Overfitting and Bag of Words

Sean Kang

What is overfitting

- Overfitting occurs when the model too closely fits the training data, and then resulting in the model that cannot make accurate predictions with new data.
- When the model learns its training data too much, or learns the irrelevant noise of the data, then the model becomes too overfitted and cannot generalize to new data.
- If the model generalize to new data, then it is unable to predict or classify the task that it was designed for.

Underfitting versus Overfitting

- If overfitting is detected, perhaps the training should stop, or reduce complexity by removing some irrelevant inputs.
- If you pause too early in the model training, then the model is not trained enough on input variables that have a significant relationship between the input and output variables.

What happens with underfitting

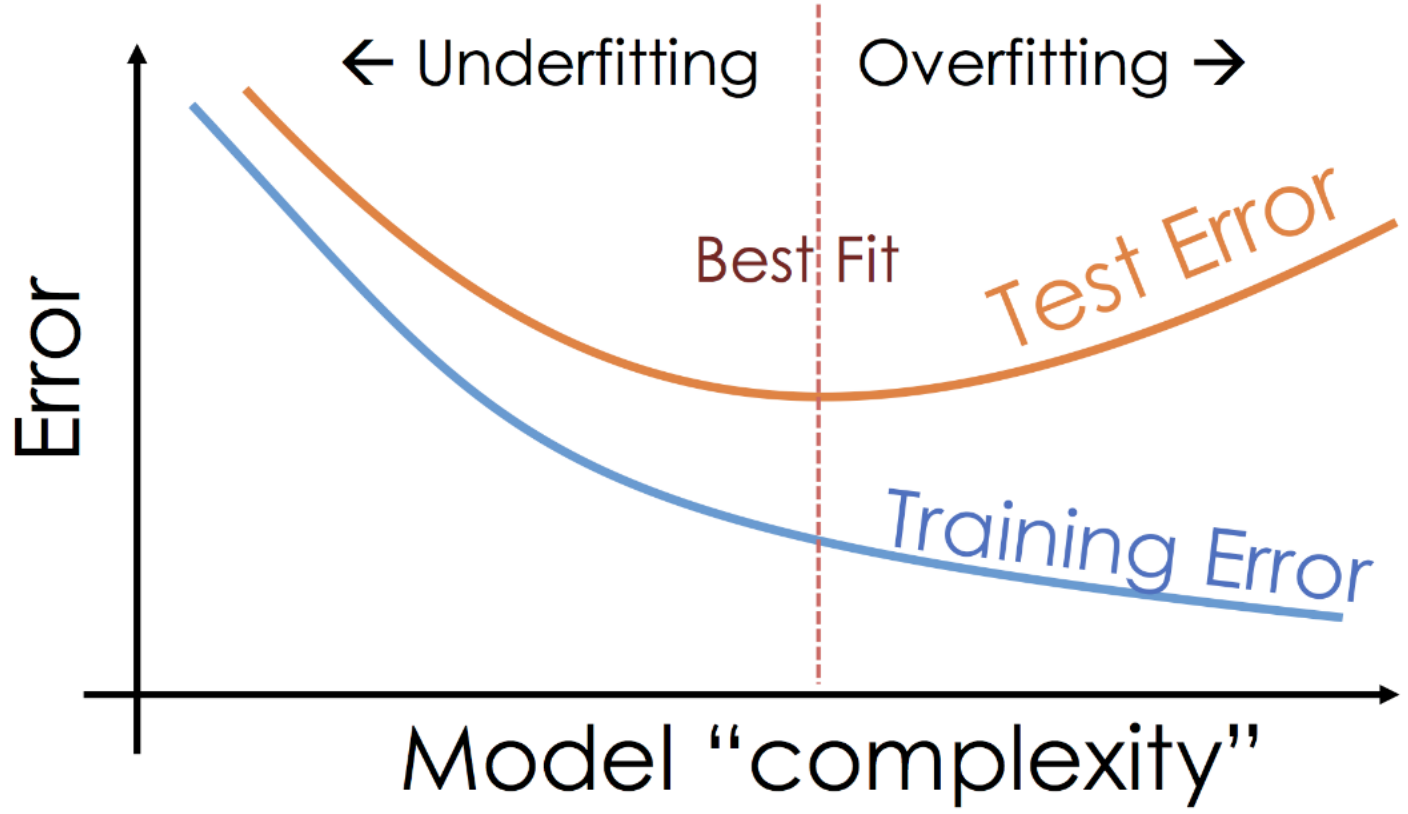
- Generalizes poorly on unseen data
- High bias
- Less variance

When the model learns

- Bias decreases
- Its variance can increase as it gets overfitted

We need to find “sweet spot”

Finding the Sweet Spot



How to detect overfitting?

- How to check for model fitness so that your model accuracy is understood.
- If the training data has no error rates (ie. low RMSE) and the test data has high error rates, then this is a sign of overfitting.
- The most popular technique is to use k-folds cross validation.

What is k-fold cross validation?

- Split the data into k pieces or folds.
- One of the folds is kept separate and used later for testing.
- The remaining folds are used for training.
- This process is repeated until each fold is used as the testing fold.
- At each iteration of the test, a score is kept.
- After all the iterations, an average score is computed to assess the model's performance.
- More on this in separate lecture.

How to avoid over-fitting?

- Using linear model helps us to avoid overfitting (however many real-world problems are non-linear)
- Train with more data – add more clean, relevant data
- Stop early – but can lead to underfitting
- Data Augmentation – some adding noisy data make model more stable – done carefully
- Feature Selection – Removing redundant, or irrelevant data, and adding most important data
- Regularization – this helps us to determine which inputs to remove from model (more later)
- Ensemble Methods – a set of classifiers (decision tree) to reduce variance in the datasets (more later)

Bag of Words – Another Feature Engineering

- A method for representing text data for machine learning or regression models
- Used with natural language processing (NLP)

Problem with Text

- Models and machine learning prefers numerical data, and fixed length structures.
- Text or language text isn’t

Bag of Words (BoW)

- Occurrence of words in a document
- Involves
  - Vocabulary of words
  - Measurement of the vocabulary words

Set of vocabulary words

- Some EDA or Data Cleaning Needed
  - Ignore capitalization
  - Ignore punctuation
  - Ignore stop words (but, or, and, of, a, etc)
  - Fix misspelling
  - Using root words (or stem words)
- Sample vocabulary
  - jump fox quick over the lazy dog cow over the moon brown

NLP

- Many open source natural language processing systems
- Java
  - Stanford NLP
    - https://nlp.stanford.edu/software/
      - POS – Part of Speech
      - This can process regular text input and tag each word with POS attributes
      - Are – to be, Were – to be, Running – to run, Ran – to run
  - Apache Open NLP
- Python
  - https://stanfordnlp.github.io/CoreNLP/

Define a vector based on vocabulary

- Based on this vocabulary:
  - jumps fox quick over the lazy dog cow moon brown
- Each document would have this vector:
  - The quick brown fox jumps over the lazy dog
    - [1, 1, 1, 1, 2, 1, 1, 0, 0, 1]
  - The cow jumps over the moon
    - [1, 0, 0, 1, 2, 0, 0, 1, 1, 0]
  - The cat in the hat
    - [0, 0, 0, 0, 2, 0, 0, 0, 0, 0] -> sample of sparse vector
    - Documents those known words are encoded. Other words are ignored
- The order of the words are ignored

How to Score

- Counts – absolute count of word in the document
- Frequency - frequency relative to other words in document (1 of 1000 words in document)