Lecture November 12

Ensemble methods

- Decision tree: simple but leads often to overfitting (high variance)

- Ensemble methods
  - Bagging (forests of same trees)
  - Random forests (forests of different trees)
  - Boosting: simple method (weak method) which is improved upon iteratively.

- models defined by

$$b_m(x) \in \mathbb{R}^n$$

$$f(x) = \sum_{m=0}^{M} \beta_m b_m(x)$$

- $h_m(x) = X_m \quad m = 1, \dots p$
- or some non-linear functions.

Three common approaches

- Restriction models

$$f(x) = \sum_{j=1}^{p} f_j(x_j)$$

$$= \sum_{j=1}^{p} \sum_{m=0}^{M_j} \beta_{jm} b_{jm}(x_j)$$

basis functions
size of the model
limited by the
set of basis
function

- selection model
  include only basis
  functions h_m that
  contribute significantly
  to fit of the model.

we end often up with
a simple decision tree

- Regularization
  Ridge is pure regularization
  Lasso is regularization
  and selection.

Boosting belongs to
selection models

$$f(x) = f_0(x) + \sum_{m=1}^{M} \beta_m b_m(x)$$

in boosting

$$\min_{f} \sum_{i=0}^{m-1} \mathcal{L}(y_i, f(x_i))$$

MSE

$$\hat{f} = \arg\min_{f(x)} \mathbb{E}\left[(y - f(x))^2\right]$$

Regression case : L2 boosting

(i) Establish a cos function

$$C(f) = \frac{1}{n} \sum_{i=0}^{n-1} (y_i - f(x_i))^2$$

(ii) $f_m(x) = f_{m-1}(x) + \beta_m b_m(x, \gamma)$

(iii) initialize $f_0(x)$

(iv) for $m = 1 : M$

$$(\hat{\beta}_m, \hat{\gamma}_m) = \underset{\gamma, \beta}{\arg\min} \frac{1}{n} \sum_{i=0}^{n-1} (y_i - f_{m-1}(x_i) - \beta b_m(x_i; \gamma))^2$$

(v) gives $\hat{\beta}_m, \hat{\gamma}_m$

(vi) determine

$$f_m(x) = f_{m-1}(x) + \beta_m b(x; \gamma_m)$$

end

Return $f_M(x)$

Example

$$f_0(x) = 0$$

$$b(x;\gamma) \quad 1+\gamma x$$

$$(\hat{\beta}_m, \hat{\gamma}_m) = \underset{\gamma\, \beta}{\arg\min}$$

$$\frac{1}{n}\sum_{i=0}^{n-1} \underbrace{(y_i - f_{m-1}(x_i) - \beta b(x_i;\gamma))^2}_{(y_i - f_{m-1}(x_i) - \beta(1+\gamma x_i))^2}$$

$$\hat{\beta}_1 : \quad \frac{\partial C}{\partial \beta} = -\frac{2}{n}\sum_i (1+\gamma x_i)(y_i - \beta(1+\gamma x_i)) = 0$$

$$\hat{\gamma}_1 : \quad \frac{\partial C}{\partial \gamma} = -\frac{2}{n}\sum_i \beta x_i (y_i - \beta(1+x_i\gamma)) = 0$$

$$m = 1 \qquad f_0(x) = 0$$

$$f_m(x) = f_{m-1}(x) + \beta_m b_m(x, \gamma_m)$$

$$f_1(x) = \beta_1 (1 + \gamma_1 x)$$

repeat till $m = M$,

return $f_M(x)$

- Classification

Adaboost: puts an emphasis on misclassified events.

start with giving all classifications the same weight $w_i = \frac{1}{N}$

$$f_M(x) = \text{classifier}$$

$$f_m(x) = f_{m-1}(x) + \beta_m f_m(x; \gamma_m)$$

$$f_m(x) = \{-1, 1\}$$

$$f(x) = \text{sign}\left(\sum_{m=0}^{M} f_m(x)\right)$$

$i = 0, 1, \ldots, n-1$

$f_0(x) =$ initial guess

for $m = 1 : M$

- fit a classifier $f_m(x)$

- compute error

$$err_m = \frac{\sum_{i=0}^{n-1} w_i I(y_i \neq f_m(x_i))}{\sum_{i=0}^{n-1} w_i}$$

- find the parameter $\gamma_m, \beta_m \ldots$

- update weights

end for

output final $f_M(x)$

Adaboost uses the following cost function

$$C(f) = \sum_{i=0}^{n-1} \mathcal{L}(y_i, f(x_i))$$

$$= \sum^{n-1} \exp(-y_i f(x_i))$$

$i = 0$

$$f_m(x) = f_{m-1}(x) + \beta_m \cancel{b_m}(x, \gamma_m)$$

$$f_m(x) = \{-1, +1\}$$

if correctly classified

$$y_i = f_m(x_i) = \pm 1$$

$$e^{-y_i f_m(x_i)} = e^{-1}$$

if wrongly classified

$$y_i \neq f_m(x_i)$$

$$e^{+1}$$

Finding weights:

$$f_m(x) = f_{m-1}(x) + \beta_m \underbrace{b_m(x; \gamma_m)}_{g_m(x)}$$

$$\hat{\beta}_m, \hat{g}_m = \arg\min_{\beta\ g}$$

$$\sum^{n-1}$$

$$\sum_{i=0} \exp\left[-y_i\left(f_{m-1}(x_i) + \beta g(x_i)\right)\right]$$

$$= \sum_i w_i^{(m)} \exp\left(-y_i \beta g(x_i)\right)$$

$$w_i^{(m)} = e^{-y_i f_{m-1}(x_i)}$$

Think of this as a weight applied to each observation.

$$\sum_i w_i^{(m)} e^{-\beta y_i g(x_i)}$$

$$= e^{-\beta} \sum_i w_i^{(m)} +$$

$$\left(e^{-\beta} + e^{\beta}\right) \sum_{i=0}^{n-1} w_i^{(m)} I(y_i \neq g(x_i))$$

solve for β

$$err_m = \sum_i w_i^{(m)} I(y_i \neq g(x_i))$$

$$i \qquad \frac{\qquad}{\sum_i w_i^{(m)}}$$

$$\Rightarrow \frac{e^{-\beta}}{e^{-\beta}+e^{\beta}} = err_m$$

$$\beta_m = \frac{1}{2} \log \frac{1-err_m}{err_m}$$

$$f_m(x) = f_{m-1}(x) + \beta_m G_m(x_i)$$

$$w_i^{(m+1)} = w_i^{(m)} e^{-\beta_m y_i g_m(x_i)}$$