

**MA477: Data Science**  
**Lesson 10 Outline — 02 February 2026**  
United States Military Academy, West Point  
Instructor: MAJ Patrick Kuiper

---

## 1 Administrative

- Student review
- Classification I Questions
- Classification I Lecture

## 2 Shrinkage Lesson Objectives

- Understand common performance measures used in classification to include accuracy, precision, recall, and  $F_1$  score.
- Use a receiver operating characteristic (ROC) curve to assess classifiers.
- Understand the difference between multiclass and multioutput classification.

## 3 Student Review

## 4 Shrinkage Methods Lecture

### Shrinkage Methods (Ridge Regression and the Lasso)

## 5 Motivating Classification with Logistic Regression

### Opening Discussion Questions

1. Why might accuracy be a misleading measure of performance in a classification problem?
2. What does it mean to say a model outputs a probability rather than a label, and why might that be useful?
3. How should we decide what “good” predictions mean when different types of errors have different consequences?
4. Give an example of a scenario where high precision is appropriate.
5. Give an example of a scenario where high recall is appropriate.
6. Why do we need an objective function to train a classifier, and how should it relate to probability?
7. How are the quantities we optimize during training connected to the metrics we use to evaluate a model afterward?

### Basic Notation for Classification

Consider a supervised classification problem with input–output pairs

$$\{(x_i, y_i)\}_{i=1}^n,$$

where  $x_i \in \mathbb{R}^p$  denotes a vector of  $p$  features and  $y_i \in \{0, 1\}$  denotes the class label for the  $i$ th observation.

A classifier is a function

$$f : \mathbb{R}^p \rightarrow \{0, 1\}$$

that assigns a class label to each input. In probabilistic classification, the model estimates the conditional probability

$$P(Y = 1 | X = x).$$

In logistic regression, this probability is modeled as

$$P(Y = 1 | X = x) = \sigma(z), \quad \text{where } z = w^\top x + b,$$

with coefficient vector  $w \in \mathbb{R}^p$ , intercept  $b \in \mathbb{R}$ , and sigmoid function

$$\sigma(z) = \frac{1}{1 + e^{-z}}.$$

Predicted class labels are obtained by applying a decision threshold  $\tau \in (0, 1)$ :

$$\hat{y}_i = \begin{cases} 1 & \text{if } P(Y = 1 | x_i) \geq \tau, \\ 0 & \text{otherwise.} \end{cases}$$

The collection of predicted labels  $\{\hat{y}_i\}_{i=1}^n$  is compared to the true labels  $\{y_i\}_{i=1}^n$  to compute evaluation quantities such as true positives, false positives, true negatives, and false negatives.

## 5.1 Why Accuracy Is Often Insufficient

### Confusion Matrix

For a binary classification problem, the confusion matrix summarizes the relationship between true class labels and predicted class labels. It is given by

	Predicted 0	Predicted 1
Actual 0	$TN$	$FP$
Actual 1	$FN$	$TP$

where  $TP$  denotes true positives,  $TN$  denotes true negatives,  $FP$  denotes false positives, and  $FN$  denotes false negatives. The confusion matrix forms the basis for commonly used classification metrics such as accuracy, precision, recall, and the F1 score.

In binary classification, accuracy is defined as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}.$$

Accuracy measures overall correctness, but it treats all errors equally. In practice, the two types of classification errors often have very different consequences.

- **False positives (FP):** predicting a positive outcome when none exists.
  - Medical screening: unnecessary follow-up tests.
  - Fraud detection: legitimate transactions incorrectly flagged.
  - Security systems: false alarms.
- **False negatives (FN):** failing to detect a true positive.
  - Medical diagnosis: missed disease.
  - Fraud detection: undetected fraud.
  - Safety or surveillance systems: missed threats.

Because accuracy cannot reflect these asymmetric costs, alternative metrics that focus on specific error types are often required.

## 5.2 Logistic Regression as a Probabilistic Classifier

Logistic regression models the conditional probability of the positive class as

$$P(Y = 1 | X = x) = \sigma(z), \quad \text{where } z = w^\top x + b.$$

Here,

- $w \in \mathbb{R}^p$  is the vector of coefficients,
- $x \in \mathbb{R}^p$  is the feature vector,
- $b \in \mathbb{R}$  is the intercept,
- $z$  is the linear predictor (or logit).

The sigmoid function is defined as

$$\sigma(z) = \frac{1}{1 + e^{-z}}.$$

Predicted class labels are obtained by thresholding the probability:

$$\hat{y} = \begin{cases} 1 & \text{if } \sigma(z) \geq \tau, \\ 0 & \text{otherwise.} \end{cases}$$

Logistic regression assumes a linear relationship between the predictors and the log-odds:

$$\log \left( \frac{P(Y = 1 | x)}{1 - P(Y = 1 | x)} \right) = z = w^\top x + b,$$

which implies a linear decision boundary defined by  $z = 0$ .

## 5.3 Precision, Recall, and Error Interpretation

From the confusion matrix, we define the following metrics.

### Precision

$$\text{Precision} = \frac{TP}{TP + FP}.$$

Precision measures the proportion of predicted positives that are correct and is sensitive to false positives.

### Recall

$$\text{Recall} = \frac{TP}{TP + FN}.$$

Recall measures the proportion of actual positives that are correctly identified and is sensitive to false negatives.

By adjusting the decision threshold  $\tau$ , one can trade precision against recall, reflecting different application priorities.

## Receiver Operating Characteristic (ROC) Curve

The Receiver Operating Characteristic (ROC) curve evaluates the performance of a binary classifier across all possible decision thresholds. It is a plot of the true positive rate (TPR) against the false positive rate (FPR), where

$$\text{TPR} = \frac{TP}{TP + FN}, \quad \text{FPR} = \frac{FP}{FP + TN}.$$

Each point on the ROC curve corresponds to a specific classification threshold applied to the model's predicted probabilities. The area under the ROC curve (AUC) provides a threshold-independent measure of a model's ability to distinguish between the two classes.

## 5.4 The F1 Score

The F1 score is the harmonic mean of precision and recall:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

The harmonic mean penalizes extreme imbalance between precision and recall, making the F1 score useful when both types of errors matter.

## 5.5 Maximum Likelihood Estimation

Logistic regression parameters are estimated using maximum likelihood. Assume

$$Y_i | X_i \sim \text{Bernoulli}(p_i), \quad p_i = \sigma(z_i), \quad z_i = w^\top x_i + b.$$

The likelihood function is

$$L(w, b) = \prod_{i=1}^n \sigma(z_i)^{y_i} (1 - \sigma(z_i))^{1-y_i}.$$

Taking logs yields the log-likelihood

$$\ell(w, b) = \sum_{i=1}^n [y_i \log \sigma(z_i) + (1 - y_i) \log (1 - \sigma(z_i))].$$

Maximizing this expression is equivalent to minimizing the negative log-likelihood, also known as the binary cross-entropy loss.

The gradient with respect to  $w$  is

$$\nabla_w \ell = \sum_{i=1}^n (y_i - \sigma(z_i)) x_i,$$

showing that parameter updates are driven by differences between observed labels and predicted probabilities.

## 5.6 Connecting Training and Evaluation

Maximum likelihood estimation focuses on learning well-calibrated probabilities through the linear predictor  $z = w^\top x + b$ . Precision, recall, and the F1 score evaluate how these probabilities are converted into decisions via thresholding.

Thus, the training objective and evaluation metrics serve distinct but complementary roles in classification.