Topics

**Supervised Learning**

**Regression**

* **Definition**: Predicting continuous values.
* Linear Regression, Ridge Regression.

**Classification**

* **Definition**: Categorizing data into predefined labels.

1. **KNN (k-Nearest Neighbours)**: A non-parametric algorithm that classifies data points based on the majority class of their nearest neighbours in feature space.
2. **Logistic Regression**: A linear model used to predict probabilities using the sigmoid function **for binary / multi class classification**
3. **LDA (Linear Discriminant Analysis)**: A classification technique that assumes linear boundaries and models data by maximizing class separability.
4. **QDA (Quadratic Discriminant Analysis)**: Similar to LDA but allows for quadratic decision boundaries by assuming separate covariance matrices for each class.
5. **Decision Trees**: A tree-structured algorithm that splits data based on feature thresholds to predict target variables.

**Entropy Formula**

1. **Pruning**: The process of reducing a decision tree's size to prevent overfitting by removing less significant splits.
2. **Bagging**: An ensemble method that combines predictions of multiple models trained on different bootstrap samples to reduce variance.
3. **Random Forest**: An ensemble of decision trees built using bagging, adding randomness in feature selection to improve performance. **Used for classification and regression**
4. **Boosting**: An ensemble method that sequentially trains models to correct errors of previous models, improving accuracy iteratively.
5. **SVM (Support Vector Machine)**: A supervised algorithm that finds the optimal hyperplane to separate classes by maximizing the margin **can be used with**
6. **Kernel SVM**: An SVM that uses kernel functions to handle non-linear relationships in data by mapping it to higher dimensions.
7. **Tuning Parameters**: The process of optimizing hyperparameters to improve model performance without overfitting.
8. **Neural Nets**: Computational models inspired by the human brain, consisting of layers of interconnected nodes for feature learning.
9. **Convolutions**: Operations in convolutional neural networks (CNNs) that detect spatial patterns like edges in image data.
10. **Pooling**: A down sampling technique in CNNs that reduces spatial dimensions while preserving important features.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Method** | **Advantage** | **Technique** | **Disadvantage** | **Solution** |
| 1 | KNN |  | Distances | **Normalization**  Large Dimensions |  |
| 2 | Logistic Regression |  |  |  |  |
| 3 | LDA |  |  |  |  |
| 4 | QDA |  |  |  |  |
| 5 | Decision Trees | Catg & Cts Variables  interpretable,  Missing data, no standardization | Gini index, Misclassification error, Entropy | Instability  Additive structures | RF / Boosting  SVM |
| 6 | Pruning |  | Cross validate on cp = (1/Tree\_size) |  |  |
| 7 | Bagging | Improved variance, Bias no change | Bootstrap samples, Out Of Bag err | non interpretable, correlated variables | Variable importance |
| 8 | Random Forest | Improved variance than CART  Reduced correlation / overfitting than bagging | No of variables > tuning param |  |  |
| 9 | Boosting | Reduced Bias  Easy computing | Weighting based on classification  Minimize Exponential loss |  |  |
| 10 | SVM |  | Quadratic programming  Lagrange multipliers | Overlapping classes  **Normalization** | Tuning param SVM |
| 11 | Tuning param | Fit more data | Decrease margin | Linear fitting | Kernel SVM |
| 12 | Kernel SVM | Non-linear fitting | Hinge loss, Binomial Deviance |  |  |
| 13 | Neural Nets | Flexible,  Logistic / Linear regression | Back propagation | Multiple minima  Overfitting  **Normalization** | Weight decay  Early stopping  Scaling inputs |
| 14 | Convolutions | Fewer weights = filter\_size X filters |  |  |  |
| 15 | Pooling | shrink the layers leading to fewer weights, no weighing |  |  |  |

