ARTIFICIAL INTELLIGENCE SOFTWARE DEVELOPMENT

Week 3 Lecture 1

Dr. Hari M Koduvely

Agenda for Today

- Theory: 5:30PM – 7:30PM
  - Recap of MLOPS from Term 1
  - Feature Engineering
- Lab: 7:30PM – 9:30PM
  - Standup Meetings

MLOps Recap

Summary of topics covered in course 1

- ❑ Data Engineering
- ❑ Training Data Generation

MLOps Recap

Summary of topics covered in course 1

- ❑ Data Engineering
  - Data Sources
  - Data Formats
  - Data Models
  - Modes of Dataflow
- ❑ Training Data Generation

MLOps Recap

Summary of topics covered in course 1

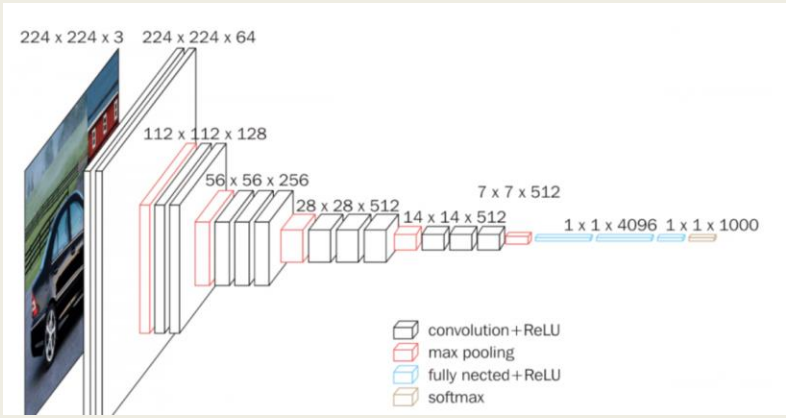
- ❑ Data Engineering
- ❑ Training Data Generation
  - Data Labelling
  - Sampling Techniques

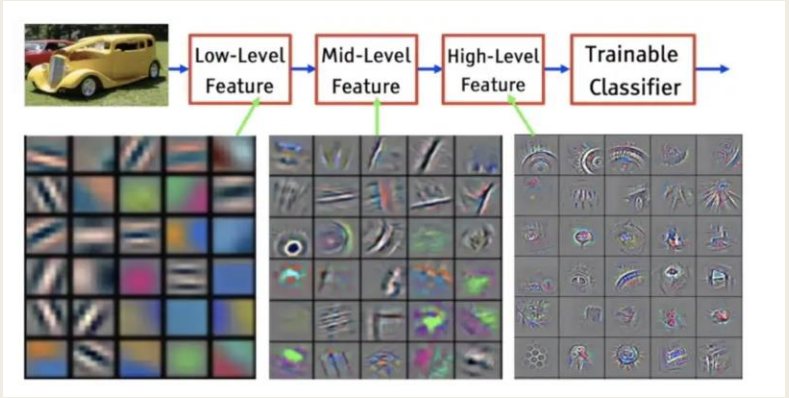
Feature Engineering

- ❑ Importance of Feature Engineering
- ❑ Handling Missing Values
- ❑ Feature Scaling
- ❑ Encoding Categorical Features
- ❑ Positional Embeddings
- ❑ Data Leakage
- ❑ Feature Selection

Importance of Feature Engineering

- Features are like Signals.
- Feature Engineering is like separating Signals from Noise.
- Deep Learning algorithms are capable of Learning Features themselves.

