STA 35C: Statistical Data Science III

Lecture 9: Logistic Regression (cont'd) & Classification Errors

Dogyoon Song

Spring 2026, UC Davis

Announcements

Homework 3 is posted (Due: Tue, Apr 21, 11:59 PM)

- Please submit on time and follow the submission instructions
- You can collaborate on Homework, but
  - All submitted work must be your own, and
  - You must clearly list the names of all students you discussed with

Midterm 1 is in class on Fri, Apr 24

- You may bring *one **hand-written** letter-sized (8.5 $\times$ 11 inches), double-sided sheet of paper* with formulas and brief notes
- **Calculator**: Simple (non-graphing) calculators only
- **No textbooks** or other materials beyond the single cheat sheet
- **SDC accommodations:** Confirm scheduling with AES online

Agenda

Last time:

- Classification: what it is and how it differs from regression
- Logistic regression for binary responses

Today:

- Extensions of logistic regression
  - More than one predictor
  - More than two classes
- Assessing classification performance
  - Error rate
  - Confusion matrix
  - ROC curve

Recap: Classification

Classification problem: Given labeled data (x_i, y_i), where $x_i \in \mathbb{R}^p$ and $y_i \in \{1, \ldots, K\}$, learn a map that assigns a new point x_{new} to one of the K classes

- Email text (X) $\rightarrow$ Spam or not (Y)
- Test score (X) $\rightarrow$ Pass or fail (Y)
- Diagnostic results $\rightarrow$ Disease or not (Y)

Typical procedure for binary classification methods: To predict the class label for x,

1. Estimate class probabilities:

$$\hat{p}_1(x) \approx \Pr(Y = 1 \mid X = x), \qquad \hat{p}_0(x) = 1 - \hat{p}_1(x)$$

2. Predict

$$\hat{y} = 1 \qquad \text{if} \qquad \hat{p}_1(x) \geq p^*$$

   - We can adjust this decision rule more aggressively or more conservatively

Recap: Simple logistic regression ($p = 1, K = 2$)

Logistic regression model:

$$\Pr(Y = 1 \mid X = x) = \sigma(\beta_0 + \beta_1 x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

Where did it come from?

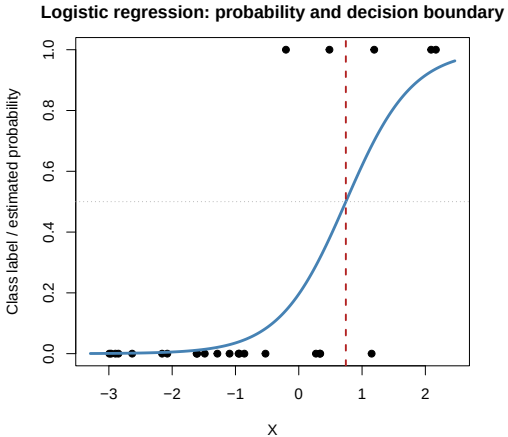
- To model $p(x) = \Pr(Y = 1 \mid X = x) \in [0, 1]$ using a quantity that varies linearly with x
- We need a monotone map $g : (0, 1) \to \mathbb{R}$ so that $g(p(x))$ can be modeled linearly in x
- We model/assume the *log-odds (logit)* is linear in X:

$$\log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x$$

Interpreting coefficients:

- β_0: log-odds at $x = 0$
- β_1: a 1-unit increase in x multiplies the *odds* by e^{β_1}

Recap: Logistic regression illustration



Making predictions: Once we have $\hat{\beta}_0, \hat{\beta}_1$,

- $\hat{p}(x) = \sigma(\hat{\beta}_0 + \hat{\beta}_1 x)$
- Predict $\hat{Y} = 1$ iff $\hat{p}(x) \geq p^*$
- Threshold $p^* \in [0, 1]$

Maximum likelihood estimation (MLE):