**Unsupervised Learning**

**Clustering**

* K-Means, Hierarchical Clustering, DBSCAN, Anomaly detection

**Dimensionality Reduction**

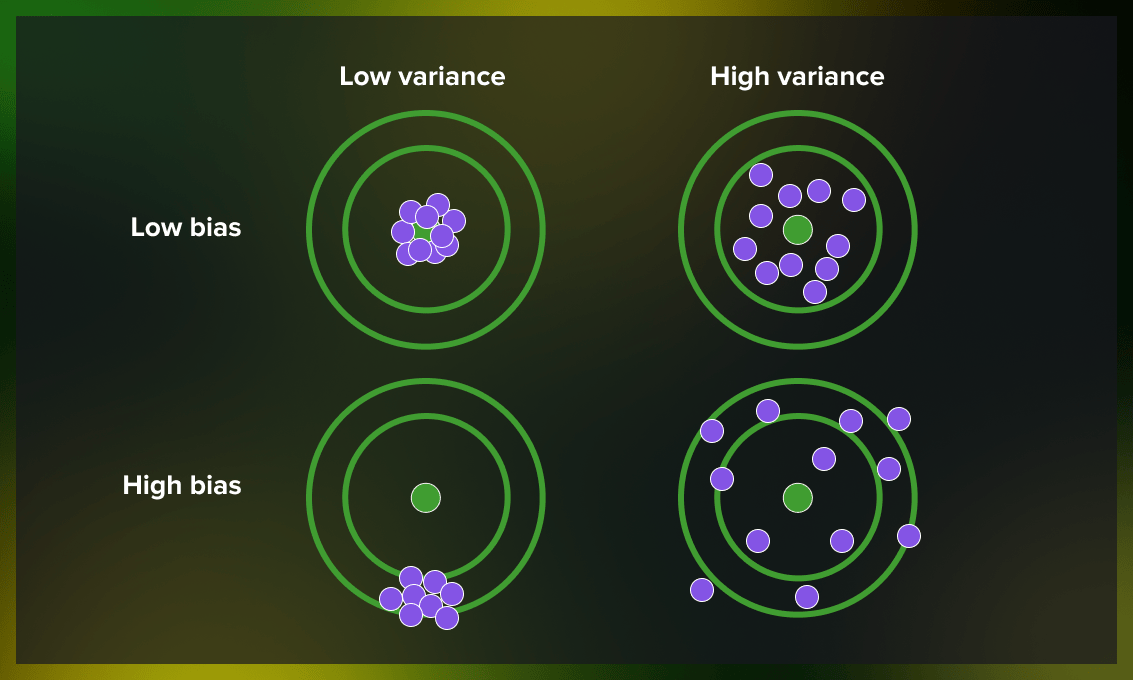
* PCA, t-SNE

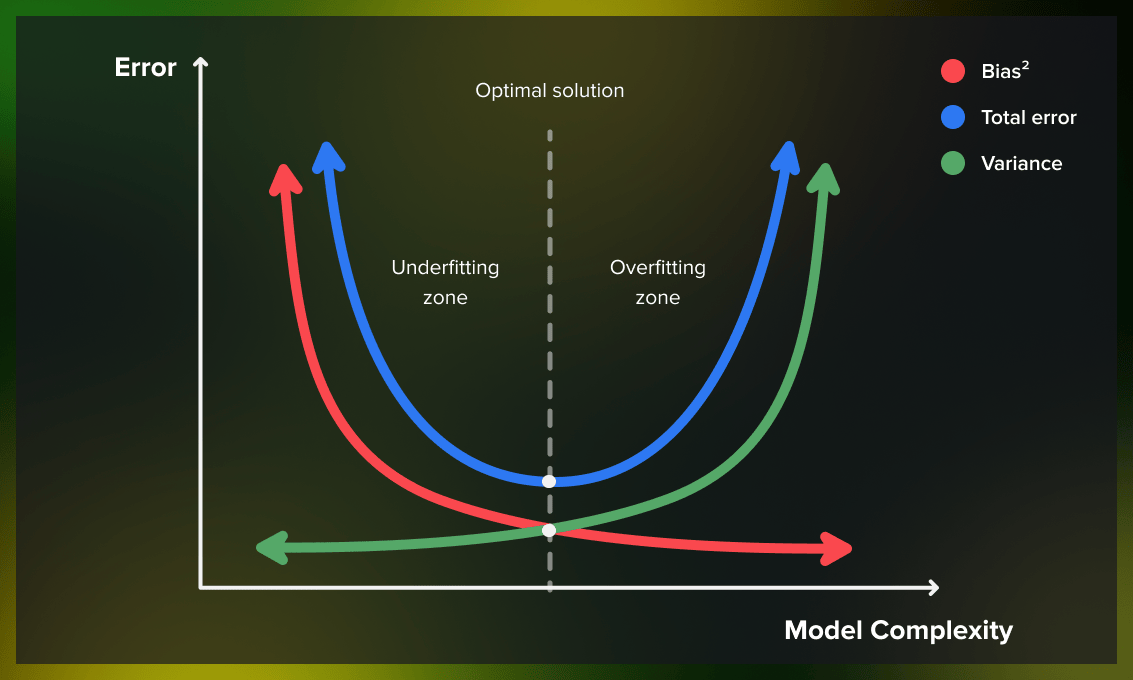
**8. Overfitting**

* **Definition**: Model fits the noise and details in the training data too closely.
* **Signs**: High accuracy on training data, low on testing.
* **Causes**: Excessively complex models with too many parameters.
* **Solution**: Use techniques like cross-validation, early stopping, and regularization.
* **Prevention**: Simplify the model or gather more data to improve generalization.

**9. Bias-Variance Trade-off**

* **Definition**: The balance between model bias (underfitting) and variance (overfitting).
* **High Bias**: Model is too simple, underfits the data.
* **High Variance**: Model is too complex, overfits the data.
* **Goal**: Minimize both to improve generalization to unseen data.
* **Solution**: Use techniques like cross-validation to find the right complexity.
* **Train error keeps decreasing**
* **Test error decreases then increases**





**10. Gradient Descent**

* **Definition**: An iterative optimization algorithm to minimize a loss function.
* **Steps**: Compute gradient, update parameters, repeat until convergence.
* **Variants**: Batch Gradient Descent (uses entire dataset), Stochastic Gradient Descent (uses one sample).
* **Learning Rate**: Controls the step size; too high can overshoot, too low may take long.
