STA 35C: Statistical Data Science III

Lecture 8: Classification Basics& Logistic Regression

Dogyoon Song

Spring 2026, UC Davis

Announcements

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

- Please submit on time and follow the submission instructions
- Please review the homework problems early in case you might have questions
- Feel free to ask questions during lecture, in office hours or post them on Piazza

Midterm 1 is in class on Fri, Apr 24

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

I'll be holding my office hours today 3:30–4:30 PM at MSB 4220

- Hope to see and chat with many of you there

Agenda

So far:

- Problem: Regression
- Method: Linear regression

Today:

- Problem: Classification
  - What is classification?
  - How it differs from regression
- Method: Logistic regression
  - Basic ideas
  - Model formulation
  - Prediction with logistic regression
  - Parameter estimation

Classification: Motivation

Classification = Supervised learning to predict *qualitative* (categorical) responses

Examples:

- Email spam vs. non-spam
- Fraudulent transaction vs. legitimate
- Medical diagnosis (multiple possible conditions)
- Handwritten digit classification (0–9)

Key difference from regression:

- Y is a *class label*, not a numeric value
- We often interpret output of classification methods as the *probability* of a class
- Accuracy metrics differ (e.g., classification error, confusion matrix)

Classification: A visual illustration

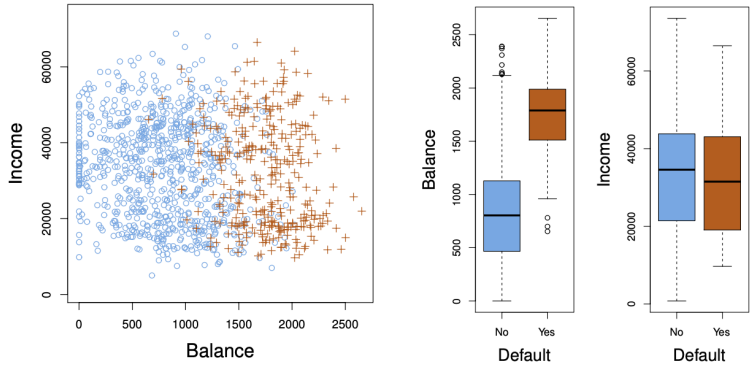


Figure: The `Default` dataset: annual incomes vs. monthly credit card balances. Orange: individuals who defaulted, Blue: those who did not [JWHT21, Figure 4.1].

Goal: Learn a decision rule (or decision boundary) that assigns a new point x_{new} to a class

Classification setting: Formal description

Goal: Given labeled data $\{(x_i, y_i)\}_{i=1}^n$, learn a rule that assigns x_{new} to one of K classes

- Often we first estimate class probabilities

$$p_k(x) = \Pr(Y = k \mid X = x), \qquad k = 1, \ldots, K,$$

  and then predict the class with the largest estimated probability

Example

- Email text $(X) \rightarrow$ spam or not (Y)
- Handwritten image $(X) \rightarrow$ digit (Y)
- Patient measurements $(X) \rightarrow$ medical condition (Y)

Question: Wait... why not just use regression methods?

Why don't we simply use regression methods?

Naive attempt:

- Assign $Y \in \{0, 1\}$ or $\{1, \ldots, K\}$ numerically
- Fit a linear model $Y \approx \beta_0 + \beta_1 X$

Issues with the naive attempt:

- For $K > 2$, coding classes as $1, 2, \ldots, K$ imposes an artificial ordering and spacing
- Even for $K = 2$, linear regression can produce fitted values outside $[0, 1]$, so they cannot be interpreted as valid probabilities
- For binary data, $\mathrm{Var}(Y \mid X = x) = p(x)(1 - p(x))$ depends on x, so the constant-variance assumptions of linear regression are mismatched

We need a method that respects the *categorical* nature of Y and keeps predicted probabilities inside $[0, 1]$

