Lecture 5
Ensembles – Part 2

(February 6, 2015)

Mu Zhu
University of Waterloo

Outline for Part 2

- gradient boosting machine (2001)
- Breiman's theorem (2001)
- some new directions (2005+)
  - variable selection ensembles (2006, 2010, 2011, 2012)
  - machine learning markets (2011)
  - magging (2014)

General Framework for Boosting

1. Pick a functional class, $\mathcal{F}$. (WHERE)
2. Pick a loss function, $L[y, F(\boldsymbol{x})]$. (WHAT)
3. Pick an optimization method. (HOW)
4. Start with $F(\boldsymbol{x}) = 0$, and sequentially build up the function,

$$F(\boldsymbol{x}) = f_1(\boldsymbol{x}) + ... + f_B(\boldsymbol{x}),$$

by repeating the following steps:

(a) apply the chosen optimization method from (3) to find

$$f \quad = \quad \arg\min_{g \in \mathcal{F}} \quad L\left[y, F(\boldsymbol{x}) + g(\boldsymbol{x})\right];$$

(b) update $F \longleftarrow F + f$.

Gradient Boosting Machine

- Friedman (2001; *Ann. Stat.*)
- free to choose $L(y, F)$
- free to choose $\mathcal{F}$ in principle, but mostly trees in practice
- choice of optimization method = take one gradient step, i.e.,

$$f = f_0 - \eta \left[\frac{\partial L}{\partial f} \right]_{f=f_0},$$

where $\eta > 0$ is step size and f_0 is initial "estimate" of f

Gradient Boosting Machine

- given current F, initial f_0 for the "next" function $= 0$
- one gradient step requires setting $f \propto -(\partial L/\partial f)|_{f=0}$
- but must do so within the constraint, $f \in \mathcal{F}$, so

$$\text{let} \quad r_i \equiv h(\boldsymbol{x}_i, y_i), \quad \text{where} \quad h \equiv -(\partial L/\partial f)|_{f=0}$$

$$\text{train} \quad f \quad = \quad \arg\min_{g \in \mathcal{F}} \quad \sum_{i=1}^{n} [r_i - g(\boldsymbol{x}_i)]^2$$

$$\text{set} \quad \eta \quad = \quad \arg\min_{\xi} \quad \sum_{i=1}^{n} L\left[y_i, F(\boldsymbol{x}_i) + \xi f(\boldsymbol{x}_i)\right]$$

- negative functional gradient, $h = -(\partial L/\partial f)|_{f=0}$, next slide

Ex: GBM with L_2 Loss Function

- consider regression instead of classification, i.e., $y_i \in \mathbb{R}$
- natural to choose

$$L\left[y, F(\boldsymbol{x}) + f(\boldsymbol{x})\right] = \frac{1}{2}\left[y - F(\boldsymbol{x}) - f(\boldsymbol{x})\right]^2$$

- then,

$$h = -\left[\frac{\partial L}{\partial f}\right]_{f=0} = \left(y - F - f\right)\big|_{f=0} = y - F$$

- thus, negative gradient = residual of the current model F, so

$$r_i = h(\boldsymbol{x}_i, y_i) = y_i - F(\boldsymbol{x}_i)$$

- so, repeatedly fit a model $\in \mathcal{F}$ to residuals of current model

Breiman's Theorem

$$\varepsilon_{RF} \quad \leq \quad \bar{\rho}\left(\frac{1-s^2}{s^2}\right)$$

- technically, requires the assumption that $s > 0$
- that is, an "average" classifier in the forest will correctly classify an "average" observation — not a very strong assumption
- next, elucidate the proof in three steps

Step 1

- target is (details later; requires $s > 0$)

$$\varepsilon_{RF} \equiv P_{(\boldsymbol{x},y)}(M(\boldsymbol{x},y) < 0) \leq \frac{\mathbb{V}\text{ar}_{(\boldsymbol{x},y)}(M(\boldsymbol{x},y))}{\mathbb{E}^2_{(\boldsymbol{x},y)}(M(\boldsymbol{x},y))}$$

