Dimensionality Reduction & Regularization Cheat Sheet

Dimensionality Reduction

Dimensionality Reduction is used to reduce the number of input variables in a dataset to improve efficiency and avoid overfitting.

1. PCA (Principal Component Analysis):
   - Linear method.
   - Projects data into principal components with highest variance.
   - Used for visualization and noise reduction.

2. TruncatedSVD:
   - Like PCA, but works on sparse matrices (e.g., TF-IDF).
   - Suitable for text data.

3. t-SNE (t-Distributed Stochastic Neighbor Embedding):
   - Non-linear technique for 2D/3D visualization.
   - Preserves local structure but not good for generalization.

4. UMAP (Uniform Manifold Approximation and Projection):
   - Faster than t-SNE, better global structure.
   - Useful for clustering and visualization.

Use Cases:

- Visualizing high-dimensional text data.

- Preprocessing step before classification.

Regularization

Regularization helps reduce overfitting by penalizing complex models.

1. L1 Regularization (Lasso):
   - Adds absolute value of coefficients to loss.
   - Can shrink some weights to zero (feature selection).

2. L2 Regularization (Ridge):
   - Adds square of coefficients to loss.
   - Shrinks all weights but doesn't eliminate features.

3. Elastic Net:
   - Combination of L1 and L2 penalties.
   - Balances feature selection and coefficient shrinkage.

4. Dropout (for neural nets):
   - Randomly sets a subset of activations to zero during training.

5. Early Stopping:
   - Stops training when validation error starts increasing.

Tips:

- Use GridSearchCV to tune regularization parameters (alpha, lambda).

- Apply regularization especially on high-dimensional data like TF-IDF vectors.