Squared-loss Impurity

As discussed in the video, we can tweak classification trees for continuous valued labels (regression). Because labels are no longer categorical, we redefine impurity such that it captures the "spread" of values in each node. Conveniently, we can capture this spread within a set by the variance of the labels within the set. The prediction made by a regression tree for a leaf with corresponding set S is simply the mean label $\bar{y}$. With this mean label as the predictor, the variance impurity is identical to the squared loss:

> ☆ **Key Points**
>
> Regression trees use a modified impurity metric appropriate for continuous label values.
>
> This impurity quantifies the "spread" or variance of the labels in a set.
>
> Each leaf in the final tree predicts the mean label of the training points inside the leaf.

$$I(S) = \frac{1}{|S|} \sum_{(\mathbf{x},y) \in S} (y - \bar{y})^2 \leftarrow \text{Average squared difference from average label}$$

$$\text{where } \bar{y} = \frac{1}{|S|} \sum_{(\mathbf{x},y) \in S} y \leftarrow \text{Average label}$$

Finding the best split

Remember, you evaluate the quality of a split of a parent set S into two sets S_L and S_R by the weighted impurity of the two branches:

$$\frac{|S_L|}{|S|} I(S_L) + \frac{|S_R|}{|S|} I(S_R)$$

In the case of the squared loss, this becomes:

$$\frac{1}{|S|} \sum_{(\mathbf{x},y) \in S_L} \left(y - \bar{y}_{S_L}\right)^2 + \frac{1}{|S|} \sum_{(\mathbf{x},y) \in S_R} \left(y - \bar{y}_{S_R}\right)^2$$

In the CART-ID3 algorithm, we evaluate all possible splits at each parent node S. As $|S|$ is constant when calculating the impurity for any S_L and S_R, you can simply pick the split of a set S into S_L, S_R that minimizes:

$$\sum_{(\mathbf{x},y) \in S_L} \left(y - \bar{y}_{S_L}\right)^2 + \sum_{(\mathbf{x},y) \in S_R} \left(y - \bar{y}_{S_R}\right)^2$$