- recall definition

$$s = \mathbb{E}_\Theta\left[\mathbb{E}_{(\boldsymbol{x},y)}(m(\Theta;\boldsymbol{x},y))\right] = \mathbb{E}_{(\boldsymbol{x},y)}\left[\mathbb{E}_\Theta(m(\Theta;\boldsymbol{x},y))\right]$$
$$= \mathbb{E}_{(\boldsymbol{x},y)}(M(\boldsymbol{x},y)),$$

  so denominator already s^2

- thus, after this step, suffices to prove numerator $\leq \bar{\rho}(1 - s^2)$

Step 2

- target is (again, details later)

$$\mathbb{V}\mathrm{ar}_{(\boldsymbol{x},y)}(M(\boldsymbol{x},y)) = \mathbb{E}_{\Theta,\Theta'}\left(\mathbb{C}\mathrm{ov}_{(\boldsymbol{x},y)}[m(\Theta;\boldsymbol{x},y), m(\Theta';\boldsymbol{x},y)]\right),$$

for $\Theta, \Theta' \overset{iid}{\sim} P_\Theta$

- definition of $\bar{\rho}$ + a (trivial) identity for $\Theta, \Theta' \overset{iid}{\sim} \mathcal{P}_\Theta$,

$$\mathbb{E}_{\Theta,\Theta'}(g(\Theta)g(\Theta')) = \mathbb{E}_\Theta(g(\Theta))\mathbb{E}_{\Theta'}(g(\Theta')) = \mathbb{E}^2_\Theta(g(\Theta)),$$

+ Jensen's inequality $\Rightarrow$

$$\begin{aligned}\mathbb{V}\mathrm{ar}_{(\boldsymbol{x},y)}(M(\boldsymbol{x},y)) &= \bar{\rho}\,\mathbb{E}^2_\Theta\left(\mathrm{Sd}_{(\boldsymbol{x},y)}m(\Theta;\boldsymbol{x},y)\right)\\ &\leq \bar{\rho}\,\mathbb{E}_\Theta\left(\mathrm{Sd}^2_{(\boldsymbol{x},y)}m(\Theta;\boldsymbol{x},y)\right) = \bar{\rho}\,\mathbb{E}_\Theta\left(\mathbb{V}\mathrm{ar}_{(\boldsymbol{x},y)}m(\Theta;\boldsymbol{x},y)\right)\end{aligned}$$

Step 3

- after step 2, remains to prove (i.e., target for step 3)

$$\mathbb{E}_{\Theta}\left(\mathbb{V}\mathrm{ar}_{(\boldsymbol{x},y)}\left(m(\Theta;\boldsymbol{x},y)\right)\right) \leq (1-s^2)$$

- simply write $m = m(\Theta;\boldsymbol{x},y)$, and get

$$\begin{aligned}\mathbb{E}_{\Theta}\left[\mathbb{V}\mathrm{ar}_{(\boldsymbol{x},y)}\left(m\right)\right] &= \mathbb{E}_{\Theta}\left[\mathbb{E}_{(\boldsymbol{x},y)}\left(m^2\right) - \mathbb{E}^2_{(\boldsymbol{x},y)}\left(m\right)\right] \\ &= \mathbb{E}_{\Theta}\left[\mathbb{E}_{(\boldsymbol{x},y)}\left(m^2\right)\right] - \mathbb{E}_{\Theta}\left[\mathbb{E}^2_{(\boldsymbol{x},y)}\left(m\right)\right] \equiv (\mathrm{I}) - (\mathrm{II})\end{aligned}$$

but

$$\begin{aligned}(\mathrm{I}) &= 1, \quad \text{since } m = \pm 1 \Rightarrow m^2 \equiv 1; \\ (\mathrm{II}) &= \mathbb{E}_{\Theta}\left(\mathbb{E}^2_{(\boldsymbol{x},y)}\left(m\right)\right) \geq \mathbb{E}^2_{\Theta}\left(\mathbb{E}_{(\boldsymbol{x},y)}\left(m\right)\right) = s^2,\end{aligned}$$

again, by Jensen's inequality

Step 1: Details

- (Lemma) If X is a random variable with $\mu \equiv \mathbb{E}(X) > 0$, then

$$P(X < 0) \leq \frac{\mathbb{V}\text{ar}(X)}{\mu^2}.$$

- Lemma above can be proved by simply taking $\epsilon = \mu$ in Chebyshev's inequality,

$$P(|X - \mu| \geq \epsilon) \leq \frac{\mathbb{V}\text{ar}(X)}{\epsilon^2}, \quad \text{for any} \quad \epsilon > 0.$$

- Step 1 can be completed by simply applying this Lemma to $X = M(\boldsymbol{x}, y)$, which requires $s = \mathbb{E}_{(\boldsymbol{x},y)}(M(\boldsymbol{x}, y)) > 0$.

