Inspecting the 2D regression dataset

Write short (~1 sentence) answers to the questions below to complete the TODOs:

1. What are your observations about where the train data and test data lie relative to each other?

   ANSWER: The train data is much more scattered than the test data, especially in the middle of the graph. At the ends of the graph, there is no data to train on.

2. What, if any, areas do you expect to have high/low aleatoric (data) uncertainty?

   ANSWER: I expect the data in the range [-2, 0] to have higher aleatoric uncertainty and the data on the ends to have less. This is because there is more noise in the range [-2, 0].

3. What, if any, areas do you expect to have high/low epistemic (model) uncertainty?

   ANSWER: I expect low epistemic uncertainty within the range [-4, 4] because there is lots of data for a model to learn from. Outside that range, there appears to be no data, and therefore there will be high epistemic uncertainty.

Analyzing the performance of standard regression model

Write short (~1 sentence) answers to the questions below to complete the TODOs:

1. Where does the model perform well?

   ANSWER: The model appears to be close to perfectly accurate anywhere where there is data to train on and even in the range [-2, 0] where there appears to be significant noise.

2. Where does the model perform poorly?

   ANSWER: Where there is no data to train on, the predictions diverge significantly. This appears in the ranges [-6, -4] and [4, 6].

Evaluating bias with wrapped regression model

Write short (~1 sentence) answers to the questions below to complete the TODOs:

1. How does the bias score relate to the train/test data density from the first plot?

   ANSWER: The higher the density of data for a given input, Capsa estimates higher bias.

2. What is one limitation of the Histogram approach that simply bins the data based on frequency?

   ANSWER: It doesn't account for other ways that the data may misrepresent reality. For example, there could be noise in the data or there could be missing data for important inputs.

Estimating aleatoric uncertainty

Write short (~1 sentence) answers to the questions below to complete the TODOs:

1. For what values of x is the aleatoric uncertainty high or increasing suddenly?

   ANSWER: The aleatoric uncertainty increases as there is visibly more noise in the model. The highest levels of aleatoric uncertainty are in the range [-2, 0].

2. How does your answer in (1) relate to how the x values are distributed?

   ANSWER: This appears to be independent of how the x values are distributed. It has more to do with how consistently given inputs are mapped to given outputs.

Estimating epistemic uncertainty

Write short (~1 sentence) answers to the questions below to complete the TODOs:

1. For what values of x is the epistemic uncertainty high or increasing

suddenly?

ANSWER: Where there is little to no input data. This is most extreme for the ranges [-6, -4] and [4, 6].

2. How does your answer in (1) relate to how the x values are distributed (refer back to original plot)? Think about both the train and test data.

   ANSWER: For the inputs where there is no data to train on but the model is tested for anyway, there is the most epistemic uncertainty.

3. How could you reduce the epistemic uncertainty in regions where it is high?

   ANSWER: Simply adding training data for those inputs would reduce the epistemic uncertainty.