Universidad Politécnica de Yucatán

MACHINE LEARNING

SOLUTION TO MOST COMMON PROBLEMS IN ML

GABRIELA ELIZABETH AVILA CHAN
2009003

IRC 9°B

15/09/2023

- **Overfiting**: occurs when the machine learning model ~~program~~ learns the training data too well, to the point that it captures noise, random fluctuations, or outliers in the data, rather than the underlying patterns. As a result, an overfiting model perfoms very well on training data but poorly on unseen data because it has essentially memorized the training examples instead of learning the true underlying relationships. Signs of it:
  - Low training error (the model fits the training data very closely)
  - High validation or test error (poor performance on new data)
  - The model's predictions are too sensitive to small changes in the input data.

- **Underfiting**: occurs when a machine learning model is too simple to capture the underlying patterns in the training data. It fails to learn the relationships between the features and the target variable, resulting on poor performance on both the training data and new data. It lacks of the capacity to represent the complexity of the data. Signs of it:
  - High training error (the model cannot even fit training data)
  - High validation or test error
  - The mode's predictions do not capture the true patterns in the data.

- **Outliers**: data points that significantly deviate from the majority of the data in a dataset. Some of its key characteristics
  - Extreme values
  - Unusual or rare
  - Impact on summary statistics (can significantly affect summary statistics)
  - Visual identification (are often visually identifiable on plots)
  - Data errors (can sometimes be the result of data entry errors, ~~mesea~~ measurement errors)
  - Influence on models
  - Domain knowledge (to be sure when is an outlier or when should be treated as an error)

- Common solutions for them
  - Over fiting
    - Regularization = (apply techniques like L1 or L2 regularization to penalize larg models coef).
    - Cross-validation (use techniques like k-fold cross-validation to asses model performance)
    - Reduce model complexity.
    - Early stopping (monitor the models performance during training data and stop when performance degrade)
    - Ensemble methods
  - Underfiting
    - Increase model complexicy
    - Feature engineering (add more relevant features)
    - Colleet more data ~~set~~
    - Hyperparameter tuning
    - Ensemble methods
  - Presence of outliers
    - transformations (log-transformations or winsorization)
    - Robust models (use models that are less sensitive to outliers)
    - Imputation
    - Data cleaning
    - Contextual understanding

- Dimensionality problem: it happens when dealing with high-dimentional datasets or feature spaces. It encompasses various issues, ~~som~~ here's an overview of the dimensionarity problem:
  - Increased computational complexity
  - Data sparsity
  - Increased risk of overfitting
  - Difficulty in visualization
  - Loss of intuition

- Dimensionality reduction: is a process used in data analysis and machine learning to reduce the number of features (dimensions) in a dataset while preserving as much meaningful information as possible.
  1. Understand the problem & prepare the data by addressing missing values, outliers, and any other data quality issues.
  2. Data preprocessing →
  3. Feature scaling: normalize your features to ensure that they have similar scales
  4. Choose the right ~~tool~~ technique: (PCA or t-SNE)
  5. Apply the dimensionality Reduction technique
  6. Select the number of components
  7. Evaluate the reductionated data
  8. Use the reducted data
  9. Monitor model performance
  10. Iterate as needed

- Bias-variance trade-off: describe the balance between two types of errors a predictive model can make: bias and variance. it can be:
  - High bias, low variance
  - Low bias, high variance
  - Balancing bias and variance

Hertel, P. (n.d.). *Common problems with machine learning that companies face*. Hyperon.Io. Retrieved September 15, 2023, from https://www.hyperon.io/blog/common-problems-with-machine-learning-that-companies-face

Issues in Machine Learning. (n.d.). Www.javatpoint.com. Retrieved September 15, 2023, from https://www.javatpoint.com/issues-in-machine-learning