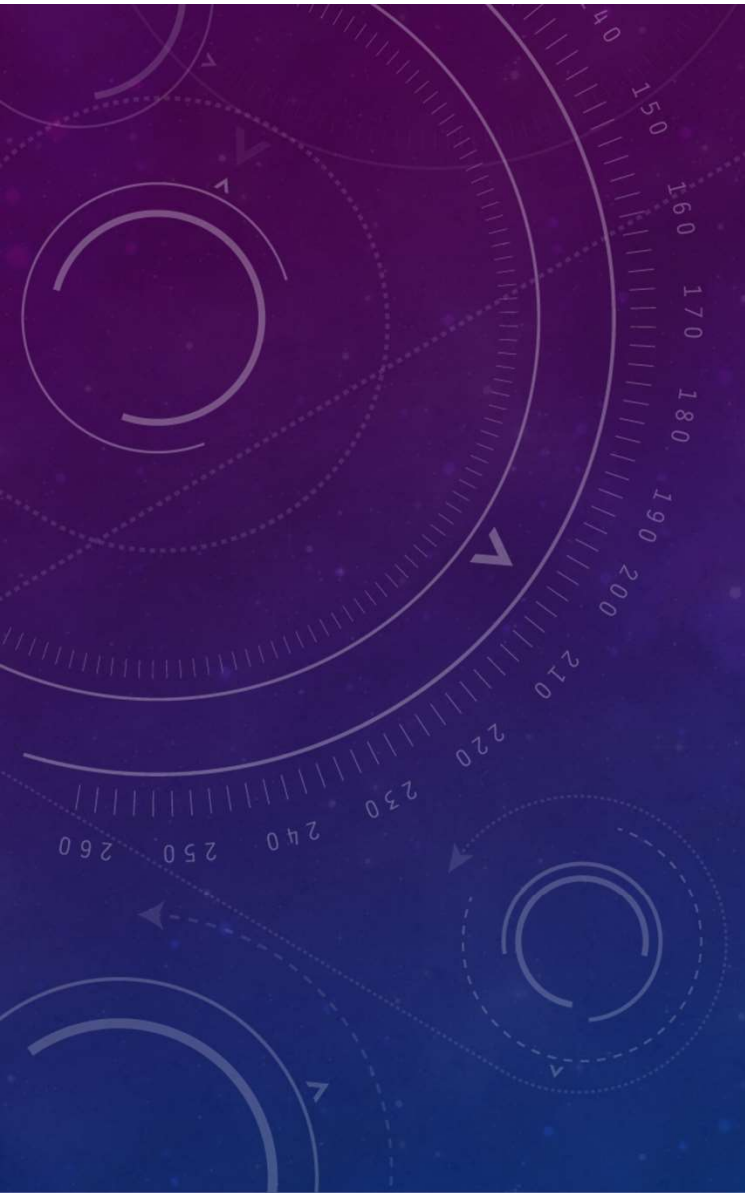

MACHINE LEARNING WITH .NET

Asmaliza Ahzan @ Emma

asmaliza@iverson.com.my

REGRESSION MODEL

CHAPTER 3

REGRESSION

- A supervised machine learning task that is used to predict the value of the label from a set of related features.
- The label can be of any real value and is not from a finite set of values as in classification tasks.
- Regression algorithms model the dependency of the label on its related features to determine how the label will change as the values of the features are varied.
- The input of a regression algorithm is a set of examples with labels of known values.
- The output of a regression algorithm is a function, which you can use to predict the label value for any new set of input features.

EXAMPLES OF REGRESSION SCENARIOS

- Predicting house prices based on house attributes such as number of bedrooms, location, or size.
- Predicting future stock prices based on historical data and current market trends.
- Predicting sales of a product based on advertising budgets.

REGRESSION TRAINERS

- LbfgsPoissonRegressionTrainer
- LightGbmRegressionTrainer
- SdcaRegressionTrainer
- OlsTrainer
- OnlineGradientDescentTrainer
- FastTreeRegressionTrainer
- FastTreeTweedieTrainer
- FastForestRegressionTrainer
- GamRegressionTrainer

CHOOSING THE TYPE OF REGRESSION MODEL

- The type of regression model you choose depends on what your expected output is.
- It also depends on the problem you are trying to solve, the characteristics of your data, and the compute and storage resources you have available.
- In summary;
  - If your output is a Boolean value, use a logistic regression model.
  - If your output is comprised of a preset range type of values (akin to an enumeration), use a logistic regression model.
  - If your output is a numeric unknown value, use a linear regression model.

CHOOSING A LINEAR REGRESSION TRAINER

- For ML.NET linear regression trainers, by and large, the most popular are FastTree and LightGBM.
- The three FastTree algorithms utilize neighbor-joining and use heuristics to quickly identify candidate joins to build out a decision tree.
- LightGBM is a very popular linear regression algorithm that utilizes a Gradient-based One Side Sampling (GOSS) to filter out the data instances for finding a split value.
- Both trainers provide both quick training and predict times while also providing very accurate model performance.

CHOOSING A LOGISTIC REGRESSION TRAINER

- Are you looking to train and predict in a low memory environment? LbfgsLogisticRegressionBinaryTrainer is a logical choice given that it was created to handle memory-restricted environments.
- Both of the SDCA-based trainers SdcaLogisticRegressionBinaryTrainer and SdcaNonCalibratedBinaryTrainer - have been optimized for scalability in training. If your training set is large and you are looking for binary classification, either of the SDCA trainers would be a good choice.
- The SymbolicSgdLogisticRegressionBinaryTrainer model is different from the other three in that it is based on a stochastic gradient descent algorithm. This means rather than looking to maximize the error function, the algorithm looks to minimize the error function.

CREATING THE LINEAR REGRESSION APPLICATION

- As mentioned earlier, the application we will be creating is an employee attrition predictor.
- Given a set of attributes tied to an employee, we can predict how long they will remain at their current job.
- The attributes included in this example aren't a definitive list of attributes, nor should be used as-is in a production environment; however, we can use this as a starting point for predicting a singular numeric output based on several attributes.

STOCHASTIC DUAL COORDINATE ASCENT (SDCA)

- Starting with Stochastic, which, in other words, means unpredictability. And in the case of machine learning, it means attempting to probabilistically predict the error function and feed random samples from your training set into the optimizer.
- The use of Dual Coordinate means two variables are coupled when training the model. As you have probably guessed, this makes the model much more complex but doesn't require any extra work to be utilized.
- Lastly, Ascent refers to maximizing the value of the error function.

EVALUATION METRICS

Metrics	Descriptions
Loss Function	The function that computes the distance between the current output of the algorithm and the expected output.
Mean Absolute Error	Finds the average absolute distance between the predicted and target values.
Mean Squared Error	Finds the average squared error between the predicted and actual values.
R-Squared	A statistical measure of how close the data are to the fitted regression line. The higher the R-squared, the better the model fits your data.
Root Mean Square Error (RMSE)	Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are.

CREATING THE LOGISTIC REGRESSION APPLICATION

- The application we will be creating to demonstrate logistic regressions is a file classifier.
- Given a file (of any type), we extract the strings from the file.
- The trainer used in this application also uses SDCA but using the logistic regression variation that was discussed earlier in this chapter.

ACTIVITY

Predict prices using regression with ml.Net

THE END