LECTURE: MODEL EVALUATION TECHNIQUES IN MACHINE LEARNING
I. INTRODUCTION TO MODEL EVALUATION
The foundation of model evaluation lies in understanding the confusion matrix, which organizes predictions into four categories:

Confusion Matrix:	**Predicted Positive**	**Predicted Negative**
Actual Positive	True Positive	False Negative
Actual Negative	False Positive	True Negative

Where: -
TP (True Positives): Correctly identified positive cases
TN (True Negatives): Correctly identified negative cases
FP (False Positives): Incorrectly identified as positive (Type I error)
FN (False Negatives): Incorrectly identified as negative (Type II error)

II. FUNDAMENTAL EVALUATION METRICS
A. Accuracy
Accuracy = (TP + TN)/(TP + TN + FP + FN)
- Represents overall correctness
Limitations: Misleading for imbalanced datasets, Doesn’t distinguish between error types
B. Precision
Precision = TP/(TP + FP)
- Measures exactness - Critical when false positives are costly -
Applications: Spam detection, Medical diagnosis, Recommendation systems
C. Recall (Sensitivity)
Recall = TP/(TP + FN)
- Measures completeness - Important when false negatives are costly
Applications: Disease detection, Fraud detection, Security systems
D. F1-Score
F1 = 2 × (Precision × Recall)/(Precision + Recall)
- Harmonic mean of precision and recall - Balanced measure for imbalanced datasets -
Properties: Ranges from 0 (worst) to 1 (best), Penalizes extreme imbalances between precision and recall
III. ROC CURVE ANALYSIS
A. ROC Curve Construction
Plot TPR vs FPR at various thresholds
TPR = TP/(TP + FN) # Sensitivity
FPR = FP/(FP + TN) # 1 - Specificity
Characteristics:
(0,0): Most conservative classifier
(1,1): Most liberal classifier
(0,1): Perfect classifierDiagonal line: Random classifier
B. Area Under the Curve (AUC)
AUC = ∫ TPR d(FPR)
- Interpretation: * AUC = 1.0: Perfect classifier * AUC = 0.5: Random classifier * AUC < 0.5: Worse than random

IV. PRACTICAL IMPLEMENTATION

A. Code Example (Python)

```
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
from sklearn.metrics import confusion_matrix, roc_curve, auc
# Basic metrics
accuracy = accuracy_score(y_true, y_pred)
precision = precision_score(y_true, y_pred)
recall = recall_score(y_true, y_pred)
f1 = f1_score(y_true, y_pred)
# Confusion matrix
conf_matrix = confusion_matrix(y_true, y_pred)
# ROC curve
fpr, tpr, thresholds = roc_curve(y_true, y_pred_proba)
roc_auc = auc(fpr, tpr)
```

B. Metric Selection Guidelines (Why/When use this metrices?)

Use Accuracy when:
- Classes are balanced
- False positives and negatives have similar cost

Use Precision when:
- False positives are more costly
- Resources for positive predictions are limited

Use Recall when:
- False negatives are more costly
- Missing positive cases is critical

Use F1-score when:
- Need balance between precision and recall
- Dataset is imbalanced

Use ROC-AUC when:
- Need threshold-independent evaluation
- Comparing different models

V. COMMON PITFALLS AND CONSIDERATIONS

Class Imbalance:
- Can skew accuracy
- May require specialized metrics
- Consider using weighted variants

Threshold Selection:
- Affects all metrics
- ROC curve helps in threshold selectionConsider business requirements

Data Leakage:
- Evaluate on unseen data
- Use proper cross-validation
- Maintain test set integrity

VI. PRACTICE PROBLEMS

Calculate all metrics for given confusion matrix:
TP = 85, FP = 15
FN = 10, TN = 90
Analyze ROC curves:
Compare two models' ROC curves
Determine optimal threshold
Calculate AUC
Case Study:
Medical diagnosis scenario
Imbalanced classes
Cost-sensitive evaluation

REFERENCES

Hastie, T., et al. “The Elements of Statistical Learning”
Géron, A. “Hands-On Machine Learning with Scikit-Learn”
James, G., et al. “An Introduction to Statistical Learning”