Gradient Methods

Wednesday, May 24, 2017 8:29 PM

$$\min_{\beta \in \mathbb{R}^p} R(\beta)$$

Directional descent master algorithm

Until stopping criteria:
- choose descent direction u_t
- choose step size η_t
- update $\beta_{t+1} \leftarrow \beta_t + \eta_t u_t$

Suppose R is convex, when we choose u_t,

$\min_{\eta \in \mathbb{R}} R(\beta_t + \eta u_t)$ is 1-D & convex

Performing this min is <u>line search</u>

<u>Interval Bisection</u>

$a_0 = L$, $b_0 = u$

while $(b_t - a_t) \cdot R'(u) > \varepsilon$:

 if $R'\left(\frac{a_t + b_t}{2}\right) > 0$ then

  $a_{t+1} = a_t$ $b_{t+1} = \frac{a_t + b_t}{2}$

 else

  $a = a_t + b_t$ ...

$$\frac{\quad}{2} \quad \beta_{t+1} - \beta_t$$

$t \leftarrow t+1$

Other step size selection:

- backtracking line search
- η_t decay according to a schedule

  e.g. $\eta_t = \frac{1}{\sqrt{t}}$

  usually schedule is chosen according to prop.s of R.

  eg. Lipschitz cont grad. w/ modulus $L \to \eta_t = \frac{1}{L}$

Coordinate descent

Until stopping crit:

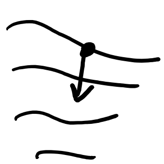
  select coord. j, $u_t = -e_j$ ✱

  set $\eta_t = \underset{\eta \in \mathbb{R}}{\operatorname{argmin}} R(\beta_t + \eta u_t)$

  update $\beta_{t+1} \leftarrow \beta_t + \eta_t u_t$

✱ j is selected either greedy, random, sequential

Gradient descent



update $\beta_{t+1} \leftarrow \beta_t - \eta_t \nabla R(\beta_t)$

$u_t = -\nabla R(\beta_t)$

ERM

$$R_n(\beta_t) = \frac{1}{n} \sum_{i=1}^{n} l(y_i, x_i, \beta_t)$$

(could add regularizer to this)

$$\nabla_\beta R_n(\beta_t) = \frac{1}{n}\sum_{i=1}^{n} \nabla_\beta l(y_i, x_i, \beta_t)$$

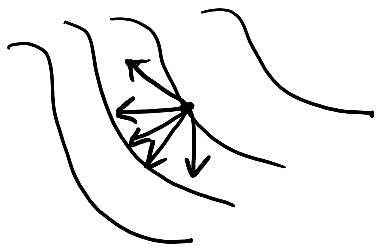
Stochastic Gradient Descent

Until Stopping Crit:

$S \leftarrow$ Subsample $\{1, \dots, n\}$ (mini batch)

$$u_t = -\frac{1}{|S|}\sum_{i \in S} \nabla_\beta l(y_i, x_i, \beta_{t-1})$$

$$\beta_{t+1} \leftarrow \beta_t + \eta_t u_t$$



Stochastic Gradient and Online Learning

Wednesday, May 24, 2017 11:11 PM

Online learning

See sample x_t

Predict $\hat{y}_t$

See truth y_t

Incur loss $\ell(y_t, \hat{y}_t)$

SGD w/ single sample

while :

$i \leftarrow$ sample $1, \dots, n$

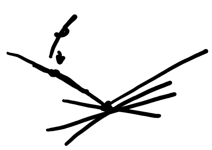
$u_t = -\nabla_\beta \ell(y_i, x_i, \beta_t)$

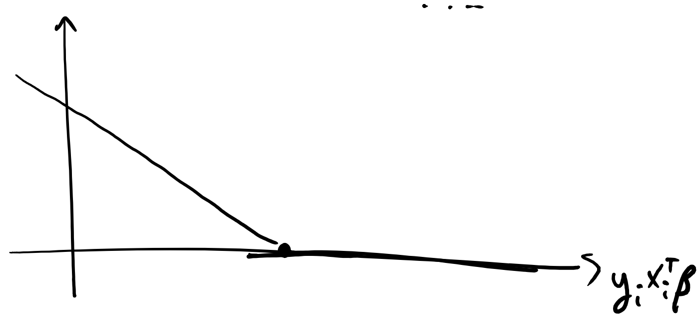
$\beta_{t+1} \leftarrow \beta_t + \eta_t u_t$

Apply to SVM :

Objective: $\frac{1}{n} \sum_{i=1}^{n} (1 - y_i x_i^T \beta)_+ + \lambda \|\beta\|_2^2$

Subgradient: $\frac{\partial}{\partial \beta} (1 - y_i x_i^T \beta)_+ \leftarrow \begin{cases} -y_i x_i & 1 - y_i x_i^T \beta > 0 \\ 0 & 1 - y_i x_i^T \beta = 0 \\ 0 & \dots \end{cases}$

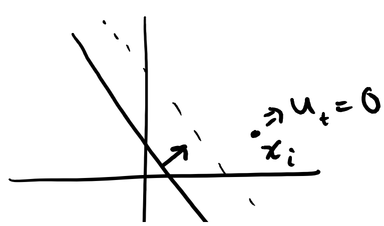




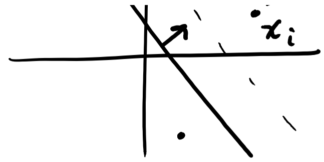
$\beta = 0$

For $i = 1, \dots, n$

$u_t = \begin{cases} y_i x_i & \text{if } y_i x_i^T \beta < 1 \\ 0 & \text{otherwise} \end{cases}$



$$u_t = \begin{cases} y_i x_i & \text{if } \; y_i x_i \beta < 1 \\ 0 & \text{otherwise} \end{cases}$$



$u_t = x_i$

$y_i = 1$

$-\lambda\beta$ (for ridge term)

$$\beta \leftarrow \beta + \eta_t \left(y_i x_i \, 1\{ y_i x_i^T \beta < 1 \} - 2\lambda\beta \right)$$

$$= (1 - \eta_t \lambda)\beta + \eta_t y_i x_i \, 1\{ y_i x_i^T \beta < 1 \}$$

$\equiv$ linear perceptron