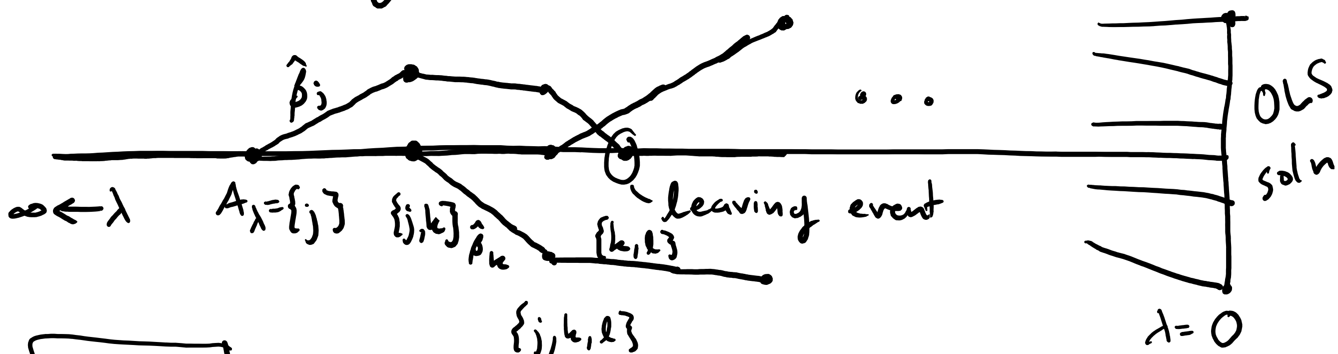
More about the Lasso

Wednesday, April 19, 2017 5:47 PM

Recall the lasso: $\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$

Define $A_\lambda = \text{supp}(\hat\beta_\lambda) = \{j : \hat\beta_{\lambda j} \neq 0\}$.

As $\lambda \to \infty$, $\beta \to 0$ so start at $\lambda = \infty$ and eventually as λ decreases $A_\lambda \neq \{\}$.



$\boxed{\text{fact 1}}$ $\hat\beta_j$ is continuous, piecewise linear in λ

$\boxed{\text{fact 2}}$ Slope of $\hat\beta_j$ is regression coef on residual for OLS on A_λ.

$\boxed{\text{fact 3}}$ LAR with lasso modification solves the lasso.

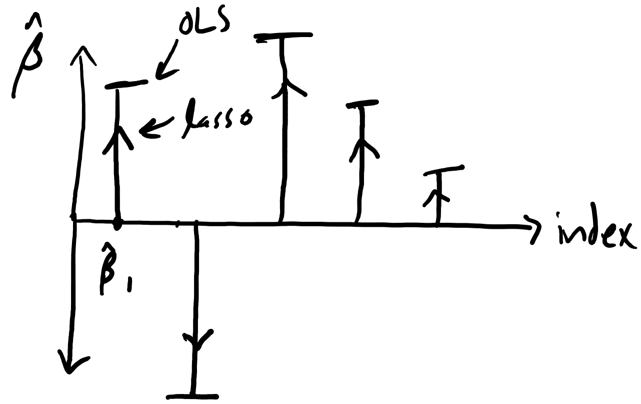
We can think of the lasso path as a sequence of models $A_{\lambda_1}, A_{\lambda_2}, \ldots$ for $\lambda_1 > \lambda_2 \ldots$ knots.

<u>Lasso introduces a bias</u>

So, for selected

$\hat\beta \uparrow \hat\beta^{OLS}$ T

model A ...



so, for selected model A, solve restricted OLS,

$$\tilde\beta = (X_A^T X_A)^{-1} X_A^T y$$

Clarification: $\beta_j = \beta_{j+} - \beta_{j-}$ for $\beta_{j+}, \beta_{j-} \geq 0$

QP: $$\min_{\beta_+, \beta_-} \|y - X(\beta_+ - \beta_-)\|_2^2 \quad \text{s.t.} \quad \sum_j \beta_{j+} + \beta_{j-} \leq C$$

$$\beta_{j+}, \beta_{j-} \geq 0$$

if β_+, β_- feasible then $\sum_j |\beta_j| \leq \sum_j \beta_{j+} + \beta_{j-} \leq C$

for $\beta := \beta_+ - \beta_-$ $\Rightarrow$ QP $\geq$ Lasso

if β is Lasso sol$^{\underline{n}}$ $\beta_{j+} = (\beta_j)_+$, $\beta_{j-} = (-\beta_j)_+$

$\sum_j \beta_{j+} + \beta_{j-} = \sum_j |\beta_j| \leq C$ $\Rightarrow$ QP $\leq$ Lasso

Logistic Regression

Wednesday, April 19, 2017 5:47 PM

Recall Empirical Risk Minimization:

$$\min_{\beta \in D} \frac{1}{n}\sum_{i=1}^{n} l(y_i, x_i; \beta_i) \quad \text{for some loss, } l.$$

Regression: $l(y_i, x_i; \beta_i) = (y_i - x_i^T \beta)^2$.