- Given data (x_i, y_i) with $y_i \in \{0, 1\}$, let
  $p_i = \Pr(Y_i = 1 \mid X_i = x_i) = \sigma(\beta_0 + \beta_1 x_i)$
- The likelihood function of (β_0, β_1) is
  $$L(\beta_0, \beta_1) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{(1-y_i)}$$
- Choose $\hat{\beta}_0, \hat{\beta}_1$ that maximizes $L(\beta_0, \beta_1)$

Multiple logistic regression ($p > 1$)

Model: Linear model for log-odds with more predictors

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p, \qquad p(x) = \Pr(Y = 1 \mid X = x)$$

- The log-odds (=logit) is linear in $X_1, \ldots, X_p$

Coefficients:

- Each β_j measures the change in log-odds for a one-unit increase in x_j, holding the others
- A 1-unit increase in X_i multiplies the odds by e^{β_i} when other predictors are controlled
- The coefficients are estimated by maximum likelihood, just as in simple logistic regression

Prediction rule: Once we estimate $p(x) = \Pr(Y = 1 \mid X = x)$, classify via

$$\hat{Y} = \begin{cases} 1, & \text{if } \hat{p}(x) \geq p^*, \\ 0, & \text{otherwise.} \end{cases}$$

Decision boundary

Question: For which predictor values x do we predict $\hat{Y} = 1$?

Decision boundary: For a logistic regression model, the class-1 region is determined by

$$p(X) \geq p^* \iff \log\left(\frac{p(x)}{1-p(x)}\right) \geq \log\left(\frac{p^*}{1-p^*}\right)$$
$$\iff \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \geq \log\left(\frac{p^*}{1-p^*}\right)$$

- The decision boundary is the hyperplane $\left\{\vec{x} \in \mathbb{R}^p \mid \beta_0 + \sum_{i=1}^{p} \beta_i x_i = \log\left(\frac{p^*}{1-p^*}\right)\right\}$

In 2-dimensional case ($p = 2$)

- Rearranging the terms gives the equation of a line in $\mathbb{R}^2$:

$$\beta_2 x_2 = -\beta_0 - \beta_1 x_1 + \log\left(\frac{p^*}{1-p^*}\right) \overset{\text{if } \beta_2 \neq 0}{\Longrightarrow} x_2 = -\frac{\beta_1}{\beta_2} x_1 - \frac{\beta_0}{\beta_2} + \frac{1}{\beta_2} \log\left(\frac{p^*}{1-p^*}\right)$$

- The decision boundary is a line, and we predict $\hat{Y} = 1$ on one side of that line (a half-space)
  - If $\beta_2 > 0$, this is the region above the line; if $\beta_2 < 0$, it is below the line

Decision boundary: Illustration

Example

Suppose the fitted model is

$$\log\left(\frac{p(x)}{1-p(x)}\right) = -5 + 3x_1 + 2x_2.$$

For $p^* = 0.5$, $\log(\frac{p^*}{1-p^*}) = 0$, so

$$\hat{y} = 1 \qquad \Longleftrightarrow \qquad -5 + 3x_1 + 2x_2 \geq 0 \qquad \Longleftrightarrow \qquad x_2 \geq 2.5 - 1.5x_1.$$

Thus, the decision boundary is the line

$$x_2 = 2.5 - 1.5x_1,$$

and we predict class 1 for points above this line.

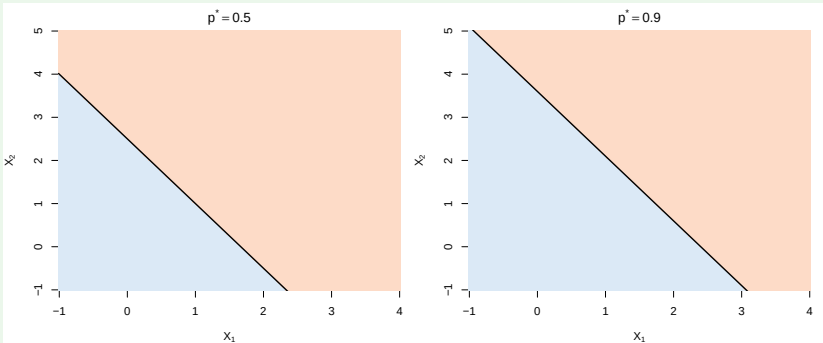
Decision boundary: Illustration

Example

If we increase $p^* = 0.5$ to $p^* = 0.9$, then $\log\left(\frac{0.9}{0.1}\right) = \log 9 \approx 2.20$, and we predict $\hat{y} = 1$ if and only if

