Bayes Model Averaging

Stephen Vardeman

Analytics Iowa LLC

ISU Statistics and IMSE

Averaging of predictors

SEL bagging averages an "ensemble" of predictors consisting of versions of a single predictor computed from different bootstrap samples. An alternative might be to somehow weight together (or otherwise combine) different predictors (potentially even based on different models or methods). Here we consider a theoretically unimpeachable (but ultimately typically impractical) version of this basic idea of combining an ensemble of predictors to produce one better than any indiviual element of the ensemble. That is "Bayes model averaging."

Bayes multiple model scenario

One theoretically straightforward way to justify this kind of enterprise is through a Bayes "multiple model" scenario. Suppose that M models $P_1, P_2, \ldots, P_M$ for $(\mathbf{x}, y)$ are under consideration, the mth of which has parameter vector $\boldsymbol{\theta}_m$ and corresponding density $p_m(\mathbf{x}, y|\boldsymbol{\theta}_m)$. Then for the mth model (repeatedly abusing notation by using p to name many different functions) the training set $\mathbf{T}$ has density

$$p_m(\mathbf{T}|\boldsymbol{\theta}_m) = \prod_{i=1}^{N} p_m(\mathbf{x}_i, y_i|\boldsymbol{\theta}_m)$$

We'll suppose here that $\boldsymbol{\theta}_m$ is not known and that it has prior density $g_m(\boldsymbol{\theta}_m)$ (for the mth model) and that and a prior probability for model m is $\pi(m)$. Then a joint distribution for $m, \mathbf{T}, \boldsymbol{\theta}_m$, and $(\mathbf{x}, y)$ has density

$$p_m(\mathbf{x}, y|\boldsymbol{\theta}_m)\, p_m(\mathbf{T}|\boldsymbol{\theta}_m)\, g_m(\boldsymbol{\theta}_m)\, \pi(m)$$

A joint marginal and conditional mean

This has a marginal density for $y, \mathbf{x}, \mathbf{T}$ that is

$$\sum_{m=1}^{M} \pi(m) \int p_m(\mathbf{x}, y|\boldsymbol{\theta}_m)\, p_m(\mathbf{T}|\boldsymbol{\theta}_m)\, g_m(\boldsymbol{\theta}_m)\, d\boldsymbol{\theta}_m$$

from which the conditional mean of $y|\mathbf{x}, \mathbf{T}$ is

$$\mathrm{E}[y|\mathbf{x}, \mathbf{T}] = \frac{\sum_{m=1}^{M} \pi(m) \int\int y p_m(\mathbf{x}, y|\boldsymbol{\theta}_m)\, p_m(\mathbf{T}|\boldsymbol{\theta}_m)\, g_m(\boldsymbol{\theta}_m)\, d\boldsymbol{\theta}_m dy}{\sum_{m=1}^{M} \pi(m) \int\int p_m(\mathbf{x}, y|\boldsymbol{\theta}_m)\, p_m(\mathbf{T}|\boldsymbol{\theta}_m)\, g_m(\boldsymbol{\theta}_m)\, d\boldsymbol{\theta}_m dy}$$

A model m conditional mean

Given m (the identity of the "correct" model) the variables $\mathbf{T}$, $\boldsymbol{\theta}_m$, and $(\mathbf{x}, y)$ have joint density

$$p_m\left(\mathbf{x}, y|\boldsymbol{\theta}_m\right) p_m\left(\mathbf{T}|\boldsymbol{\theta}_m\right) g_m\left(\boldsymbol{\theta}_m\right)$$

for which the conditional mean of $y|\mathbf{x}, \mathbf{T}, m$ is, say,

$$\mathrm{E}\left[y|\mathbf{x}, \mathbf{T}, m\right] = \frac{\int\int y p_m\left(\mathbf{x}, y|\boldsymbol{\theta}_m\right) p_m\left(\mathbf{T}|\boldsymbol{\theta}_m\right) g_m\left(\boldsymbol{\theta}_m\right) d\boldsymbol{\theta}_m dy}{\int\int p_m\left(\mathbf{x}, y|\boldsymbol{\theta}_m\right) p_m\left(\mathbf{T}|\boldsymbol{\theta}_m\right) g_m\left(\boldsymbol{\theta}_m\right) d\boldsymbol{\theta}_m dy}$$

so that

$$\int\int y p_m\left(\mathbf{x}, y|\boldsymbol{\theta}_m\right) p_m\left(\mathbf{T}|\boldsymbol{\theta}_m\right) g_m\left(\boldsymbol{\theta}_m\right) d\boldsymbol{\theta}_m dy$$
$$= \mathrm{E}\left[y|\mathbf{x}, \mathbf{T}, m\right] \cdot \int\int p_m\left(\mathbf{x}, y|\boldsymbol{\theta}_m\right) p_m\left(\mathbf{T}|\boldsymbol{\theta}_m\right) g_m\left(\boldsymbol{\theta}_m\right) d\boldsymbol{\theta}_m dy$$

A posterior weighted average of conditional means

Thus

$$\mathrm{E}\left[y|\mathbf{x},\mathbf{T}\right] = \frac{\sum_{m=1}^{M}\mathrm{E}\left[y|\mathbf{x},\mathbf{T},m\right]\pi\left(m\right)\int\int p_m\left(\mathbf{x},y|\boldsymbol{\theta}_m\right)p_m\left(\mathbf{T}|\boldsymbol{\theta}_m\right)g_m\left(\boldsymbol{\theta}_m\right)d\boldsymbol{\theta}_m dy}{\sum_{m=1}^{M}\pi\left(m\right)\int\int p_m\left(\mathbf{x},y|\boldsymbol{\theta}_m\right)p_m\left(\mathbf{T}|\boldsymbol{\theta}_m\right)g_m\left(\boldsymbol{\theta}_m\right)d\boldsymbol{\theta}_m dy}$$