Gaussian error model: $Y = X^T\beta + \varepsilon_i$

for $\varepsilon_i \sim N(0, \sigma^2)$ then density is

$$f_{Y|X}(y_i|\beta, x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - x_i^T\beta)^2}$$

$$-2\sigma^2 \log f_{Y|X}(y_i|\beta, x_i) = (y_i - x_i^T\beta)^2 + C$$

Maximum likelihood is empirical risk minimization when the loss is the negative log-likelihood (under iid).

<u>Logistic model</u> Classification ($Y \in \{0,1\}$)

$Y|X$ is Binomial (not much choice there)

but how does $\mathbb{P}\{Y=1 \mid X=x\}$ depend on x?

The logistic model assumes that

$$\log \frac{\mathbb{P}\{Y=1 \mid X=x\}}{\mathbb{P}\{Y=0 \mid X=x\}} = x^T\beta .$$

Logit function is $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$, so,

$$\text{logit}\left(\mathbb{P}\{Y=1 \mid X=x\}\right) = x^T\beta .$$

Claim: $\text{logit}^{-1}(a) = \frac{e^a}{1+e^a}$

$e^{\text{logit}(p)} = \frac{p}{1-p}$ so

$$\text{logit}^{-1}(\text{logit}(p)) = \frac{\frac{p}{1-p}}{1+\frac{p}{1-p}} = p \checkmark$$

$$\Rightarrow \mathbb{P}\{Y=1 \mid X=x\} = \frac{e^{x^T\beta}}{1+e^{x^T\beta}}$$

Also, $\mathbb{P}\{Y=0 \mid X=x\} = \frac{1}{1+e^{x^T\beta}}$ so

$$\log \mathbb{P}\{Y=y \mid X=x\} = y x^T\beta - \log\left(1+e^{x^T\beta}\right)$$

and the loss (neg. log-likelihood) is

$$\ell(y, x; \beta) = -y x^T\beta + \log\left(1+e^{x^T\beta}\right)$$

note: in ESL they maximize log-likelihood.

Logistic Regression (should be called Logistic Classification)

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} -y_i x_i^T \beta + \log\left(1 + e^{x_i^T \beta}\right)$$

Fitting Logistic Regression

Wednesday, April 19, 2017 8:47 PM

$$\ell(y,x;\beta) = -yx^T\beta + \log\left(1+e^{x^T\beta}\right)$$

$$\frac{\partial}{\partial\beta}\ell(y,x;\beta) = -yx + \frac{e^{x^T\beta}\cdot x}{1+e^{x^T\beta}} = -yx + \text{logit}^{-1}(x^T\beta)\cdot x$$

$$= (p-y)x \quad \text{if } p = \mathbb{P}\{Y=1 \mid X=x,\beta\}$$

$$\frac{\partial^2}{\partial\beta\partial\beta^T}\ell(y,x;\beta) = \frac{e^{x^T\beta}xx^T}{1+e^{x^T\beta}} - \frac{e^{2x^T\beta}xx^T}{(1+e^{x^T\beta})^2} = \frac{e^{x^T\beta}}{(1+e^{x^T\beta})^2}xx^T$$

$$= p(1-p)xx^T \succeq 0 \quad \text{so } \ell \text{ is convex!}$$

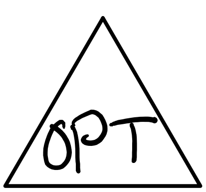
Empirical risk: $R_n(\beta) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, x_i;\beta)$

$$\frac{\partial}{\partial\beta}R_n(\beta) = \frac{1}{n}\sum_{i=1}^n (p_i-y_i)x_i = \frac{1}{n}X^Tr \quad \text{where}$$

$$p_i = \text{logit}^{-1}(x^T\beta), \quad r_i = p_i - y_i$$

$$\frac{\partial^2}{\partial\beta\partial\beta^T}R_n(\beta) = \frac{1}{n}\sum_{i=1}^n p_i(1-p_i)x_ix_i^T = \frac{1}{n}X^TWX$$

$$W_{ii} = p_i(1-p_i)$$



Newton-Raphson

until convergence criteria

$$\beta_{t+1} \leftarrow \beta_t + H^{-1}g$$

Hessian $\uparrow$ gradient at β_t

Idea: Approximate $R_n(\beta)$ by local quadratic

$$R_n(\beta) \approx R_n(\beta_t) + g^T(\beta-\beta_t) + \frac{1}{2}(\beta-\beta_t)^T H(\beta-\beta_t)$$

$\hookrightarrow$ argmin $= \beta_t + H^{-1} g$.

Logistic: $H^{-1} g = \underbrace{(X^T W X)^{-1} X^T r}_{\text{weighted least squares}}$

$$\boxed{\text{Newton Raphson} \longrightarrow \text{iteratively re-weighted least squares}}$$