Step 2: Details

- shorthand: $m = m(\Theta; \boldsymbol{x}, y)$; $m' = m(\Theta'; \boldsymbol{x}, y)$
- if $\Theta, \Theta' \overset{iid}{\sim} \mathcal{P}_\Theta$, $\mathbb{E}_\Theta(m) = \mathbb{E}_{\Theta'}(m')$ [trivial]
- from the definition of M, get

$$
\begin{aligned}
& \mathbb{V}\mathrm{ar}_{(\boldsymbol{x},y)}(M(\boldsymbol{x},y)) \\
= \; & \mathbb{E}_{(\boldsymbol{x},y)}(M^2(\boldsymbol{x},y)) - \mathbb{E}^2_{(\boldsymbol{x},y)}(M(\boldsymbol{x},y)) \\
= \; & \mathbb{E}_{(\boldsymbol{x},y)}(\mathbb{E}^2_\Theta[m]) - \mathbb{E}^2_{(\boldsymbol{x},y)}(\mathbb{E}_\Theta[m]) \\
= \; & \mathbb{E}_{(\boldsymbol{x},y)}(\mathbb{E}_\Theta[m]\mathbb{E}_{\Theta'}[m']) - \mathbb{E}_{(\boldsymbol{x},y)}(\mathbb{E}_\Theta[m])\mathbb{E}_{(\boldsymbol{x},y)}(\mathbb{E}_{\Theta'}[m']) \\
= \; & \mathbb{E}_{\Theta,\Theta'}(\mathbb{E}_{(\boldsymbol{x},y)}[mm']) - \mathbb{E}_{\Theta,\Theta'}(\mathbb{E}_{(\boldsymbol{x},y)}[m]\mathbb{E}_{(\boldsymbol{x},y)}[m']) \\
= \; & \mathbb{E}_{\Theta,\Theta'}\left(\mathbb{C}\mathrm{ov}_{(\boldsymbol{x},y)}[m,m']\right)
\end{aligned}
$$

Diversity and Ensembles

- importance of diversity
- power of ensembles
- testimony
  - high-profile, million-dollar Netflix contest (2006-09)
  - prediction contests on `www.kaggle.com`

Some New Directions

1. variable selection ensembles
   - use ensembles for a problem other than prediction
2. machine learning markets
   - how to aggregate other than taking averages
3. magging
   - how to aggregate, but for a different problem

<u>disclaimer</u>

selection of directions purely dictated by *personal taste*

Variable Selection Ensembles

Idea

- standard variable selection algorithms all deterministic
- do multiple runs
  - use a stochastic mechanism to "force" standard algorithms to give different answers for each run
- aggregate the results

Challenges

- delicate balance between false positives and false negatives
- characterizing/finding the "right" stochastic mechanism

Stochastic Mechanisms

	Stochastic Mechanism	Type
Z&C	Darwinian evolution in parallel universes	SO
M&B	Bootstrap + random scaling	Bt/Sub
W+	Bootstrap + random subset	Bt/Sub
X&Z	Stochastic stepwise search	SO

