MA477: Data Science

Lesson 11 Outline — 04 February 2026

United States Military Academy, West Point
Instructor: MAJ Patrick Kuiper

---

1 Administrative

- Student review
- Classification II Questions
- Classification II Lecture

2 Classification 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 Motivating Classification with Logistic Regression

Opening Discussion Questions

1. Give an example of a scenario where high precision is appropriate.
2. Give an example of a scenario where high recall is appropriate.
3. Describe an example of a multiclass problem?
4. Describe an example of a multilabel problem?
5. Why do we need an objective function to train a classifier, and how should it relate to probability / threshold?
6. How are the quantities we optimize during training connected to the metrics we use to evaluate a model afterward?

4.1 Logistic Regression as a Probabilistic Classifier

Logistic regression models the conditional probability of the positive class as

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

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:

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

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

4.2 Understanding Error - 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 τ, 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

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

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.

4.3 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.

Multiclass, Multilabel, and Multioutput Classification

Following the classification framework described by :contentReference[oaicite:0]index=0, we distinguish between the following settings.

Multiclass Classification Each observation belongs to exactly one of $K > 2$ classes:

$$Y \in \{1, 2, \ldots, K\}.$$

The classifier predicts a single class label

$$\hat{y} = \arg\max_k P(Y = k \mid X = x).$$

Multilabel Classification Each observation may belong to multiple classes simultaneously:

$$Y = (Y_1, \ldots, Y_K), \quad Y_k \in \{0, 1\}.$$

Each label is predicted independently:

$$\hat{Y}_k \in \{0, 1\}.$$

Multioutput Classification The response consists of multiple outputs, possibly with different label spaces:

$$Y = \left(Y^{(1)}, \ldots, Y^{(m)}\right), \quad Y^{(j)} \in \mathcal{Y}^{(j)}.$$

Each output is predicted jointly by a single model.

4.4 Maximum Likelihood Estimation

Logistic regression parameters are estimated using maximum likelihood. Assume

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

The likelihood function is

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

Taking logs yields the log-likelihood

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

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.

4.5 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.