COURSE GLOSSARY

Supervised Learning with scikit-learn

Classification: A supervised learning task that assigns discrete labels or categories to observations, for example predicting churn vs. no churn

Confusion matrix: A table summarizing classification outcomes by counting true positives, true negatives, false positives, and false negatives for predicted vs. actual labels

Cross-validation: A resampling technique that splits the dataset into multiple training and validation folds to produce more reliable estimates of model performance

Decision boundary: The surface or line in feature space that separates regions where a classifier assigns different labels to observations

F1 score: The harmonic mean of precision and recall that balances both metrics into a single number, especially useful on imbalanced classes

Feature: An input variable used by a model to make predictions, also called a predictor or independent variable

Hyperparameter: A configuration value set before training (e.g., k in KNN or alpha in regularization) that controls aspects of a model's learning behavior and must be tuned externally

Entity: A real-world object or concept (for example, a professor or university) that is represented in a database by a table and its rows

ETL (Extract, Transform, Load): A data integration approach where data is extracted from sources, transformed into a target schema or cleaned format, and then loaded into the destination storage (commonly a data warehouse).

Lasso regression: A form of linear regression that adds an L1 penalty (absolute coefficients scaled by alpha) to the loss function and can set some coefficients exactly to zero, enabling feature selection

Machine learning: The process by which computers learn patterns and make decisions from data without being explicitly programmed to perform each task

Overfitting: When a model learns noise or idiosyncrasies in the training data and therefore performs well on training data but poorly on unseen data

Pipeline: A scikit-learn object that chains preprocessing transformers and an estimator into a single reproducible workflow so that data transformations and model fitting/prediction are applied consistently across training and evaluation

Precision: The fraction of predicted positive cases that are actually positive, measuring how many positive predictions are correct

Recall: The fraction of actual positive cases that are correctly identified by the model, also called sensitivity

Regression: A supervised learning task that predicts continuous numeric target values, such as predicting house prices or blood glucose levels

Regularization: A technique that modifies the model's loss function to penalize large parameter values in order to reduce overfitting and improve generalization

Ridge regression: A form of linear regression that adds an L2 penalty (squared coefficients scaled by alpha) to the loss function to shrink weights toward zero

ROC curve and AUC: The ROC curve plots true positive rate against false positive rate across decision thresholds, and AUC (area under the curve) summarizes this plot as a single value between 0 and 1 representing overall discriminative ability

Supervised learning: A type of machine learning where models are trained on labeled data to learn a mapping from input features to known target values for predicting unseen examples

Target variable: The output value the model is trying to predict, also called the response or dependent variable

Test set: A held-out portion of data used to evaluate a trained model's performance and estimate how well it generalizes to new data

Training set: The portion of data used to fit a model's parameters and learn relationships between features and the target

Underfitting: When a model is too simple to capture the underlying patterns in the data, resulting in poor performance on both training and test data

Unsupervised learning: A type of machine learning that discovers patterns or structure in unlabeled data, such as clustering similar observations without predefined targets