$$-5 + 3x_1 + 2x_2 \geq \log 9 \qquad \Longleftrightarrow \qquad x_2 \geq 3.60 - 1.5x_1.$$

Raising the threshold shifts the boundary upward in parallel and makes the class-1 region smaller.



Pop-up quiz: Logistic regression boundary in 2D

Suppose a fitted logistic model is

$$\log\left(\frac{\hat{p}(x)}{1-\hat{p}(x)}\right) = -2 + x_1 + x_2.$$

We predict class 1 when $\hat{p}(x) \geq 0.8$. Note that

$$\log\left(\frac{0.8}{0.2}\right) = \log 4.$$

Question: Which statement is correct?

A) The boundary is $x_1 + x_2 = 2$, and class 1 is predicted when $x_1 + x_2 \geq 2$.
B) The boundary is $x_1 + x_2 = 2 + \log 4$, and class 1 is predicted when $x_1 + x_2 \geq 2 + \log 4$.
C) The boundary is curved because the logistic function is nonlinear.
D) The model predicts class 1 only when both x_1 and x_2 are positive.

Answer: B. The threshold 0.8 corresponds to log-odds $\log 4$, so the boundary satisfies $-2 + x_1 + x_2 = \log 4$.

Multinomial logistic regression ($K > 2$)

Question: What if the response has more than two classes?

Idea: Choose one class as a **baseline**, say class K, and model the log-odds of each other class relative to that baseline:

$$\log\left(\frac{p_k(x)}{p_K(x)}\right) = \beta_{k0} + \beta_k^\top x, \qquad k = 1, \ldots, K-1.$$

Example: $K = 3$, $p = 1$, baseline class 3

$$\log\left(\frac{p_1(x)}{p_3(x)}\right) = \beta_{10} + \beta_{11}x, \qquad \log\left(\frac{p_2(x)}{p_3(x)}\right) = \beta_{20} + \beta_{21}x.$$

Then

$$p_k(x) = \frac{e^{\beta_{k0}+\beta_{k1}x}}{1 + e^{\beta_{10}+\beta_{11}x} + e^{\beta_{20}+\beta_{21}x}}, \quad k \in \{1, 2\}, \qquad p_3(x) = \frac{1}{1 + e^{\beta_{10}+\beta_{11}x} + e^{\beta_{20}+\beta_{21}x}}.$$

Predict the class k with the largest $p_k(x)$: $\hat{y}(x) = \arg\max_{k \in \{1,\ldots,K\}} \hat{p}_k(x)$

- Changing the baseline class changes the coefficients, but not the fitted probabilities or predictions

Multinomial logistic regression: Illustration

Example

Suppose class 3 is the baseline and we fit

$$\log\left(\frac{p_1(x)}{p_3(x)}\right) = 2 - 2x, \qquad \log\left(\frac{p_2(x)}{p_3(x)}\right) = -6 + 2x.$$

Then

$$p_1(x) : p_2(x) : p_3(x) = e^{2-2x} : e^{-6+2x} : 1.$$

$$p_1(x) = \frac{e^{2-2x}}{1 + e^{2-2x} + e^{-6+2x}}, \qquad p_2(x) = \frac{e^{-6+2x}}{1 + e^{2-2x} + e^{-6+2x}}, \qquad p_3(x) = \frac{1}{1 + e^{2-2x} + e^{-6+2x}}.$$

The predicted class is the one with the largest probability:

$$\hat{Y}(x) = \begin{cases} 1, & x < 1, \\ 3, & 1 \le x \le 3, \\ 2, & x > 3. \end{cases}$$

At $x = 1$ and $x = 3$, there is a tie, which can be broken arbitrarily.