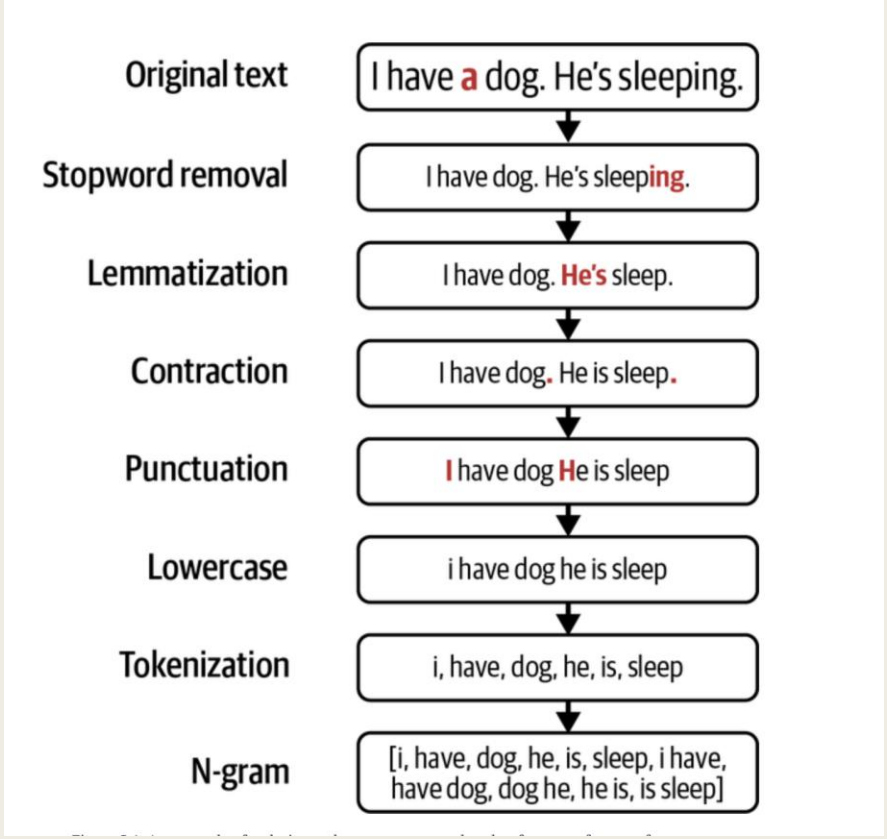




Source https://medium.com/analytics-vidhya/the-world-through-the-eyes-of-cnn-5a52c034dbeb

Importance of Feature Engineering

- Features are like Signals.
- Feature Engineering is like separating Signals from Noise.
- Deep Learning algorithms are capable of Learning Features themselves.
- Classical ML algorithms requires manual creation of useful features



Source https://medium.com/analytics-vidhya/the-world-through-the-eyes-of-cnn-5a52c034dbeb

Handling Missing Values

❑ Not all types of missing values are the same

- ✅ **Missing Completely at Random (MCAR)**
  - The probability of a value being missing is **unrelated** to both the **observed** and **unobserved** data.
  - This means the missingness is purely random.
  - **Generally the least problematic type**, as the missing data does not introduce bias and can be safely ignored under many standard analyses (though it reduces power).

- ✅ **Missing at Random (MAR)**
  - The probability of a value being missing is **related to other observed variables** in the dataset, but **not** to the value of the missing data itself.
  - Can be handled using statistical methods like **multiple imputation** or **maximum likelihood**, which model the observed relationships.

- ✅ **Missing Not at Random (MNAR)**
  - The probability of a value being missing is **related to the unobserved (missing) data itself**.
  - This is the **most challenging type**, as the missingness introduces **systematic bias** and can't be addressed using only the observed data.
  - **Requires more complex statistical methods**, such as **Bayesian models**, **pattern-mixture models**, or sensitivity analyses.

ID	Age	Gender	Annual income	Marital status	Number of children	Job
1		A	150,000		1	Engi
2	27	B	50,000			Teac
3		A	100,000	Married	2	
4	40	B			2	Engi
5	35	B		Single	0	Doct
6		A	50,000		0	Teac
7	33	B	60,000	Single		Teac
8	20	B	10,000			Stud

Handling Missing Values

- Not all types of missing values are the same

- ✅ **1. Missing Completely at Random (MCAR)**
  - **Example:**
    In a survey, some respondents accidentally skip a question about their favorite color due to a printing error on some forms.
  - **Key point:** The missingness is completely unrelated to any variable — observed or unobserved.
  - **Effect:** No bias is introduced; the missingness is pure randomness.

- ✅ **2. Missing at Random (MAR)**
  - **Example:**
    In a medical study, younger participants are more likely to skip reporting their income. However, the participants' ages are recorded.
  - **Key point:** The missingness (of income) depends on an **observed variable** (age), but not on the actual income value.
  - **Effect:** Bias can be corrected using statistical methods that incorporate the observed variable (age).

- ✅ **3. Missing Not at Random (MNAR)**
  - **Example:**
    In a mental health survey, people with severe depression are less likely to answer questions about their mental state.
  - **Key point:** The missingness depends on the **missing value itself** (level of depression).
  - **Effect:** This introduces bias that **cannot be corrected** using just the available data. Advanced modeling or strong assumptions are needed.

