Example models

Outline
- the basic idea of models
- history
- Bayesian statistics and gradient boosting

Interpretable supervised prediction by combining models instead of using a single model

Bagged regression ensemble

Training a subset, but weight/score/vote/combine instances of the same datapoint

i.e. bootstrap aggregating
Breiman (1994) for a linear regression ensemble
Breiman (1996) for classification

Construct regression models $M_1^{(b)}, \dots, M_n^{(b)}$ by training the corresponding ensemble member on a bootstrapped sample from D

Bagged model for each model member is

$$\hat{M}_i^{(b)}(x) = \frac{1}{|D^{(b)}|} \sum_{(x',y') \in D^{(b)}} M_i^{(b)}(x')$$

$$\hat{M}_{bag}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{M}_i^{(b)}(x)$$

Under what circumstances does a bagged ensemble do better than a single predictor?

Assume the true function is $f(x)$. Then,

$$M_i^{(b)}(x) = f(x) + \epsilon_i^{(b)}(x) \text{ where}$$

$$\epsilon_i^{(b)}(x) \sim N(0, \sigma^{(b)})$$

Expected error of each model member is

$$E_{\epsilon_i}\left[\left(M_i(x) - f(x)\right)^2\right] = E_{\epsilon_i}\left[\left(\epsilon_i^{(b)}(x)\right)^2\right]$$

Average error over the whole ensemble

$$E_{bag} = \frac{1}{B} \sum_{b=1}^{B} E_{\epsilon_i}\left[\left(\epsilon_i^{(b)}(x)\right)^2\right]$$

The expected error of the bagged ensemble is

$$E_{bag} = E_{\epsilon_i}\left[\left(\frac{1}{B}\sum_{b=1}^{B} \epsilon_i^{(b)}(x)\right)^2\right]$$

You can show that when $E_{\epsilon_i}[\epsilon_i^{(b)}(x)] = N(0, \sigma^{(b)})$ and when errors made by the individual members are uncorrelated then

$$E_{bag} = \frac{1}{B} \sum_{b=1}^{B} E_{\epsilon_i}\left[\left(\epsilon_i^{(b)}(x)\right)^2\right]$$

$$E_{bag} = \frac{1}{B} E_{bag}$$

$$\therefore E_{bag} \ll E_{bagg}$$

Bagged classification ensemble

train 1 classifier on a bootstrapped sample
test classifier on bagging set

$$\hat{M}_i^{(b)}(x) = \begin{cases} 1 & \text{if } \sum_{(x',y') \in D^{(b)}} \mathbb{1}(M_i^{(b)}(x') = y') \geq \frac{|D^{(b)}|}{2} \\ 0 & \text{otherwise} \end{cases}$$

$\hat{M}_i^{(b)}(x)$ = majority vote over $\hat{M}_i^{(b)}(x)$, $\hat{M}_i^{(b)}(x)$

$$\hat{M}_{bag}(x) = \sum_{b=1}^{B} \mathbb{1}(\hat{M}_i^{(b)}(x) = 1)$$

Under what circumstances does an ensemble of classifiers perform better than a single classifier?

(1) Assume each member of the ensemble does worse than α $\left(\frac{1}{2} - \frac{1}{2}\sqrt{\frac{\pi}{2}}\right)$

(2) Assume each ensemble member errors are uncorrelated with any other

$$E_{bag} = \sum_{i=\frac{B}{2}+1}^{B} \binom{B}{i} \alpha^i (1-\alpha)^{B-i}$$

α = error rate of each classifier.

Suppose $\alpha = 0.3$ and $B = 21$ (Breiman 1996)

The error rate of bagging is then 0.026 which is a 10 fold improvement over the single classifier.

Bagging is a simple yet powerful ensemble method, which reduces variance of the individual models by positively biased members of the ensemble.

To do this, we have sampled $D^{(b)}$ uncorrelated with each other, $D^{(b)}$ is in fact uncorrelated with and $\hat{M}_i^{(b)}(x)$ is uncorrelated with each other.

Bagging is a simple yet powerful ensemble method, which reduces variance of the individual models by positively biased members of the ensemble.

Bagging is strongly predictive with unstable models, where the bagged ensemble can be learned independently, also numbers get an equal vote in the final classification.

Boosting (Freund and Schapire 1996)

The members of the ensemble are learned sequentially, with each member focusing on the instances that the previously learned members of the ensemble

To do this, we have sampled $D^{(b)}$ uncorrelated with each $D^{(b)}$, $D^{(b)}$ is in fact uncorrelated with and $\hat{M}_i^{(b)}(x)$ is not uncorrelated with each other.

Bagging is a simple yet powerful ensemble method, which reduces variance of the individual models by positively biased members of the ensemble.