Tree-based methods for classification and regression

DATA 607 — Session 3 — 04/03/2019

Regression trees

Dataset: $(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots, (\vec{x}_n, \vec{y}_n)$, where $x_i \in R \subseteq \mathbb{R}^p$ and $y_i \in \mathbb{R}$

Regression tree construction:

- Divide the predictor space, R, into disjoint regions, $R_1, \ldots, R_J$
- If a new observation, $\vec{x}$, falls into region R_j, set $\widehat{r}(\vec{x})$ equal to the average y_i such that $x_i \in R_j$.

How do we subdivide the predictor space? **recursive bisection**

Given an index j and a *cutpoint* $s \in \mathbb{R}$, bisect R into two pieces:

$$R_0 = \{\vec{x} \in R : x_j \leq s\}, \qquad R_1 = \{\vec{x} \in R : x_j > s\}$$

Then bisect R_0 and R_1...

Given an index j_0 and a cutpoint $s_0 \in \mathbb{R}$, bisect R_{00} into two pieces:

$$R_{00} = \{\vec{x} \in R_0 : x_{j_0} \leq s_0\}, \qquad R_{01} = \{\vec{x} \in R_0 : x_{j_0} > s_0\}.$$

Given an index j_1 and a cutpoint $s_1 \in \mathbb{R}$, bisect R_{01} into two pieces:

$$R_{10} = \{\vec{x} \in R_1 : x_{j_1} \leq s_1\}, \qquad R_{11} = \{\vec{x} \in R_1 : x_{j_1} > s_1\}$$

Then bisect R_{00}, R_{01}, R_{01}, and R_{10}...

Etc...

Questions:

1. How do we find the indices j, j_0, j_1, ... and the cutpoints s, s_0, s_1, ...?
2. When do we stop?

Answers:

1. For each pair (j, s), let $\text{Var}^-(j, s)$ and $\text{Var}^+(j, s)$ be the variances of

$$\{y_i : x_{i,j} \leq s\} \quad \text{and} \quad \{y_i : x_{i,j} > s\},$$

respectively.
Take (j, s) to be the pair that *minimizes the total variance*,

$$\text{Var}^-(j, s) + \text{Var}^+(j, s).$$

2. Don't bisect a region if the number of training points it contains is smaller than some threshold.

Classification trees

Dataset $(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots, (\vec{x}_n, \vec{y}_n)$, where

$$x_i \in R \subseteq \mathbb{R}^p, \quad y_i \in \{1, \ldots, K\}.$$

Classification tree construction:

- Divide the predictor space, R, into disjoint regions, $R_1, \ldots, R_J$
- If a new observation, $\vec{x}$, falls into region R_j, set $\widehat{r}(\vec{x})$ equal to most common class label, y_i, for which $x_i \in R_j$.

Partition R via the same recusrive bisection procedure described for regression trees, except we replace the variance minimization criterion with one more suited to categorical response variables.

To bisect a region, R:

For class labels k, indices j, and cutpoints s, let

$$\widehat{p}_k(j,s) = \frac{|\{i : x_i \in R,\, y_i = k\}|}{|\{i : x_i \in R\}|}$$

$$G(j,s) = \sum_{k=1}^{K} \widehat{p}_k(j,s)(1 - \widehat{p}_k(j,s)) \qquad \text{(Gini index)}$$

$$D(j,s) = -\sum_{k=1}^{K} \widehat{p}_k(j,s) \log \widehat{p}_k(j,s) \qquad \text{(cross entropy)}$$

Choose (j,s) to minimize either $G(j,s)$ or $D(j,s)$.

Gini index and cross entropy are examples of *(im)purity measures*.

Bagging

Decision tree regression and classification suffer from high variance, in the sense that variance among trees fitted with different training subsets of the same dataset will be quite high.

We can bring down this variance by *bagging* (**B**ootstrap **AGG**regat**ING**):

- Generate B bootstrapped samples from your dataset.
- For each $b \in B$, fit a decision tree to b to get a regression function, $\widehat{r}_b$.
- Let $\widehat{r}(\vec{x})$ be the average of the $\widehat{r}_b(\vec{x})$, in the case of continuous response, or the most commonly occuring class label among the $\widehat{r}_b(\vec{x})$, in the case of a categorical response variable.