* **Applications**: Widely used in machine learning, especially neural networks.

Loss=Error Term+λ∑wi2​ L2 Regularization (Ridge Regression)

Loss=Error Term+λ∑∣wi​∣ L1 Regularization (Lasso Regression)

|  |  |  |
| --- | --- | --- |
| **Feature Selection** | Shrinks some coefficients to zero (selects features). | Retains all features by shrinking their magnitude. |

|  |  |  |
| --- | --- | --- |
| **Model Interpretability** | Higher (sparse model). | Lower (dense model). |

|  |  |  |
| --- | --- | --- |
| **Handling Correlated Features** | Arbitrarily chooses one feature among correlated ones. | Distributes weights across correlated features. |

|  |  |  |
| --- | --- | --- |
| **Optimization** | Non-differentiable at zero (slightly harder). | Fully differentiable (easier to optimize). |

|  |  |  |
| --- | --- | --- |
| **Best Use Case** | High-dimensional data with irrelevant features. | Problems with correlated features or dense data. |

**12. Cross-Validation**

* **Definition**: A technique to evaluate a model’s performance by splitting the data into subsets.
* **K-Fold Cross-Validation**: Split data into k subsets, train on k-1 subsets, test on the remaining one.
* **Leave-One-Out Cross-Validation (LOOCV)**: Special case of k-fold where k equals the number of data points.
* **Purpose**: Prevent overfitting and ensure model generalization.
* **Advantage**: Gives more reliable performance estimates than using a single training/test split.