ID	Age	Gender	Annual income	Marital status	Number of children	Job
1		A	150,000		1	Engi
2	27	B	50,000			Teac
3		A	100,000	Married	2	
4	40	B			2	Engi
5	35	B		Single	0	Doct
6		A	50,000		0	Teac
7	33	B	60,000	Single		Teac
8	20	B	10,000			Stud

Handling Missing Values

- How to treat missing values in the data?
- Deletion:
  - Column Deletion
  - Row Deletion
  - Easy to implement
  - Can leads to accuracy loss

ID	Age	Gender	Annual income	Marital status	Number of children	Job
1		A	150,000		1	Engi
2	27	B	50,000			Teac
3		A	100,000	Married	2	
4	40	B			2	Engi
5	35	B		Single	0	Doct
6		A	50,000		0	Teac
7	33	B	60,000	Single		Teac
8	20	B	10,000			Stud

Handling Missing Values

- How to treat missing values in the data?
- Imputation:
  - Mean
  - Median
  - Mode
  - Interpolation (e.g. KNN)
  - Can create Bias in the data
  - Can cause Data Leakage

ID	Age	Gender	Annual income	Marital status	Number of children	Job
1		A	150,000		1	Engi
2	27	B	50,000			Teac
3		A	100,000	Married	2	
4	40	B			2	Engi
5	35	B		Single	0	Doct
6		A	50,000		0	Teac
7	33	B	60,000	Single		Teac
8	20	B	10,000			Stud

Feature Scaling

- Natural scale of different features are not same
- ML algorithm does not know this
- Make all features in the same numerical range
- For the range [0,1] scale factor:

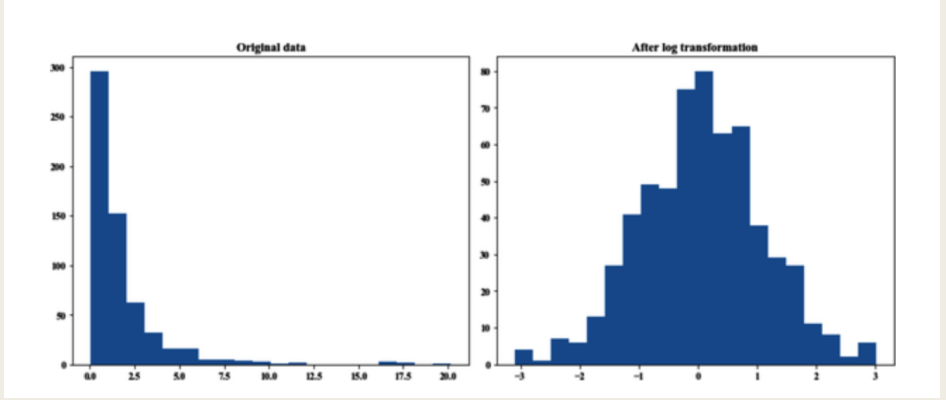
  [x – min(x)] / [max(x) - min(x)]

ID	Age	Gender	Annual income	Marital status	Number of children	Job
1		A	150,000		1	Engi
2	27	B	50,000			Teac
3		A	100,000	Married	2	
4	40	B			2	Engi
5	35	B		Single	0	Doct
6		A	50,000		0	Teac
7	33	B	60,000	Single		Teac
8	20	B	10,000			Stud

Feature Scaling

- Not all features would be having Normal Distribution
- Transform the features to make them Normal
  - Box-Cox transformation

    x' = (x^a - 1) / a   for a != 0
       = log(x) for a = 0



Discretization

- Process of converting continuous features to discrete features
- Aka Quantization or Binning
- Example:
  - Create bucket for ages: 0-10, 10-18, 18-30, 30-50, 50-65, 65-80, 80+
  - Need to be careful when choosing value of the boundaries
  - Use Histogram plots

ID	Age	Gender	Annual income	Marital status	Number of children	Job
1		A	150,000		1	Engi
2	27	B	50,000			Teac
3		A	100,000	Married	2	
4	40	B			2	Engi
5	35	B		Single	0	Doct
6		A	50,000		0	Teac
7	33	B	60,000	Single		Teac
8	20	B	10,000			Stud

Encoding of Categorical Features

- Some categorical features can have very large number of values
- And new values can appear in the production scenario unseen during training
- Examples:
  - IP addresses
  - Zip codes
  - Brand names

Embeddings

- Numerical vector representation of a categorical variable
- Words having similar semantics would be closer
- Preserves semantic relationships