Multinomial logistic regression: Illustration (cont'd)

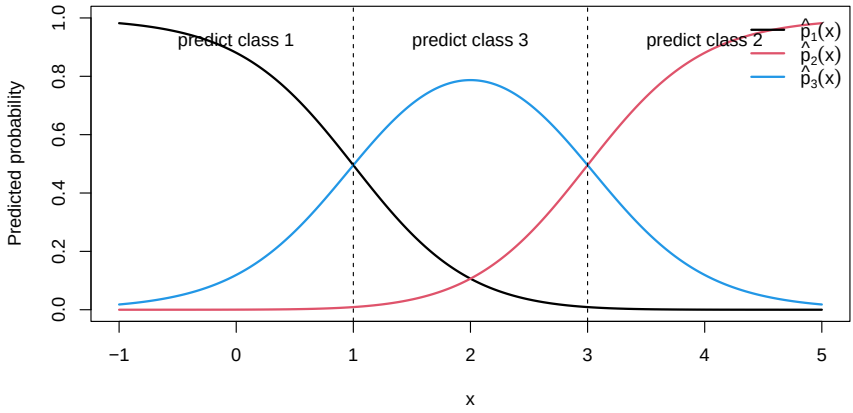


Figure: Predicted class probabilities under the multinomial logistic model $\log\left(\frac{p_1(x)}{p_3(x)}\right) = 2 - 2x$ and $\log\left(\frac{p_2(x)}{p_3(x)}\right) = -6 + 2x$. The predicted class changes at the vertical boundaries $x = 1$ and $x = 3$.

Error rate

Definition: Fraction of observations that are misclassified

$$\text{Error rate} = \frac{1}{n}\sum_{i=1}^{n} I(\hat{y}_i \neq y_i)$$

Bayes classifier:

$$x \mapsto \arg\max_k \Pr(Y = k \mid X = x)$$

- Minimizes the expected classification error rate
- Usually impossible to compute *in practice*, since $\Pr(Y \mid X)$ is unknown

Question: Even if we could compute Bayes classifier, is the error rate the best measure?

- Some classification errors could be costlier than others
- e.g., missing a cancer is worse than a false alarm

Confusion matrix: Binary classification

Let's consider **binary** classification ($Y = 0$ or 1)

		True default status		
		No	Yes	Total
Predicted default status	No	9432	138	9570
	Yes	235	195	430
	Total	9667	333	10000

Figure: An example confusion matrix for the `Default` dataset [JWHT21, Table 4.5].

Four possible outcomes:

- True positive (TP): predicted $\hat{Y} = 1$ when $Y = 1$ is true
- False negative (FN): predicted $\hat{Y} = 0$ when $Y = 1$ is true
- False positive (FP): predicted $\hat{Y} = 1$ when $Y = 0$ is true
- True negative (TN): predicted $\hat{Y} = 0$ when $Y = 0$ is true

Minimizing total error rate can be suboptimal if FP and FN have different costs

More on error metrics

For binary classification, the confusion matrix gives several useful summaries:

$$\text{Error rate} = \frac{FP + FN}{TP + TN + FP + FN}, \qquad \text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},$$

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

- **Error rate:** Among all observations, how many are misclassified?
- **False positive rate (FPR):** Among actual negatives, how many are falsely claimed positive?
- **False negative rate (FNR):** Among actual positives, how many are missed?

		True class		
		− or Null	+ or Non-null	Total
Predicted class	− or Null	True Neg. (TN)	False Neg. (FN)	N*
	+ or Non-null	False Pos. (FP)	True Pos. (TP)	P*
	Total	N	P	

Figure: Summary of possible classification outcomes in a population [JWHT21, Table 4.6].

