Machine Learning Study Guide (Theory-Only)

1. Outlier Detection

Concepts & Explanations:

• Outliers are points far from the bulk of data; may be errors or rare events.
• Boxplot/IQR: Outliers < Q1–1.5×IQR or > Q3+1.5×IQR.
• Z-Score: Outliers when |Z| > 3.
• Histogram: Reveals gaps in distribution.
• Scatterplot: Finds unusual relationships.
• One-Shot: Isolation Forest, One-Class SVM, LOF.

Worked Example:

Dataset of scores {65,70,75,78,100,0} → 0 and 100 are outliers.

Real-World Example:

Fraud detection: $50,000 purchase flagged among typical <$200 purchases.

Supplemental Resources:

• Book: An Introduction to Statistical Learning – James et al.
• Book: Practical Statistics for Data Scientists – Bruce & Gedeck.
• YouTube: StatQuest – Outliers and Influence
• Article: https://towardsdatascience.com/outlier-detection-methods
• Website: https://scikit-learn.org/stable/modules/outlier_detection.html

2. Feature Scaling

Concepts & Explanations:

• Ensures features contribute equally in distance/gradient models.
• Standardization: Mean=0, variance=1.
• Min-Max Scaling: Rescales to [0,1].
• Robust Scaling: Uses median & IQR; resistant to outliers.

Worked Example:

Age (0–100) vs Income (10k–100k). Without scaling, income dominates.

Real-World Example:

Recommender systems require scaling for meaningful similarity.

Supplemental Resources:

• Book: Hands-On Machine Learning – Aurélien Géron.

• YouTube: Krish Naik – Feature Scaling in Machine Learning
• Article: https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/
• Website: https://scikit-learn.org/stable/modules/preprocessing.html

3. Feature Engineering

Concepts & Explanations:

• Transform raw data into useful features.
• Transformations: Log, sqrt, Box-Cox.
• Encoding: OHE, Label, Target encoding.
• Feature Creation: Ratios, interaction terms, domain features.
• Binning: Converts continuous → categories.
• Selection: Remove redundancy via correlation, VIF, embedded models.
• Dimensionality Reduction: PCA, autoencoders.
• Temporal: Lag values, rolling averages, seasonality.

Worked Example:

Create income-to-expense ratio as new feature.

Real-World Example:

Retail: Days since last purchase is more predictive than raw date.

Supplemental Resources:

• Book: Feature Engineering for Machine Learning – Zheng & Casari.
• Book: Applied Predictive Modeling – Kuhn & Johnson.
• YouTube: Data School – Feature Engineering in Machine Learning
• Website: https://www.kaggle.com/learn/feature-engineering
• Article: https://towardsdatascience.com/the-art-of-feature-engineering

4. Encoding Categorical Variables

Concepts & Explanations:

• OHE: Expands categories → binary columns.
• Label Encoding: Assigns integers, good for ordinals.
• Target Encoding: Replace category with mean target.

Worked Example:

Colors {Red,Blue,Green} → OHE: [1,0,0],[0,1,0],[0,0,1].

Real-World Example:

Airline data with 100 airports: OHE creates 100 columns; target encoding more efficient.

Supplemental Resources:

• Book: Python Machine Learning – Sebastian Raschka.
• YouTube: Krish Naik – Encoding Categorical Variables
• Article: https://www.analyticsvidhya.com/blog/2020/08/types-of-categorical-data-encoding/
• Website: https://scikit-learn.org/stable/modules/preprocessing.html#encoding-categorical-features

5. Embedded Methods

Concepts & Explanations:

• Regularization: L1 shrinks coefficients to zero, L2 reduces variance.
• Tree-Based Models: Robust to scaling and outliers.
• Robust Loss: Huber, Quantile reduce extreme error effects.
• Anomaly-Aware Models: Isolation Forest, One-Class SVM.

Worked Example:

Lasso shrinks irrelevant coefficients to zero.

Real-World Example:

Healthcare: Tree models identify key symptoms while ignoring irrelevant features.

