Data621 - Blog4

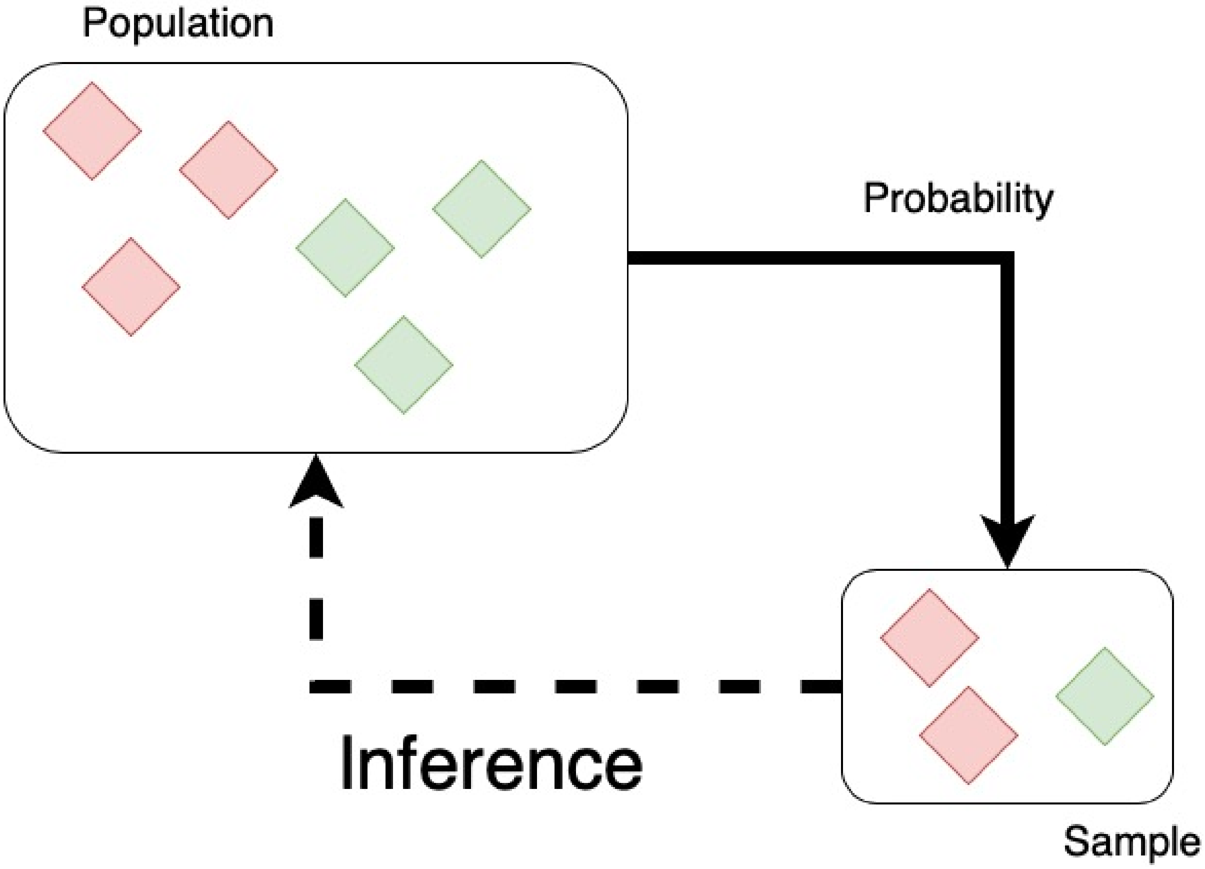
Amit Kapoor

4/15/2021

Inference vs Prediction

The terms Inference and Prediction are used extensively in data science community. Inference uses the model to understand and learn about the data generation process while Prediction uses the model to predict the outcomes for new sample. Inference is the information learned about the data generating process. On the other hand, prediction represents the output produced by a model of a data generating process in response to a specific set of inputs.

Inference



As shown above, lets consider a given population and we use probability to sample from that population. Once we get that sample, we're going to try to say something (or infer) about this global population. In other words, given a set of data inference is how the output is generated as a function of the data.

