MA477: Data Science

Lesson 12 Outline — 06 February 2026

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

---

1 Administrative

- Student review
- Logistic Regression Questions
- Logistic Regression Lecture

2 Logistic Regression Lesson Objectives

- Understand why linear regression is not appropriate for a qualitative response.
- Interpret the coefficients of a multiple logistic regression model and obtain predictions for a set of inputs.
- Use functions in the sklearn module to assess logistic regression models.

3 Student Review

Logistic Regression: Core Concepts and Discussion

1. **Describe some of the issues with using a linear regression type model for classification problems - what portion of logistic regression is linear?**

   Linear regression does not have an expressly probabilistic measure associated with discrete outcomes.?

   The logit is the linear predictor

   $$z = w^\top x + b,$$

   where $w \in \mathbb{R}^p$ and $b \in \mathbb{R}$. It aggregates feature information into an unconstrained real-valued score that serves as the input to the nonlinear transformation.

2. **What function is applied to the linear transformation of the observed features, and how does this function operate?**

   Linear regression does not have an expressly probabilistic measure associated with discrete outcomes.

   The sigmoid function

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

   maps the real-valued logit $z \in \mathbb{R}$ to a value in $(0, 1)$, enabling probabilistic interpretation while preserving score ordering and differentiability.

3. **What probability distribution do we use to model do we use to define a error function with Logisitc Regression?**

   Logistic regression assumes a Bernoulli response:

   $$Y \mid X = x \sim \text{Bernoulli}(p), \quad p = \sigma(z).$$

   Thus,

   $$P(Y = 1 \mid X = x) = p, \qquad P(Y = 0 \mid X = x) = 1 - p,$$

   which satisfies non-negativity, normalization, and completeness of probabilities.

   Where the Bernoulli PMF is defined as follows:

   $$P(Y = y) = p^{\,y}(1 - p)^{\,1-y}, \qquad y \in \{0, 1\},\ p \in [0, 1].$$

4. **What is the loss function used for Logistic Regression?** In logistic regression, each response Y_i is modeled as a Bernoulli random variable with success probability

$$p_i = \sigma(z_i), \qquad z_i = w^\top x_i + b,$$

where $\sigma(\cdot)$ denotes the sigmoid function. The joint probability mass function of the data is therefore

$$P(y_1, \ldots, y_n \mid w, b) = \prod_{i=1}^{n} \sigma(z_i)^{y_i} \big(1 - \sigma(z_i)\big)^{1-y_i}, \qquad y_i \in \{0, 1\}, \tag{1}$$

$$\ell(w, b) = \log P(y_1, \ldots, y_n \mid w, b) = \sum_{i=1}^{n} \left[y_i \log \sigma(z_i) + (1 - y_i) \log \big(1 - \sigma(z_i)\big) \right]. \tag{2}$$

The negative log-likelihood of the Bernoulli model yields the cross-entropy loss:

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

Minimizing this loss is equivalent to maximum likelihood estimation and strongly penalizes confident incorrect predictions.

5. **How do we minimize this loss?**

The gradient of the log-likelihood with respect to the weight vector $w \in \mathbb{R}^p$ is obtained by applying the chain rule to the linear predictor $z_i = w^\top x_i + b$:

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

Here, $(y_i - \sigma(z_i))$ represents the prediction error for observation i, and x_i scales this error by the corresponding feature values. The resulting gradient points in the direction that increases the log-likelihood by reducing the discrepancy between observed labels and predicted probabilities.

Gradient ascent updates the parameters according to

$$w^{(t+1)} = w^{(t)} + \eta \sum_{i=1}^{n} (y_i - \sigma(z_i)) x_i,$$

where $\eta > 0$ is the learning rate. Equivalently, minimizing the negative log-likelihood via gradient descent yields the same update with a negative sign. This update rule produces stable, interpretable learning dynamics due to the convexity of the logistic loss.

Multiclass Classification Strategies

6. **What do we do when we what to consider multiple classes?**

$$\mathrm{softmax}(z)_k = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}}, \qquad k = 1, \ldots, K.$$

   (a) **Softmax (Multinomial Logistic Regression).** Trains a single model that outputs a normalized probability distribution over all classes. *Strengths:* statistically principled, enforces probability normalization, enables direct class competition, and produces well-calibrated probabilities.

   (b) **One-vs-All (One-vs-Rest).** Trains one binary classifier per class against all remaining classes. *Strengths:* simple to implement, flexible, compatible with any binary classifier, and easy to interpret per class.

   (c) **One-vs-One.** Trains a binary classifier for every pair of classes and combines predictions via voting or aggregation. *Strengths:* effective for small numbers of classes, handles class overlap well, often performs strongly with margin-based models.

Summary

Logistic regression combines a linear logit, a sigmoid transformation, a Bernoulli probability model, cross-entropy loss derived from maximum likelihood, and gradient-based optimization into a coherent probabilistic classification framework.