Quantile Regression Forests Report

[Team Name etc etc]

Introduction:

Let Y be a real-valued response variable and X a covariate or predictor variable, possibly high-dimensional. A standard goal of statistical analysis is to infer, in some way, the relationship between Y and X. Standard regression analysis tries to come up with an estimate $\widehat{\mu}(x)$ of the conditional mean E(Y | X = x). The conditional mean minimizes the expected squared error loss,

$$E(Y \,|\, X = x) = arg\, min_z E\{(Y - z)^2 \,|\, X = x\}$$

and approximation of the conditional mean is typically achieved by minimization of a squared error type loss function over the available data.

The conditional mean illuminates just one aspect of the conditional distribution of a response variable Y , yet neglects all other features of possible interest. This led to the development of quantile regression.

For a continuous distribution function, the α-quantile $Q_\alpha(x)$ is defined such that the probability of Y being smaller than $Q_\alpha(x)$ is, for a given X = x, exactly equal to α.

$$Q_\alpha(x) = inf\{y : F(y \,|\, X = x) \ \geq \ \alpha\}$$

The quantiles give more complete information about the distribution of Y as a function of the predictor variable X than the conditional mean alone.

Applications:

1) Prediction Intervals

Quantile regression can be used to build prediction intervals. A 95% prediction interval for the value of Y is given by

$$I(x) = [Q_{.025}(x),\ Q_{.975}(x)]$$

That is, a new observation of Y , for X = x, is with high probability in the interval I(x).

2) Outlier Detection

Quantile regression can be used for outlier detection.

3) Estimating Quantiles From Data

Quantile regression can be used to estimate quantiles from data.

Random Forests:

The prediction of a single tree T(θ) for a new data point X = x is obtained by averaging over the observed values in leaf $l(x,\ \theta)$. Let the weight vector $w_i(x,\ \theta)$ be given by a positive constant if observation X_i is part of leaf $l(x,\ \theta)$ and 0 if it is not. The weights sum to one, and thus

$$w_i(x,\ \theta) = \frac{1_{\{X_i \in R_l(x,\theta)\}}}{\#\{j : X_j \in R_l(x,\theta)\}}$$

Where R_l is a rectangular subspace of the space to which X belongs, which leaf l corresponds to.

The prediction of a single tree, given covariate X = x, is then the weighted average of the original observations Y_i, i = 1, . . . , n,

$$\hat{\mu}(x) = \sum_{i=1}^{n} w_i(x,\ \theta) Y_i$$

Using random forests, the conditional mean E(Y | X = x) is approximated by the averaged prediction of k single trees, each constructed with an i.i.d. vector θ_t, t = 1, . . . , k. Let $w_i(x)$ be the average of $w_i(\theta)$ over this collection of trees,

$$w_i(x) = k^{-1} \sum_{t=1}^{k} w_i(x, \theta_t)$$

The prediction of random forests is then

$$\widehat{\mu}(x) = \sum_{i=1}^{n} w_i(x)Y_i$$

Quantile Regression Forests:

Just as E(Y | X = x) is approximated by a weighted mean over the observations of Y, define an approximation to $E[1_{\{Y\leq y\}} \mid X = x]$ by the weighted mean over the observations of $1_{\{Y\leq y\}}$,

$$\widehat{F}(y \mid X = x) = \sum_{i=1}^{n} w_i(x)1_{\{Y_i\leq y\}}$$

Algorithm:

a) Grow k trees $T(\theta_t)$, t = 1, . . . , k, as in random forests. However, for every leaf of every tree, take note of all observations in this leaf, not just their average.

b) For a given X = x, drop x down all trees. Compute the weight $w_i(x, \theta_t)$ of observation i ∈ {1, . . . , n} for every tree. Compute weight $w_i(x)$ for every observation i ∈ {1, . . . , n} as an average over $w_i(x, \theta_t)$, t = 1, . . . , k.

c) Compute the estimate of the distribution function for all y ∈ R, using the weights.

Now estimates of $\widehat{Q}_\alpha(x)$ can be obtained.

The key difference between quantile regression forests and random forests is as follows: for each node in each tree, random forests keeps only the mean of the observations that fall into this node and neglects all other information. In contrast, quantile regression forests keeps the value of all observations in this node, not just their mean, and assesses the conditional distribution based on this information.

References:

[1] Nicolai Manhausen. Quantile Regression Forests