Visual illustration: Linear vs. logistic regression

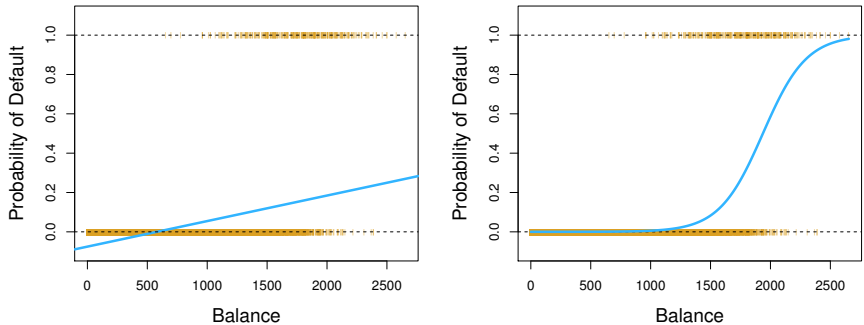


Figure: **Left:** Estimated "probability" of `default` using linear regression. **Right:** Probability estimated via logistic regression [JWHT21, Figure 4.2].

Therefore: Classification-specific methods are typically more appropriate and preferred

Pop-up quiz #1: Classification vs. regression

Suppose we code a binary response as $Y \in \{0, 1\}$ and fit a linear regression model. For some predictor value x, the fitted value is $\hat{y} = -0.15$.

Question: What is the best conclusion?

A) The estimated probability of class 1 is -15%.

B) Linear regression can produce fitted values outside $[0, 1]$, so it is not ideal for modeling class probabilities.

C) The model predicts class 0 with certainty.

D) The data must contain mislabeled observations.

Answer: B. A fitted value of -0.15 cannot be a valid probability, which is one reason we prefer classification-specific methods.

Roadmap

We will learn two types of classification methods

- **Logistic regression**: a *discriminative* approach that models $P(Y = 1|X)$
- **Generative models**: model $P(Y)$ and $P(X \mid Y)$, then apply Bayes' rule
  - Example: Linear discriminant analysis (LDA), Naive Bayes

Today: Logistic regression with one predictor X ($p = 1$) for binary ($K = 2$) classification

Logistic regression: Basic ideas

For binary classification ($Y \in \{0, 1\}$), let

$$p(x) = \Pr(Y = 1 \mid X = x)$$

- We want to model $p(x) \in [0, 1]$ using a function $f : x \mapsto p(x)$
- Then predict the class using an estimated probability $\hat{p}(x)$:

$$Y = \begin{cases} 1, & \text{if } \hat{p}(x) \geq p^* \quad (\text{often } p^* = 0.5), \\ 0, & \text{otherwise} \end{cases}$$

How do we get there from linear regression?

- A naive use of linear model for $Y \in \{0, 1\}$ can yield $\hat{y} \notin [0, 1]$
- The *odds*: $\frac{p(X)}{1-p(X)} \in [0, \infty)$
- The *log-odds* (also called logit): $\log\left(\frac{p(X)}{1-p(X)}\right) \in \mathbb{R}$

Logistic regression: Model formulation

Key idea: Model the log-odds as a linear function of x:

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

- Positive log-odds $\iff p(x) > 0.5$
- Negative log-odds $\iff p(x) < 0.5$

Question: How do we write $p(X)$ as a function of β_0, β_1, X?