Supplemental Resources:

• Book: Elements of Statistical Learning – Hastie, Tibshirani, Friedman.
• YouTube: StatQuest – Regularization
• Article: https://medium.com/@jonathansharman/lasso-vs-ridge-regression-explained
• Website: https://scikit-learn.org/stable/modules/linear_model.html

6. Variance Inflation Factor (VIF)

Concepts & Explanations:

• Measures multicollinearity between features.
• Formula: VIF = 1 / (1 – R^2).
• VIF > 5–10 indicates problematic multicollinearity.

Worked Example:

Height in cm and inches → redundant, high VIF.

Real-World Example:

Marketing: Ad spend and impressions often highly correlated.

Supplemental Resources:

• Book: Applied Linear Regression – Kutner.
• YouTube: StatQuest – Multicollinearity and VIF

• Article: https://medium.com/analytics-vidhya/variance-inflation-factor-vif-93b3c9a9e6b0

7. Overfitting

Concepts & Explanations:

• Model fits noise instead of signal.
• Symptoms: low training error, high test error.
• Causes: too many features, outliers, high VIF, overly complex models.
• Prevention: regularization, cross-validation, simpler models.

Worked Example:

Tree fits training data perfectly but fails on test set.

Real-World Example:

Stock prediction models often overfit historical data.

Supplemental Resources:

• Book: An Introduction to Statistical Learning – James et al.
• Book: Hands-On Machine Learning – Aurélien Géron.
• YouTube: StatQuest – Overfitting and Underfitting
• Article:
https://towardsdatascience.com/overfitting-in-machine-learning-and-how-to-avoid-it-7f0e3d10a1f0

8. Bias–Variance Tradeoff

Concepts & Explanations:

• Bias: Error from oversimplification (underfitting).
• Variance: Error from sensitivity to fluctuations (overfitting).
• Good models balance bias and variance.

Worked Example:

Linear regression underfits (bias), deep tree overfits (variance).

Real-World Example:

Healthcare: Simple model misses diagnoses, complex model memorizes rare cases.

Supplemental Resources:

• Book: Elements of Statistical Learning – Hastie et al.
• YouTube: StatQuest – Bias and Variance
• Article: https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229

9. PCA (Dimensionality Reduction)

Concepts & Explanations:

• PCA reduces dimensionality by projecting onto fewer components.
• Captures maximum variance while removing noise and multicollinearity.

Worked Example:

100 features → 2 PCs explaining 95% variance.

Real-World Example:

Facial recognition reduces thousands of pixels into key PCA features.

Supplemental Resources:

• Book: Pattern Recognition and Machine Learning – Bishop.
• YouTube: StatQuest – PCA Clearly Explained
• Article: https://towardsdatascience.com/principal-component-analysis-explained-832e3f2c09ab

10. Model Evaluation Metrics

Concepts & Explanations:

• Accuracy: Proportion of correct predictions.
• Precision: TP / (TP+FP).
• Recall: TP / (TP+FN).
• F1 Score: Harmonic mean of precision & recall.
• ROC-AUC: Probability positive ranked higher than negative.

Worked Example:

Fraud detection: 95% accuracy but low recall (frauds missed).

Real-World Example:

Spam filters: Precision matters to avoid flagging real emails.

Supplemental Resources:

• Book: Introduction to Information Retrieval – Manning et al.
• YouTube: StatQuest – Precision, Recall, and F1
• Article:
https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

11. Cross-Validation

Concepts & Explanations:

• Cross-validation tests generalization across multiple splits.
• k-fold CV: Data split into k parts; train on k-1, test on 1, repeat k times.

Worked Example:

5-fold CV: Dataset split into 5 chunks, each tested once.

Real-World Example:

Credit scoring uses CV to ensure fairness across customer groups.

Supplemental Resources:

• Book: An Introduction to Statistical Learning – CV chapter.
• YouTube: StatQuest – Cross-Validation
• Article: https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f