This is the average of $\mathrm{E}[y|\mathbf{x},\mathbf{T},m]$ with respect to the conditional distribution (the "posterior" distribution) of $m|\mathbf{x},\mathbf{T}$ specified by

$$\pi\left(m|\mathbf{x},\mathbf{T}\right) = \frac{\pi\left(m\right)\int\int p_m\left(\mathbf{x},y|\boldsymbol{\theta}_m\right)p_m\left(\mathbf{T}|\boldsymbol{\theta}_m\right)g_m\left(\boldsymbol{\theta}_m\right)d\boldsymbol{\theta}_m dy}{\sum_{m=1}^{M}\pi\left(m\right)\int\int p_m\left(\mathbf{x},y|\boldsymbol{\theta}_m\right)p_m\left(\mathbf{T}|\boldsymbol{\theta}_m\right)g_m\left(\boldsymbol{\theta}_m\right)d\boldsymbol{\theta}_m dy}$$

That is, optimal SEL prediction weights optimal predictors of y from the M constituent models by the relevant (updated from $\pi\left(m\right)$ by the information in $\mathbf{x}$ and $\mathbf{T}$ about the relevant density $p_m\left(\mathbf{x},y|\boldsymbol{\theta}_m\right)$) conditional probabilities of the M components.

0-1 loss and another view

Essentially the same argument pertains in cases where y takes values in $\mathcal{G} = \{1, 2, \ldots, K\}$ and 0-1 loss is involved. Under the same model as above, $P[y = k|\mathbf{x}, \mathbf{T}]$ is a $\pi(m|\mathbf{x}, \mathbf{T})$-weighted average of $P[y = k|\mathbf{x}, \mathbf{T}, m]$s appropriate under the M constituent models. (Of course, integrals "dy" are sums.) Ultimately, optimal 0-1 loss classifiers then choose for input $\mathbf{x}$ (and training set $\mathbf{T}$) the class k maximizing this Bayes model average probability.

These developments of Bayes model averaging predictors explicitly involve $\mathbf{x}$ in the posterior distribution of m (given $\mathbf{x}$ and $\mathbf{T}$). This is because if one thinks of a new $\mathbf{x}$ and corresponding y as generated by the same mechanism that produces $\mathbf{T}$, the observed $\mathbf{x}$ is informative about m.

Another way of modeling and calculating follows.

Functions of $\boldsymbol{x}$ as objects of interest

One might suppose that the functions of $\mathbf{x}$,

$$\mu_m(\mathbf{x}) = \frac{\int\int y p_m(\mathbf{x}, y|\boldsymbol{\theta}_m)\, g_m(\boldsymbol{\theta}_m)\, d\boldsymbol{\theta}_m dy}{\int\int p_m(\mathbf{x}, y|\boldsymbol{\theta}_m)\, g_m(\boldsymbol{\theta}_m)\, d\boldsymbol{\theta}_m dy}$$

or

$$p_m(y|\mathbf{x}) = \frac{\int p_m(\mathbf{x}, y|\boldsymbol{\theta}_m)\, g_m(\boldsymbol{\theta}_m)\, d\boldsymbol{\theta}_m}{\sum_{y=1}^{K}\int p_m(\mathbf{x}, y|\boldsymbol{\theta}_m)\, g_m(\boldsymbol{\theta}_m)\, d\boldsymbol{\theta}_m}$$

are objects of interest, but without a necessary connection to a specific new observation $\mathbf{x}$, itself informative about m and $\boldsymbol{\theta}_m$. (These functions are the conditional means of and densities for y given $\mathbf{x}$ under particular models m.)

Another joint distribution and posterior for m

Positing a distribution specified by

$$p_m\left(\mathbf{T}|\boldsymbol{\theta}_m\right) g_m\left(\boldsymbol{\theta}_m\right) \pi\left(m\right)$$

for $m, \boldsymbol{\theta}_m, \mathbf{T}$ in the multiple model scenario, the posterior distribution for m given $\mathbf{T}$ has pmf

$$\pi\left(m|\mathbf{T}\right) = \frac{\pi\left(m\right) \int p_m\left(\mathbf{T}|\boldsymbol{\theta}_m\right) g_m\left(\boldsymbol{\theta}_m\right) d\boldsymbol{\theta}_m}{\sum_{m=1}^{M} \pi\left(m\right) \int p_m\left(\mathbf{T}|\boldsymbol{\theta}_m\right) g_m\left(\boldsymbol{\theta}_m\right) d\boldsymbol{\theta}_m}$$

Averages of functions

So the posterior mean of $\mu_m(\mathbf{x})$ given $\mathbf{T}$ is

$$\sum_{m=1}^{M} \mu_m(\mathbf{x})\, \pi(m|\mathbf{T})$$

and the posterior mean of $p_m(y|\mathbf{x})$ given $\mathbf{T}$ is

$$\sum_{m=1}^{M} p_m(y|\mathbf{x})\, \pi(m|\mathbf{T})$$

These differ from the previous "Bayes model averages," but they also represent sensible ensembles of predictors appropriate in the constituent models.