SO = stochastic optimization; Bt/Sub = bootstrap and/or sub-sampling

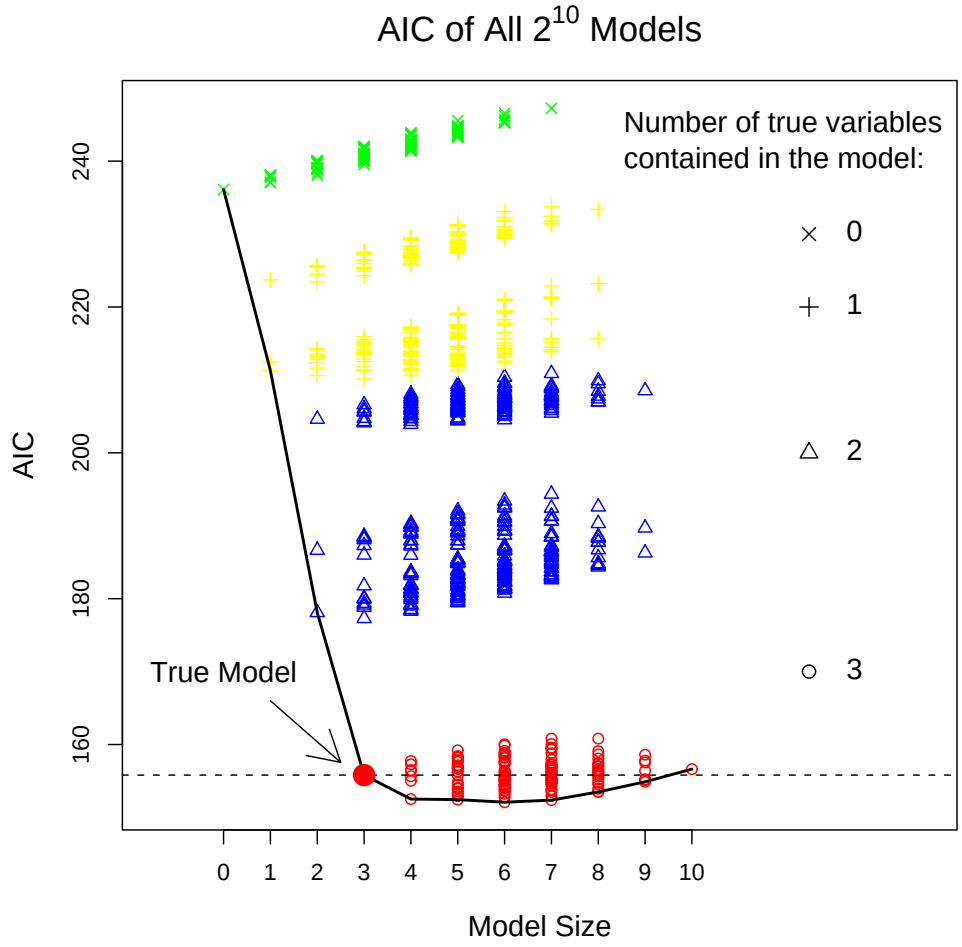
References

Z&C: Zhu & Chipman (2006; *Technometrics*)

M&B: Meinshausen & Bühlmann (2010; *JRSSB*)

W+: Wang *et al.* (2011; *Ann. Appl. Stat.*)

X&Z: Xin & Zhu (2012; *J. Comp. Graph. Stat.*)



Total of ten predictors $x_1, x_2, ..., x_{10}$. True model: $y = x_2 + x_5 + x_8 + \varepsilon$. [M. Zhu (2008; *Am Stat*)]

Machine Learning Markets

- classification outcomes: $c_1, c_2, ..., c_K$
- price vector (p_k is the price for betting on the outcome c_k):

$$\boldsymbol{p} = (p_1, p_2, ..., p_K)^{\mathrm{T}}$$

- reward vectors:

$$\begin{aligned} \boldsymbol{r}_1 &= (1, 0, 0, ..., 0, 0)^{\mathrm{T}} \\ \boldsymbol{r}_2 &= (0, 1, 0, ..., 0, 0)^{\mathrm{T}} \\ &\vdots \qquad\qquad \vdots \\ \boldsymbol{r}_K &= (0, 0, 0, ..., 0, 1)^{\mathrm{T}} \end{aligned}$$

- players $b = 1, 2, ..., B$

Each Player

- wealth: w_b
- utility function: $u_b(\cdot)$
- belief for outcome c_k: $f_b(k)$
- holding vector (fractional and negative positions allowed):

$$\boldsymbol{z}_b = [z_b(1), z_b(2), ..., z_b(K)]^{\mathrm{T}}$$

- each player maximizes his/her expected utility,

$$\max_{\boldsymbol{z}_b} \quad \sum_{k=1}^{K} u_b\left(w_b - \boldsymbol{p}^{\mathrm{T}}\boldsymbol{z}_b + \boldsymbol{r}_k^{\mathrm{T}}\boldsymbol{z}_b\right) f_b(k),$$

  under various market constraints

Equilibrium Prices

A. Storkey (2011), "Machine learning markets", in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pp. 716-724.

- computed various equilibrium prices for different utility functions
- e.g., if $u_b(\cdot) = \log(\cdot)$, then

$$p_k^* = \frac{\sum_b w_b f_b(k)}{\sum_b w_b}$$

- implication: with different $u_1, u_2, ..., u_B$, p_k^* would represent different (perhaps more complex) combinations of $f_b(k)$

Magging

- Meinshausen & Bühlmann (2014; arXiv:1406.0596) + Bühlmann & Meinshausen (2014; arXiv:1409.2638)

- suppose $\exists$ heterogeneous groups, $G_1, G_2, ..., G_B$, and

$$y_i = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_i + \varepsilon_i, \quad \text{where} \quad \boldsymbol{\beta}_i = \boldsymbol{\beta}_b \quad \forall \quad i \in G_b$$