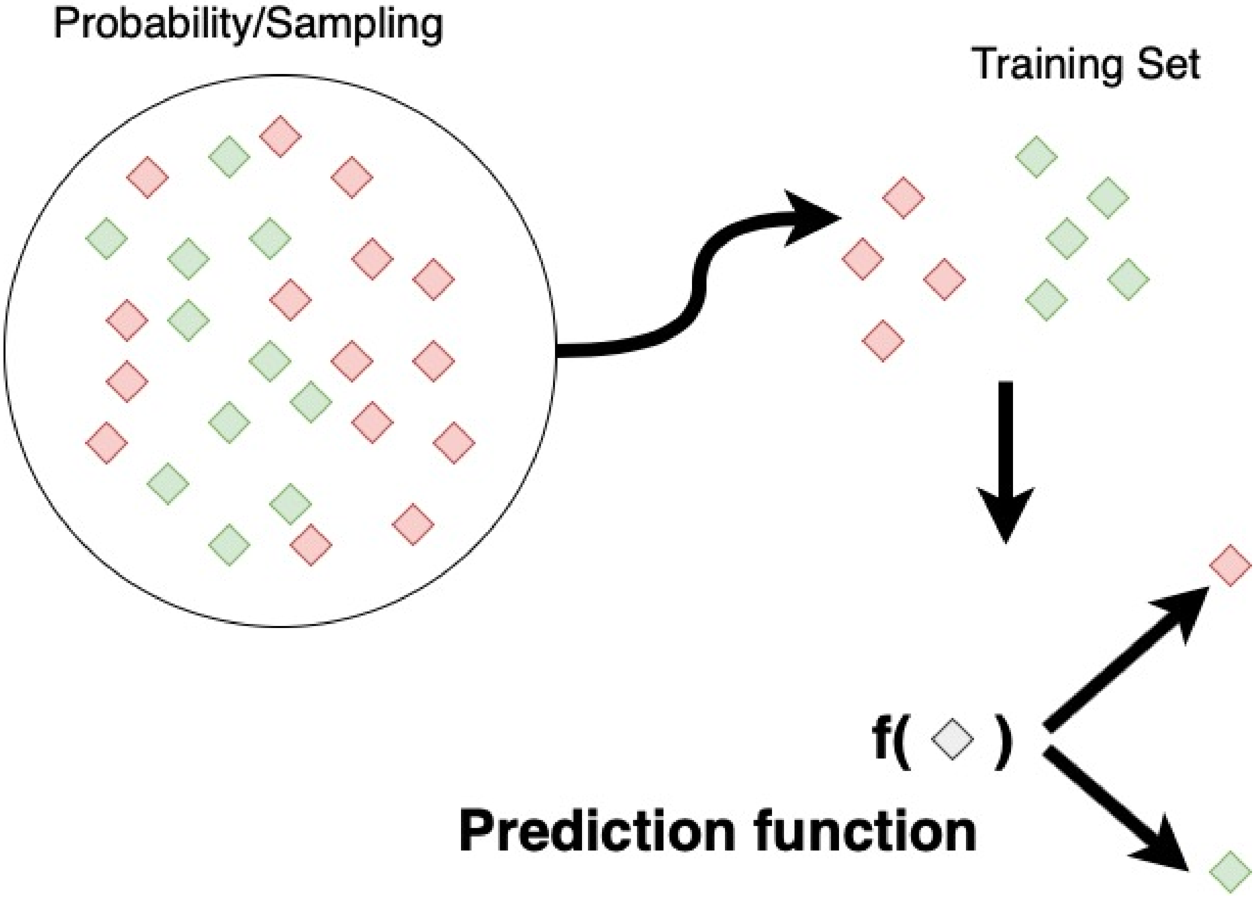
Lets consider a simple linear regression model

$\hat{y_i} = \beta_0 + \beta_1 \; x_i + \epsilon_i$

- x_i are set of independent variables that are observed and available to the model as input or predictors
- y_i is a dependent variable that is also observed and provided to the model as output or response.
- β are inferred parameters or coefficients that transform the inputs into the predictions.
- ϵ_i is the magnitude of typical errors in predictions.

Inference in this case would mean estimating the parameters of the model β_0 and β_1 that help understand the relationship between x and Y. The end goal is to estimate an association between response and predictor variables.

Prediction



Prediction is to take some sample from a population again and we build a training set, where we have different kinds of things to predict and then we use this training set data to build a prediction function. Once we have that prediction function, we provide a new sample to this function which in this case we dont know what color it is, that function assigns it to one of the two colors. Hence given a new sample, we use an existing data set to build a model that reliably chooses the correct identifier from a set of outcomes. Considering the equation again describe in previous section

$\hat{y_i} = \beta_0 + \beta_1 \; x_i + \epsilon_i$

- Predictions $\hat{y_i}$ are the output data, generated by the model they would be computed from the estimates of parameters β_0 and β_1

The goal is to tune a best model considering all predictors to predict Y from x.

Closing Thoughts

Inference based approach emphasizes more on the interpretation and understanding the results while in prediction based approach, we focus more on accurate results by predicting the output (dependent variable) using a fine tuned model.

References

Linear Models with R by Julian J. Faraway