Threshold selection

Many classifiers (e.g. logistic regression) produce $\hat{p}(x) = \Pr(Y = 1 \mid x)$

- If $\hat{p}(x) \geq p^*$, predict $Y = 1$, else 0
- Increasing p^* lowers false positives but increases false negatives

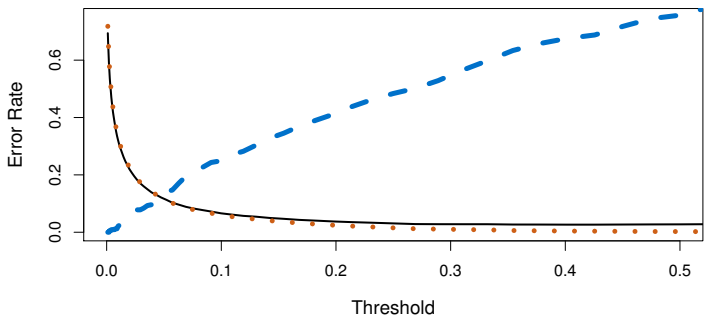


Figure: False positive (orange dotted) and false negative (blue dashed) error rates as a function of the threshold value p^* for the `Default` dataset [JWHT21, Figure 4.7].

Receiver operating characteristic (ROC) curve

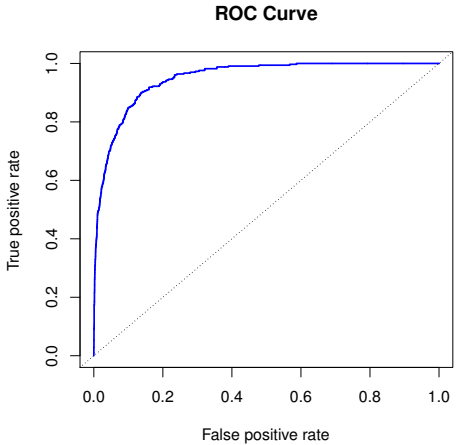


Figure: An example ROC curve, with AUC [JWHT21, Figure 4.8].

ROC curve

- Plot TPR vs. FPR as p^* moves $1 \to 0$
  - $\mathrm{TPR} = \frac{TP}{P} = \frac{TP}{TP+FN} = 1 - \mathrm{FNR}$
  - $\mathrm{FPR} = \frac{FP}{N} = \frac{FP}{TN+FP}$
- Curves closer to the upper-left corner indicate better classification performance

Area under curve (AUC)

- Reflects overall discriminative power across thresholds
  - Perfect classifier: AUC $= 1$
  - Random guess: AUC $= 0.5$

Pop-up quiz: Error metrics

A logistic regression classifier predicts class 1 when $\hat{p}(x) \geq p^*$. Suppose we lower the threshold from $p^* = 0.5$ to $p^* = 0.2$.

Question: What usually happens?

A) Fewer observations are predicted positive, so both TPR and FPR decrease.

B) More observations are predicted positive, so TPR decreases and FPR increases.

C) More observations are predicted positive, so TPR tends to increase and FPR also tends to increase.

D) TPR and FPR stay the same; only the predicted probabilities change.

Answer: C. Lowering the threshold makes it easier to predict class 1, so we usually catch more true positives but also create more false positives.

Wrap-up

- Logistic regression extends to multiple predictors by modeling the log-odds as a linear function of $x_1, \ldots, x_p$.
- A classification threshold turns estimated probabilities into class labels; changing the threshold changes the decision boundary; for logistic regression, the decision boundary is linear in the predictor space.
- Multinomial logistic regression handles $K > 2$ classes by comparing each class to a baseline and normalizing the resulting probabilities.
- Classification performance can be assessed using error rate: confusion matrices, ROC curves, and AUC give a more detailed picture.
- Changing the classification threshold changes the tradeoff between false positives and false negatives.

Next lecture: Generative models for classification (LDA, Naive Bayes)

References


Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani.
An Introduction to Statistical Learning: with Applications in R, volume 112 of *Springer Texts in Statistics*.
Springer, New York, NY, 2nd edition, 2021.