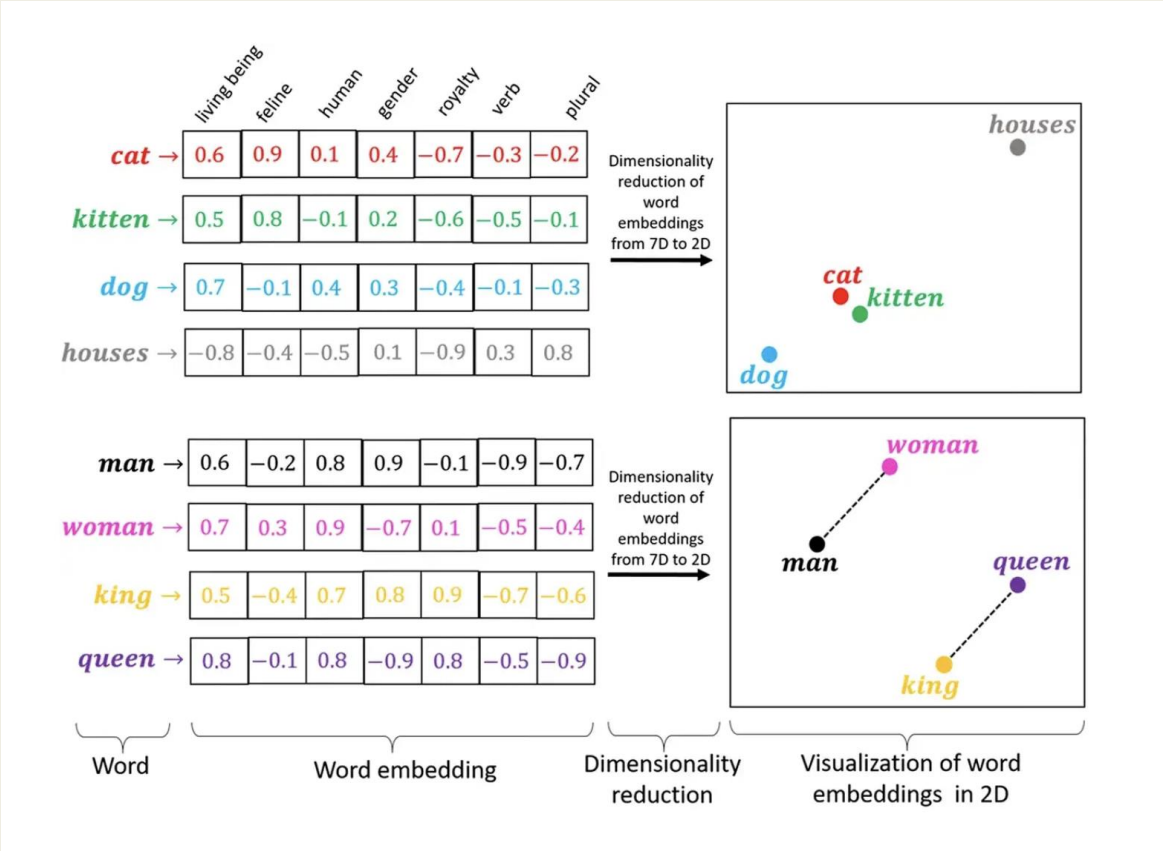


Image source https://medium.com/@hari4om/word-embedding-d816f643140

Embeddings

- Popular word embeddings for NLP
  - Word2Vec
  - Glove
  - Sentence Transformers

