STA 35C: Statistical Data Science III

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

Dogyoon Song

Spring 2025, UC Davis

Agenda

Last time: Simple logistic regression ($p = 1,\ K = 2$)

Today:

- Extensions of logistic regression
  - Multiple logistic regression ($p > 1$)
  - Multinomial logistic regression ($K > 2$)
- Assessing a classification method
  - Error rate & the Bayes classifier
  - Confusion matrix & false positives/negatives

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

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?

- We want to predict $p(X) = \Pr[Y = 1 \mid X] \in [0, 1]$ ... using a linear model of X
- We need a monotone increasing function $p(X) \in [0, 1] \to f \circ p(X) \in \mathbb{R}$
- 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: Coefficient estimation & prediction

Maximum likelihood estimation (MLE):

- Given data $(x_i, y_i) \in \{0, 1\}$, $p_i = \Pr(Y_i = 1) = \sigma(\beta_0 + \beta_1 x_i)$
- The likelihood function of (β_0, β_1) is

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

- Choose $\hat{\beta}_0, \hat{\beta}_1$ that maximizes $L(\beta_0, \beta_1)$, typically by numerical methods

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

- $\hat{p}(x) = \sigma(\hat{\beta}_0 + \hat{\beta}_1 x)$
- