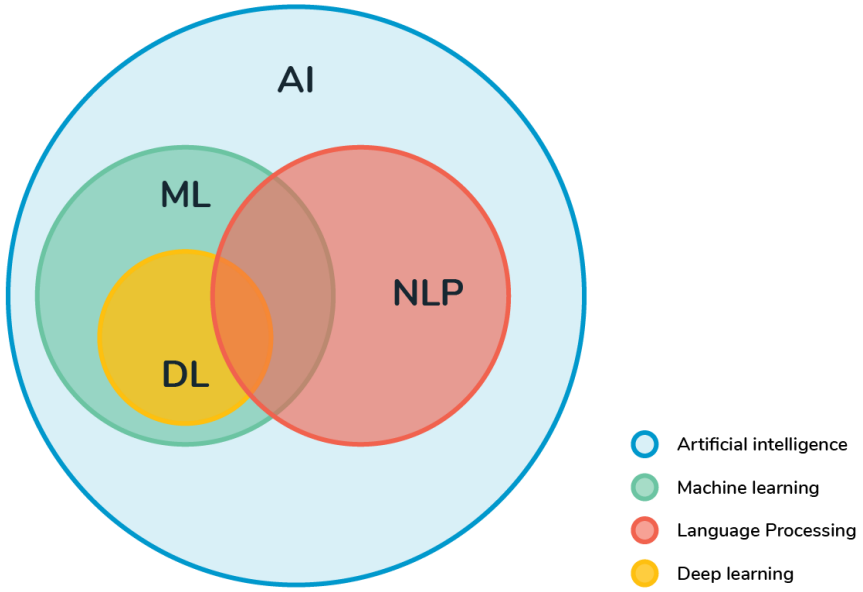
Language Detection AI

CKRAFT-BOT

AI – NLP



- Spam detection
- Fake new detection
- Language detection
- Chat bots
- Smart Assistance
- Sentiment analysis
- Text summarization
- Q&A
- Translation
- Transcription
- Spell check

Dataset

- Kaggle
- 22,000 rows
  - 22 Languages
  - 1000 texts/ lang

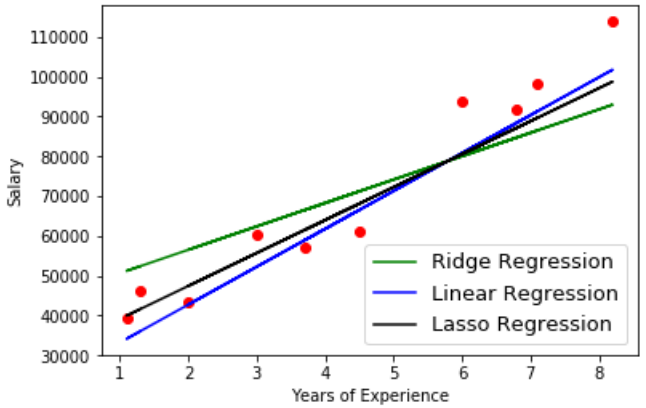
Project

- Import in dataset
- Quick review of dataset
  - Clean
- Training model
- Model eval (accuracy test)
- Observations

Training Process

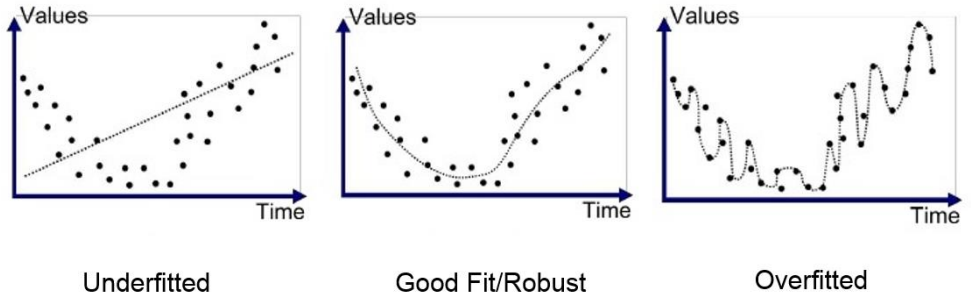
- Algs
  - Training model: Ridge Classifier
  - Model eval: (Multinominal) Naive Bayes classifier
- Model setup
  - Vectorization
    - Parsing
    - Convert text to numeric
    - Convert to matrix
      - Row X column
      - Language X unique vocabs (no dupes)
  - Matrixes → baseline for predication

Ridge Classifier



- Like linear regression
  - Best fit line
  - Strength of var relationships
    - IV → DV
  - Coefficients
    - Twinkle twinkle little stars
    - P Values
      - < 0.001 near perfect, high confidence
      - < 0.05 stat sig
      - > 0.05 not stat sig, null hyp

- But not linear regression
  - Mitigates under/overfitting by modifying weights through L2 regularization application
    - Takes the weights to near zero while not taking it to zero



Model Eval

- (Multinominal) Naive Bayes classifier (algorithm) is based off the Bayes Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- Naive Bayes classifier is based off the Bayes Theorem
- Pros of using this classifier alg: efficiency and ease
- Assumption: every value (feature) is independent or “unique”
- Calculates conditional probability
  - An event occurring given presumptions or evidence that have already occurred
  - Simply put likelihood of matching existing convention otherwise “naïve”