \ell_j(\vartheta_{j,t+1}) - \nabla \ell_j(\vartheta_{j,t}) + \frac{1}{n} \sum_{i=1}^n \nabla \ell_i(\vartheta_{i,t}) \right]$$

$$\theta_{t+1} = \arg \min_{\theta} \left\{ \lambda \|\theta\|_1 + \frac{1}{2\gamma} \|\theta - \theta_{t+1}^*\|_2^2 \right\}$$