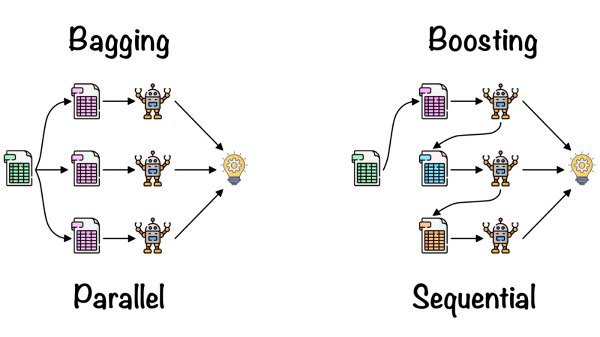
**13. Feature Engineering**

* **Definition**: The process of transforming raw data into features that better represent the problem.
* **Techniques**: Scaling, normalization, polynomial features, one-hot encoding.
* **Goal**: Improve the performance of machine learning models by adding new or refined features.
* **Impact**: Can drastically improve model accuracy.
* **Tools**: Libraries like sklearn provide preprocessing utilities for this.

**14. Hyperparameter Tuning**

* **Definition**: The process of finding the optimal set of hyperparameters that improve model performance.
* **Hyperparameters**: Parameters not learned by the model but set before training, e.g., learning rate, tree depth.
* **Techniques**: Grid Search (exhaustive), Random Search (random combinations), Bayesian Optimization (probabilistic).
* **Goal**: Improve the model's accuracy and efficiency.
* **Automation**: Can be automated using libraries like GridSearchCV, RandomizedSearchCV in sklearn.

**15. Ensemble Methods**



**1. Bagging (Bootstrap Aggregating)**

* **Definition**: An ensemble method that trains multiple models parallelly on different subsets of the data, then averages or votes their predictions. Averages predictions from several strong, diverse models.
* **Example**: Random Forest is a popular bagging algorithm.
* **Advantage**: Reduces variance and helps prevent overfitting.
* **Data Sampling**: Each model is trained on a random subset of the original dataset with replacement. Bootstrapped samples provide different perspectives on the data, allowing ensemble methods to generate diverse models that collectively perform better.
* **Usage**: Effective for high-variance models like decision trees.
* **OOB Error**: For each point, there are some trees which do not have that point in their bootstrap sample. Aggregate the result of these trees for this point and compare the prediction with the true value. This is OOB error.

**2. Boosting**

* **Definition**: Sequentially trains weak models, with each model correcting the errors of the previous one.
* **Example**: Algorithms like AdaBoost, Gradient Boosting, and XGBoost.
* **Advantage**: Reduces bias by focusing on difficult-to-predict examples.
* **Training**: Each subsequent model gives more importance to previously misclassified samples.
* **Usage**: Effective for low-bias models like decision trees; often used for both classification and regression.

**3. Stacking**

* **Definition**: Combines predictions from multiple models using a meta-model to generate a final prediction.
* **Example**: Logistic regression or linear regression used as the meta-model.
* **Advantage**: Can combine different types of base models (e.g., decision trees, SVMs, neural networks) to capture a diverse range of patterns.
* **Layers**: Predictions from base models are fed into a second-level model (meta-learner) for final predictions.
* **Usage**: Suitable for complex problems, can boost predictive power.

**4. Voting**

* **Definition**: Combines the predictions of multiple models by averaging (for regression) or majority voting (for classification).
* **Example**: Hard voting (majority rule) and soft voting (average of predicted probabilities).
* **Advantage**: Simple yet effective, can improve robustness by leveraging different models.
* **Model Diversity**: Works well when base models are diverse, capturing different aspects of the data.
* **Usage**: Commonly used in classification tasks; improves generalization.

**5. Random Forest (Bagging variant)**

* **Definition**: An ensemble of decision trees, where each tree is trained on a random subset of features and data.
* **Feature Sampling**: Each tree uses a random subset of features, increasing diversity among trees.
* **Advantage**: Decorrelate bagged trees, reduces both variance more and overfitting while maintaining interpretability.  
  A strong predictor can cause correlated trees.
* **Prediction**: Takes the majority vote (classification) or average (regression) of all decision trees.
* **Usage**: Works well with both classification and regression tasks, robust to noisy data and outliers.

**Model Evaluation**

Case Study