Python - Machine Learning

Tuesday, August 30, 2022 4:00 PM

Feature Engineering

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm.

Data preprocessing is a technique that is used to convert the raw data into a clean data set.

Whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

Feature engineering is the process of extracting features from raw data using **math, statistics and domain knowledge**

Feature Selection

Selecting features before modeling our data is important:

- Reduces overfitting
- Improving accuracy
- Reduce training time

Techniques:

- Univariate selection
  - Statistical tests can be used to select features that have strongest relationship with the output
- Feature importance
  - Feature importance gives you a score for each feature of your data, the higher the score, more important or relevant is the feature towards your output variable.
- Correlation Matrix with Heatmap
  - Correlation states how the features are related to each other or the target variable.
  - Heatmap makes it easy to identify which features are most related to the target variable.

Machine learning

Machine learning is a method of data analysis that automates analytical model building. Using models that iteratively learn from data, machine learning allows computers to find hidden insights without being explicitly programmed where to look.

Supervised learning:

- We feed set of labeled data into some ML learning algorithm which then creates a model to fit this data to some outputs.
- In the business world a lot of machine learning models fall under this type.
  - Predicting which customers are most likely to leave our business
  - Predicting who might default on a load

Unsupervised learning:

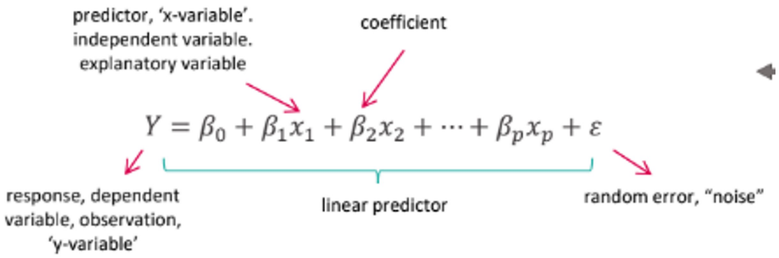
- Unsupervised learning is concerned with finding interesting clusters of input data. It does so without any help of data labelling.

Reinforcement learning:

- Reinforcement learning is a type of learning where an agent learns by receiving rewards and penalties.
- Unlike supervised learning, it is not given the correct label or answer. It is learning based on experience.

Regression Analysis

- Regression analysis is a statistical process for estimating the relationships among variables
- The predictor is continuous
- Relationship between a dependent variable and one or more independent variables(or predictors)
- Simple regression
    - Only one dependent
    - Only one independent
        - $y = b_0 + b_1 x_1$
- Multiple linear regression
    - Only one dependent
    - Many independent


predictor, 'x-variable'. independent variable. explanatory variable
coefficient
$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$
response, dependent variable, observation, 'y-variable'
linear predictor
random error, "noise"

Simple linear regression

- A linear regression is a statistical approach to modeling the relationship between a dependent variable(y) and one or more independent variables(x)
- Basically we want to know the regression equation that can be used to make predictions
- Our model uses linear predictor functions whose parameters (similar to the k and b in 'y=kx+b') are estimated using the training data
- Linear regressions are one of the most popular ML algorithms due to their simplicity and ease of implementation

Regression Model Evaluation

R^2:

- R-squared is a statistical measure of how close the data are to the fitted regression line.
- R-squared is always between 0% and 100%.
- The higher the R-squared, the better the model fits your data.

Adjusted R^2:

- The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model.
- R2 always increases when the number of variables increases.
- Adding useless variable to your model will increase the R2

Mean Absolute Error:

- The absolute error is the absolute value of the difference between the forecasted value and actual value.
- The mean absolute error(MAE) is the average of all absolute errors.
- MAE tells us how big of an error we can expect from the forecast on average.

Mean Squared Error:

- The mean squared error(MSE) of an estimator measures the average of the squares of the errors.

Root Mean Squared Error:

- Root mean squared error(RMSE) is the square root of the mean of the square of all of the error.
- The parameter indicates the standard deviation of the residuals or how far the points are from the regression or modelled line.

Overfitting

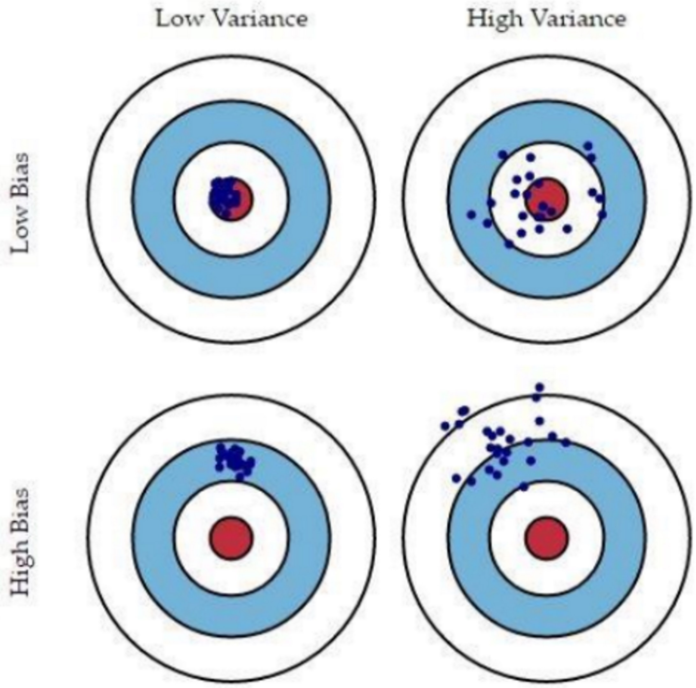
- Overfitting means that the model we trained has trained too well and fit too closely to the training dataset.
- Techniques to reduce overfitting:
  - Increase training data
  - Reduce model complexity
  - Ridge regularization and lasso regularization
  - Use dropout for neural networks to tackle overfitting

Underfitting

- Underfitting can happen when the model is too simple and means that the model does not fit the training data.
- Techniques to reduce underfitting:
  - Increase model complexity
  - Increase number of features, performing feature engineering
  - Remove noise from data
  - Increase number of epochs or increase the duration of training to get better results

Bias & Variance

- Bias is the difference between the average prediction of our model and the correct value which we are trying to predict. Model with high bias pays very little attention to the training data and oversimplifies the model. It always leads to high error on training and test data.
- Variance is the variability of model prediction for a given data point or a value which tells us spread of our data. Model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn’t seen before. As a result, such models perform very well on training data but has high error rates on test data.

- 
Low Variance
High Variance
Low Bias
High Bias