Introduction to Machine Learning
Lecture 12

Mehryar Mohri
Courant Institute and Google Research
mohri@cims.nyu.edu

Bagging

Ensemble Methods - Classification

- Problem: given T binary classification hypotheses $(h_1, \ldots, h_T)$ find combined classifier

$$f \colon x \mapsto \operatorname{sgn}\Big(\sum_{t=1}^{T} \alpha_t h_t\Big),$$

  with better performance.

- When does it work? Need diversity (e.g., different features, different training sets).

$\longrightarrow$ use different subsets of the data for training.

Bagging - Classification

(Breiman, 1996)

Bagging (Boostrap aggregating).

$\textsc{Bagging}(S = ((x_1, y_1), \ldots, (x_m, y_m)))$

for $t \leftarrow 1$ **to** T **do**

  $S_t \leftarrow \textsc{Bootstrap}(S) \triangleright$ i.i.d. sampling with replacement from S.

  $h_t \leftarrow \textsc{TrainClassifier}(S_t)$

return $h_S = x \mapsto \textsc{MajorityVote}((h_1(x), \ldots, h_T(x)))$

Bagging - Regression

(Breiman, 1996)

Bagging (Boostrap aggregating).

$\textsc{Bagging}(S = ((x_1, y_1), \ldots, (x_m, y_m)))$

for $t \leftarrow 1$ **to** T **do**

  $S_t \leftarrow \textsc{Bootstrap}(S)$ ▷ i.i.d. sampling with replacement from S.

  $h_t \leftarrow \textsc{TrainRegressionAlgorithm}(S_t)$

return $h_S = x \mapsto \textsc{Mean}((h_1(x), \ldots, h_T(x)))$

Bias-Variance Decomposition

- Proposition: for any hypothesis h_S, the following decomposition holds:

$$\operatorname*{E}_S\Big[\operatorname*{E}_{X,Y}[(h_S(X)-Y)^2]\Big]$$
$$= \underbrace{\operatorname*{E}_{S,X}\Big[\Big(h_S(X)-\operatorname*{E}_S[h_S(X)]\Big)^2\Big]}_{\text{variance}} + \underbrace{\operatorname*{E}_X\Big[\Big(\operatorname*{E}_S[h_S(X)]-\operatorname{E}[Y|X]\Big)^2\Big]}_{\text{bias}} + \underbrace{\operatorname*{E}_X\Big[(Y-\operatorname{E}[Y|X])^2\Big]}_{\text{noise}}.$$

- Bias-variance minimization trade-off:
  - small S and large H: small bias, large variance.
  - large S and small H: large bias, small variance.

Bias-Variance Decomp. Proof

- Observe that

$$\begin{aligned}
&\mathop{\mathrm{E}}_{X,Y}[(h_S(X) - \mathrm{E}[Y|X])(\mathrm{E}[Y|X] - Y)]\\
&= \mathop{\mathrm{E}}_{X}\Big[\mathop{\mathrm{E}}_{Y|X}[(h_S(X) - \mathrm{E}[Y|X])(\mathrm{E}[Y|X] - Y)]\Big]\\
&= \mathop{\mathrm{E}}_{X}\Big[(h_S(X) - \mathrm{E}[Y|X])\mathop{\mathrm{E}}_{Y|X}[(\mathrm{E}[Y|X] - Y)]\Big] = 0.
\end{aligned}$$

- Therefore,

$$\begin{aligned}
\mathop{\mathrm{E}}_{X,Y}[(h_S(X) - Y)^2] &= \mathop{\mathrm{E}}_{X,Y}\Big[\big[(h_S(X) - \mathrm{E}[Y|X]) + (\mathrm{E}[Y|X] - Y)\big]^2\Big]\\
&= \mathop{\mathrm{E}}_{X,Y}\big[(h_S(X) - \mathrm{E}[Y|X])^2\big] + \mathop{\mathrm{E}}_{X,Y}\big[(\mathrm{E}[Y|X] - Y)^2\big].
\end{aligned}$$

Bias-Variance Decomp. Proof

- Since $\mathop{\mathrm{E}}_{S}\Big[h_S(X) - \mathop{\mathrm{E}}_{S}[h_S(X)]\Big] = 0$, the following holds:

$$
\begin{aligned}
&\mathop{\mathrm{E}}_{S}\Big[\mathop{\mathrm{E}}_{X,Y}\big[(h_S(X) - \mathrm{E}[Y|X])^2\big]\Big] \\
&= \mathop{\mathrm{E}}_{X,Y}\Big[\mathop{\mathrm{E}}_{S}\Big[\Big(h_S(X) - \mathop{\mathrm{E}}_{S}[h_S(X)] + \mathop{\mathrm{E}}_{S}[h_S(X)] - \mathrm{E}[Y|X]\Big)^2\Big]\Big] \\
&= \mathop{\mathrm{E}}_{X,Y}\Big[\mathop{\mathrm{E}}_{S}\Big[\Big(h_S(X) - \mathop{\mathrm{E}}_{S}[h_S(X)]\Big)^2\Big]\Big] + \mathop{\mathrm{E}}_{X}\Big[\mathop{\mathrm{E}}_{S}\Big[\Big(\mathop{\mathrm{E}}_{S}[h_S(X)] - \mathrm{E}[Y|X]\Big)^2\Big]\Big].
\end{aligned}
$$

Ensemble Bias

- Bias of averaged hypothesis in regression:

$$\begin{aligned}
\text{Bias}\,(h_S, x) &= \mathop{\text{E}}_{S \sim D^m}\left[\frac{1}{T}\sum_{t=1}^{T} h_t(x)\right] - \text{E}[Y|x] \\
&= \frac{1}{T}\sum_{t=1}^{T}\left[\mathop{\text{E}}_{S \sim D^m}[h_t(x)] - \text{E}[Y|x]\right] \\
&= \frac{1}{T}\sum_{t=1}^{T} \text{Bias}\,(h_t, x)\,.
\end{aligned}$$

→ Thus, relatively unbiased base hypotheses lead to relatively unbiased ensemble.

Ensemble Variance

- Proposition: for any x, the variance of the ensemble hypothesis at x is given by

$$\mathrm{Var}\left(h_S, x\right) = \frac{1}{T^2} \sum_{t=1}^{T} \mathrm{Var}\left(h_t, x\right) + \frac{1}{T^2} \sum_{t \neq t'} \mathrm{Cov}[h_t(x), h_{t'}(x)].$$

  - thus, if approximately uncorrelated base hypotheses $\mathrm{Cov}[h_t(x), h_{t'}(x)] \approx 0$.
  - assume approximately equal variances, then

$$\mathrm{Var}\left(h_S, x\right) \approx \frac{1}{T} \mathrm{Var}\left(h_1, x\right).$$

$\longrightarrow$ reduction by $1/T$.

Bagging - Regression

- Regression properties:
  - small covariances (different subsets).
  - similar variances (on average).
  - similar biases.
- Classification: unclear explanation.

References

- Leo Breiman. Bagging Predictors. *Machine Learning*, 24:123-140, 1996.

- Peter Bühlmann and Bin Yu. Analyzing Bagging. *The Annals of Statistics*. 30(4): 927-961, 2002.

- Andreas Buja and Werner Stuetzle. "The effects of bagging on variance, bias, and mean squared error", Preprint, AT&T Labs-Research, 2000.