Calibrated Explanation with Local Interpretable Model-Agnostic Explanations

May 23, 2023

Algorithm 1: Calibrated Explanations using LIME with Conformal Prediction

Input : instance_of_interest
Output: calibrated_explanations

```
// Step 1: Generate local explanations using LIME
local_explanations ←
 LIME.generate_explanations(instance_of_interest);

// Step 2: Generate perturbed instances
perturbed_instances ←
 create_perturbed_instances(instance_of_interest);

// Step 3: Generate explanations for perturbed instances
   using LIME
perturbed_explanations ← [];
foreach perturbed_instance in perturbed_instances do
    perturbed_explanation ←
     LIME.generate_explanations(perturbed_instance);
    perturbed_explanations.append(perturbed_explanation);
end

// Step 4: Apply conformal prediction to the collection of
   local explanations
prediction_regions ← confor-
 mal_prediction.construct_prediction_regions(perturbed_explanations);


// Step 5: Calibrate the explanations using prediction
   regions
calibrated_explanations ← [];
foreach local_explanation in local_explanations do
    calibration_level ←
     prediction_regions.estimate_calibration_level(local_explanation);
    calibrated_explanation ← calibrate_explanation(local_explanation,
     calibration_level);
    calibrated_explanations.append(calibrated_explanation);
end

Function calibrate_explanation(local_explanation, calibration_level):
    calibrated_explanation ← adjust_explanation(local_explanation,
     calibration_level);
    return calibrated_explanation;

Function adjust_explanation(local_explanation, calibration_level):
    adjusted_explanation ← make_adjustments(local_explanation,
     calibration_level);
    return adjusted_explanation;
```