Diversity Metrics for Regression Ensembles

Hosni M.

October 2024

Contents

Diversity among base learners is a key factor in the success of ensemble methods for regression tasks. Diverse models are less likely to make the same errors, leading to improved overall performance when their predictions are combined.
This document lists various diversity metrics used to measure diversity in regression ensembles, categorized into **pairwise** and **non-pairwise** metrics.

1 Pairwise Metrics

These metrics are calculated between pairs of regressors to assess the diversity between them.

1.1 Correlation Coefficient (Pearson's ρ)

Measures the linear correlation between the outputs (or errors) of two regressors.

Formula:

$$\rho_{ij} = \frac{\text{cov}(y_i, y_j)}{\sigma_{y_i}\sigma_{y_j}} \tag{1}$$

where y_i and y_j are the outputs of regressors i and j, and σ_{y_i} is the standard deviation of y_i.

1.2 Mean Squared Difference (MSD)

Calculates the average squared difference between the outputs of two regressors.

Formula:

$$\text{MSD}_{ij} = \frac{1}{N}\sum_{k=1}^{N}(y_{i,k} - y_{j,k})^2 \tag{2}$$

where N is the number of data points.

1.3 Mean Absolute Difference (MAD)

Computes the average absolute difference between the outputs of two regressors.

Formula:

$$\text{MAD}_{ij} = \frac{1}{N}\sum_{k=1}^{N}|y_{i,k} - y_{j,k}| \tag{3}$$

1.4 Error Correlation

Calculates the correlation between the errors of two regressors.

Formula:

$$\rho_{\text{error},ij} = \frac{\text{cov}(e_i, e_j)}{\sigma_{e_i}\sigma_{e_j}} \tag{4}$$

where $e_i = y_{\text{true}} - y_i$.

1.5 Pairwise Diversity Measures

General term for metrics computed between pairs of regressors (e.g., Pairwise Correlation, MSD, MAD).

1.6 Disagreement Measure (Thresholded)

Calculates the proportion of instances where the difference between two regressors exceeds a certain threshold θ.

Formula:

$$\text{Disagreement}_{ij} = \frac{1}{N}\sum_{k=1}^{N} \delta\left(|y_{i,k} - y_{j,k}| > \theta\right) \tag{5}$$

where $\delta(\cdot)$ is the indicator function.

1.7 Rank Correlation (Spearman's ρ)

Measures the correlation between the ranks of the outputs of two regressors.

Formula:

$$\rho_{\text{rank},ij} = 1 - \frac{6\sum_{k=1}^{N} d_k^2}{N(N^2-1)} \tag{6}$$

where d_k is the difference between the ranks of $y_{i,k}$ and $y_{j,k}$.

1.8 Mutual Information

Quantifies the amount of information obtained about one regressor's output through the other.

Conceptual Approach: Estimate the mutual information $I(y_i; y_j)$ between the outputs.

1.9 Kullback-Leibler Divergence (KL Divergence)

Measures the difference between the probability distributions of the outputs of two regressors (applicable when outputs are probabilistic).

Formula:

$$D_{\text{KL}}(P_i||P_j) = \sum_k P_i(k)\log\left(\frac{P_i(k)}{P_j(k)}\right) \tag{7}$$

where $P_i(k)$ is the probability of regressor i predicting value k.

1.10 Yule's Q Statistic (Adapted for Regression)

Measures the association between the outputs of two regressors using a contingency table approach.

Adapted Formula:

$$Q_{ij} = \frac{ad - bc}{ad + bc} \tag{8}$$

where a, b, c, and d are counts based on thresholded outputs.

1.11 Covariance Error Measure

Computes the covariance of errors between two regressors.

Formula:

$$\text{Cov}(e_i, e_j) = \frac{1}{N}\sum_{k=1}^{N}(e_{i,k} - \bar{e}_i)(e_{j,k} - \bar{e}_j) \tag{9}$$

where $\bar{e}_i$ is the mean error of regressor i.

1.12 Partial Correlation Coefficient

Measures the correlation between two regressors' outputs while controlling for the target variable.

Formula:

$$\rho_{ij\cdot y_{\text{true}}} = \frac{\rho_{ij} - \rho_{iy_{\text{true}}}\rho_{jy_{\text{true}}}}{\sqrt{(1-\rho_{iy_{\text{true}}}^2)(1-\rho_{jy_{\text{true}}}^2)}} \tag{10}$$

1.13 Double-Fault Measure (Adapted for Regression)

Counts instances where both regressors have large errors simultaneously.

Conceptual Approach: Identify and count co-occurring high-error cases where $|e_{i,k}| > \epsilon$ and $|e_{j,k}| > \epsilon$ for a threshold ϵ.

2 Non-Pairwise Metrics

These metrics assess the diversity of the ensemble as a whole without focusing on specific pairs of regressors.

2.1 Variance of Outputs

Measures the variance among the predictions of all base regressors for each data point.

Formula:

$$\mathrm{Var}(Y_k) = \frac{1}{M} \sum_{i=1}^{M} (y_{i,k} - \bar{y}_k)^2 \tag{11}$$

where M is the number of regressors and $\bar{y}_k$ is the mean prediction for data point k.

2.2 Ensemble Variance (Ambiguity)

Represents the expected disagreement between base learners and the ensemble prediction.

Formula:

$$\text{Ambiguity} = \frac{1}{N} \sum_{k=1}^{N} \left(\frac{1}{M} \sum_{i=1}^{M} (y_{i,k} - \bar{y}_k)^2 \right) \tag{12}$$

2.3 Negative Correlation Learning (NCL)

Encourages base learners to be negatively correlated in terms of their errors through modifications in the learning process.

Conceptual Approach: Adjust the loss function to include terms that penalize positive correlations between errors:

$$L_{\text{NCL}} = \sum_{i=1}^{M} L_i - \lambda \sum_{i=1}^{M} \sum_{j \neq i} \mathrm{cov}(e_i, e_j) \tag{13}$$

where L_i is the loss of regressor i and λ is a regularization parameter.

2.4 Coefficient of Variation (CV)

The ratio of the standard deviation to the mean of the outputs for each data point, highlighting relative variability.

Formula:

$$\mathrm{CV}_k = \frac{\sigma_{Y_k}}{\bar{Y}_k} \tag{14}$$

where σ_{Y_k} is the standard deviation of predictions for data point k.

2.5 Diversity Density

Considers the density of models' outputs in the output space; high density implies low diversity.

Conceptual Approach: Estimate the density of predictions in the output space and assess diversity inversely proportional to this density.

2.6 Error Variance

Variance of the errors across all base regressors, reflecting diversity in model errors.

Formula:

$$\mathrm{Var}(E_k) = \frac{1}{M}\sum_{i=1}^{M}(e_{i,k} - \bar{e}_k)^2 \tag{15}$$

where $\bar{e}_k$ is the mean error for data point k.

2.7 Ambiguity Decomposition

Decomposes ensemble error into bias, variance, and ambiguity components, with ambiguity representing diversity.

Formula:

$$E\left[(\bar{y}_k - y_{\text{true},k})^2\right] = \frac{1}{M}\sum_{i=1}^{M} E\left[(y_{i,k} - y_{\text{true},k})^2\right] - \text{Ambiguity} \tag{16}$$

where

$$\text{Ambiguity} = \frac{1}{M}\sum_{i=1}^{M} E\left[(y_{i,k} - \bar{y}_k)^2\right] \tag{17}$$