Ensemble methods for classification and regression

DATA 607 — Session 4 — 06/03/2019

Ensemble learning

Improve accuracy of a regressor by constructing $N \gg 0$ regression functions $\widehat{r}_i$, for the same dataset, and setting

$$\widehat{r}(x) = \frac{1}{N} \sum_i \widehat{r}_i(x).$$

Improve accuracy of a classifier by constructing $N \gg 0$ different classification functions, $\widehat{r}_i$, for the same dataset, and setting

$$\widehat{r}(x) = \underset{k=1,\ldots,K}{\operatorname{argmax}} \left| \{ i : \widehat{r}_i(x) = k \} \right|.$$

Resampling via Bootstrap Sampling

We make different decision trees from the same dataset by resampling – sampling from our sample (i.e., our dataset).

Similar, in spirit, to splitting a sample into training and testing subsets, or into folds for cross validation.

We use bootstrap samples.

A **bootstrap sample** from a dataset D of size n is a dataset D' constructed by selecting n elements from D, *with replacement*. It's almost certain that D' will contain the same observations multiple times.

We expect D' to contain a proportion $1 - 1/e \approx 0.632$ of the elements of D.

Example: $\{x_1, x_2, x_1, x_3\}$ is a bootstrapped sample from $\{x_1, x_2, x_3, x_4\}$.

Bagging Classifiers

Bagging is short for **B**ootstrap **AGG**regat**ING**. It's also reasonable because, loosely speaking, you're using a "bag" of classifiers.

Procedure:

1. Draw $N \gg 0$ bootstrap samples $D_1, \ldots, D_n$ from your dataset, D.
2. Train a regressor/classifier, $\widehat{r}_i$, on each D_i.
3. To predict the target value for a new observation, x:
   - average the $\widehat{r}_i(x)$ in the regression case.
   - assign the most common class label among the $\widehat{r}_i(x)$ in the classification case.

This technique can be applied to any regressor or classifier, but is most useful for those that have high variance, i.e., that have tend to overfit. Decision trees, for example. (Finding a toy dataset that makes this point isn't so easy, though!)

Random forests

Highly correlated predictors can lead to instability and overfitting. They also limit the effectiveness of bagging the bagged estimates will inherit the correlation.

Random forests is an ensemble technique that uses a simple — but effective — trick to mitigate this issue.

Procedure:

1. Draw $N \gg 0$ bootstrap samples $D_1, \ldots, D_n$ from your dataset, D.
2. Train a regressor/classifier, $\widehat{r}_i$, on each D_i, using only a random sample of $m \approx \sqrt{p}$ of the p features.
3. Predict target values for new observations just like for bagging.