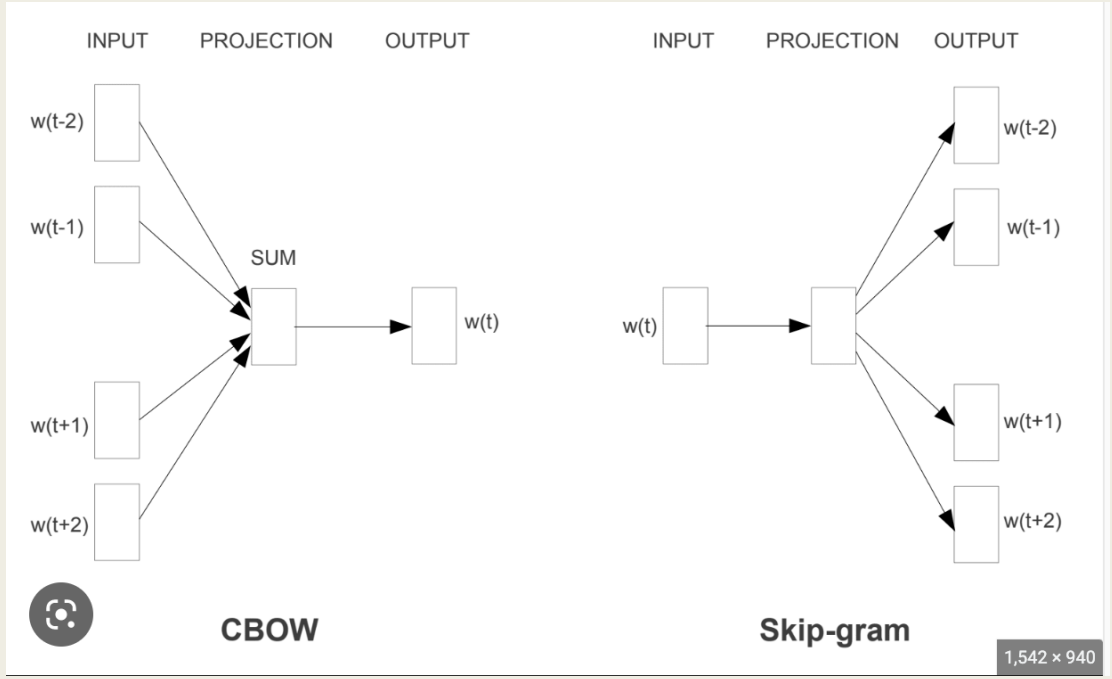


Image source https://medium.com/@hari4om/word-embedding-d816f643140

Data Leakage

- Training ML model using information not expected to be available during prediction.
- Examples:
  - **Feature Leakage** – caused by a feature which is a duplicate or proxy of the target variable.

    Monthly salary as feature to predict yearly salary.
  - **Sample Leakage** – Duplicate samples between train and test datasets.
  - **Non iid Data –** Splitting a time series dataset randomly

Common Causes of Data Leakage

- ❑ Filling in missing data before splitting.
- ❑ Not removing duplicates before splitting.
- ❑ Scaling before splitting.
- ❑ Splitting time-correlated data randomly instead of by time.
- ❑ Group leakage.

How to Detect Data Leakage

- ❑ Measure the correlation between each feature and target variable
- ❑ Investigate cases of very high correlation
- ❑ Measure the temporal correlation between train and test split

Feature Selection

- ❑ Adding more features leads to better model performance.
- ❑ Having too many features can have negative impacts also:
  - More chances of data leakage
  - Could cause overfitting
  - May require more memory to serve the model
  - Could increase latency at inference

Feature Selection

- Two factors to consider while selecting a feature:
  - Importance to the model
  - Generalization to unseen data

Feature Importance

Shapley Values

- Concept borrowed from Co-operative Game Theory (1950s).
- Invented by Lloyd Shapley.
- In ML also known as SHAP (SHapley Additive exPlanations)
- Used for fairly attributing a player's contribution to the end result of a game.
- Think of ML as a co-operative game by all the features to make a prediction.

Feature Importance

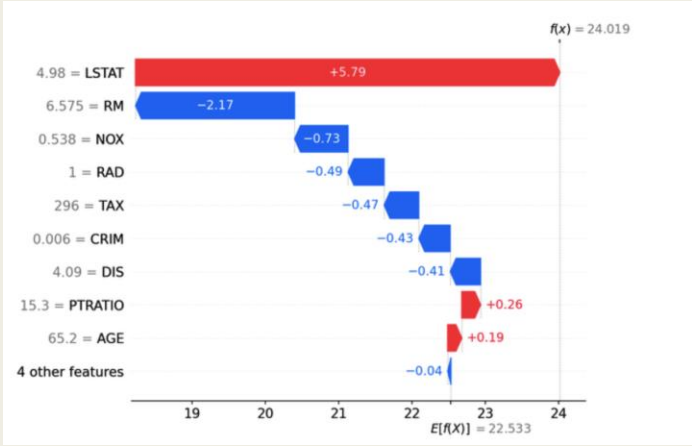
Shapley Values

- Computed by perturbing values of input features and measuring how it is changing the model prediction.
- The Shapley value of a given feature is the average marginal contribution to the overall model score.
- Can be used for both global importance and single prediction

Feature Importance

Shapley Values

Global Feature Importance



Single Prediction Feature Importance

Feature Generalization

- ML model should make accurate predictions on unseen data
- Measuring generalization capability of features is more difficult
- Two factors to consider for feature generalization:
  - Feature coverage
  - Distribution of feature values

Summary of Best Practices

- Split data by time into train/valid/test splits instead of doing it randomly.
- If you oversample your data, do it after splitting.
- Scale and normalize your data after splitting to avoid data leakage.
- Use statistics from only the train split, instead of the entire data, to scale your features and handle missing values.
- Understand how your data is generated, collected, and processed. Involve domain experts if possible.
- Keep track of your data's lineage.
- Understand feature importance to your model.
- Use features that generalize well.
- Remove no longer useful features from your models.

Feature Importance

Shapley Values

- Google Colab Notebook - Credit Risk Score Prediction