- don't know which group $\boldsymbol{x}_{new}$ belongs to @ prediction time

- use a regression estimate $\boldsymbol{\beta}^*$ that maximizes the "explained variance in the most adversarial scenario" instead of, say, averaging individual estimates from each group

Explained Variance

- suppose true coefficient $= \boldsymbol{\beta}$ and estimate $= \boldsymbol{\gamma}$
- then, explained variance is

$$\begin{aligned}\Omega(\boldsymbol{\beta},\boldsymbol{\gamma}) &= \mathbb{V}\mathrm{ar}(y) - \mathbb{V}\mathrm{ar}(y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\gamma}) \\ &= \mathbb{V}\mathrm{ar}(y) - \left[\mathbb{V}\mathrm{ar}(y) - 2\mathbb{C}\mathrm{ov}(y, \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\gamma}) + \mathbb{V}\mathrm{ar}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\gamma})\right] \\ &= 2\boldsymbol{c}^{\mathrm{T}}\boldsymbol{\gamma} - \boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\gamma}\end{aligned}$$

where

$$\boldsymbol{c} \equiv \mathbb{C}\mathrm{ov}(y, \boldsymbol{x}) = \mathbb{C}\mathrm{ov}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \boldsymbol{x}) \quad \text{and} \quad \boldsymbol{\Sigma} \equiv \mathbb{V}\mathrm{ar}(\boldsymbol{x})$$

- under (common) assumption of $\mathbb{C}\mathrm{ov}(\boldsymbol{\varepsilon}, \boldsymbol{x}) = 0$, get

$$\boldsymbol{c} = \boldsymbol{\Sigma}\boldsymbol{\beta} \quad \Rightarrow \quad \Omega(\boldsymbol{\beta},\boldsymbol{\gamma}) = 2\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\gamma} - \boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\gamma}$$

The Maximin Effect

- assume $\boldsymbol{\beta}_i \sim \mathcal{P}_\beta$ [e.g., for heterogeneous groups, $\mathcal{P}_\beta$ = finite mixture of point masses]

- go after the quantity,

$$\boldsymbol{\beta}^* \quad = \quad \max_{\boldsymbol{\gamma}} \quad \min_{\boldsymbol{\beta} \in \text{Supp}(\mathcal{P}_\beta)} \quad \Omega(\boldsymbol{\beta}, \boldsymbol{\gamma}),$$

  which they called the "maximin effect"

- the term "most adversarial scenario" refers to the case in which our estimate is least capable of explaining the variance of y

Estimating the Maximin Effect

- first, generate different individual estimates $\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, ..., \widehat{\boldsymbol{\beta}}_B$ using random sub-samples of the training data

- then, aggregate these individual estimates by taking the point in their convex hull $\mathcal{H}(\widehat{\boldsymbol{\beta}}_1, ..., \widehat{\boldsymbol{\beta}}_B)$ that is closest to the origin (in a particular norm), i.e.,

$$\widehat{\boldsymbol{\beta}^*} = \underset{\boldsymbol{u} \in \mathcal{H}(\widehat{\boldsymbol{\beta}}_1, ..., \widehat{\boldsymbol{\beta}}_B)}{\arg\min} \|\boldsymbol{u}\|^2_{\widehat{\boldsymbol{\Sigma}}}$$

where $\|\boldsymbol{u}\|^2_{\widehat{\boldsymbol{\Sigma}}} = \boldsymbol{u}^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}} \boldsymbol{u}$

called Magging ... for maximin aggregating

Our Journey: Bagging → Magging

- many successes, but successes also perpetuate myths,
  - e.g., "if a collection of classifiers are 'nearly independent' of each other, then taking majority votes over their results will always achieve greater accuracy"
- for a counter example, see
  - S. B. Vardeman & M. D. Morris (2013), "Majority voting by independent classifiers can increase error rates", *The American Statistician* **67**, pp. 94–96.
- for today's lecture + more on such myths, watch out for
  - M. Zhu (forthcoming), "Use of majority votes in statistical learning", *Wiley Interdisciplinary Reviews: Comp. Stat.*

Summary

- key ideas:
  - ensembles
  - strength-diversity trade-off
- specific methods:
  - bagging; AdaBoost
  - random forest; gradient boosting machine
  - variable selection ensemble
  - machine learning market
  - magging