- Solving for $p(x)$, we observe

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X \quad \iff \quad \frac{p(X)}{1-p(X)} = \exp(\beta_0 + \beta_1 X)$$

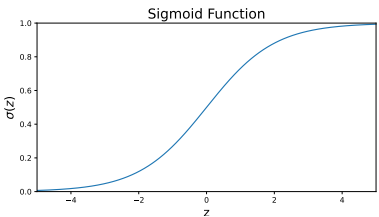
$$\iff \quad p(X) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}$$

Logistic regression model

Logistic regression model:

$$p(X) = \Pr[Y = 1 \mid X] = \sigma(\beta_0 + \beta_1 X) := \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}$$

- $\sigma : z \mapsto \frac{e^z}{1+e^z}$ is called the logistic function (=sigmoid function)
- $p(x) \in (0, 1)$ for all x
- A one-unit increase in X changes the *log-odds* by β_1, so the *odds* are multiplied by e^{β_1}



Decision rule:

- If the threshold is $p^* = 0.5$, predict class 1 when $\beta_0 + \beta_1 x \geq 0$
- More generally, compare $\beta_0 + \beta_1 x$ to $\log(\frac{p^*}{1-p^*})$

An example in R: : Fraud or not?

Scenario: Predict if a transaction is fraud ($Y = 1$) or not ($Y = 0$) using its amount (X)

```
# Simulate toy data:
set.seed(123)
n <- 100
X <- runif(n, 1, 500)  # transaction amount
# true logistic function: p = 1 / [1 + exp(-(-5 + 0.02*X))]
p <- 1 / (1 + exp(-(-5 + 0.02*X)))
Y <- rbinom(n, 1, prob=p)

# Fit logistic regression:
model <- glm(Y ~ X, family=binomial)
summary(model)

# Probability of fraud at X=300:
predict(model, data.frame(X=300), type="response")
```

```
Call:
glm(formula = Y ~ X, family = binomial)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.084377   1.242746  -4.896 9.79e-07 ***
X            0.024664   0.004923   5.010 5.44e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ 

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 138.629  on 99  degrees of freedom
Residual deviance:  53.504  on 98  degrees of freedom
AIC: 57.504

Number of Fisher Scoring iterations: 6

> # Probability of fraud at X=300:
> predict(model, data.frame(X=300), type="response")
        1
0.7883033
```

Check:

- Interpret $\hat{\beta}_0, \hat{\beta}_1$
- $\exp(\hat{\beta}_1)$: multiplicative change in the odds for a \$1 increase in amount

Pop-up quiz #2: Logistic regression coefficients

Suppose you fit the model

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

and obtain $\hat{\beta}_1 = -2.0$.

Question: Which interpretation is most accurate?

A) The estimate is invalid because logistic regression requires $\beta_1 > 0$.

B) For each one-unit increase in X, the probability $p(x)$ decreases by exactly 2.

C) For each one-unit increase in X, the odds of $Y = 1$ are multiplied by $e^{-2} \approx 0.14$.

D) If X increases enough, the predicted probability becomes negative.

Answer: C. A one-unit increase in X changes the log-odds by -2, so the odds are multiplied by e^{-2}.

Estimating logistic regression coefficients

Maximum likelihood estimation (MLE):

- Conditional on the predictors x_i, logistic regression assumes the responses are independent with

$$Y_i \sim \text{Bernoulli}(p_i), \qquad p_i = \sigma(\beta_0 + \beta_1 x_i)$$

- The likelihood function:

$$L(\beta_0, \beta_1) = \Pr\big(\underbrace{(x_i, y_i)_{i=1}^n}_{\text{data at hand}}; \underbrace{\beta_0, \beta_1}_{\text{logistic model}}\big) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}$$

- We choose $\hat{\beta}_0, \hat{\beta}_1$ to maximize L (the parameters for which the given data are most likely)

Why MLE?

- MLE reflects the Bernoulli nature of the response
- Unlike simple linear regression, there is no closed-form solution though

Wrap-up

- Classification vs. regression: Categorical vs. numerical Y
- Linear regression is usually a poor choice for classification because fitted values can fall outside $[0, 1]$ and class coding can be artificial
- Logistic regression models the log-odds linearly:

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

- The logistic function converts the linear predictor into a valid probability $p(x) \in (0, 1)$
- Logistic regression is fit by maximum likelihood, and class predictions come from thresholding estimated probabilities

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.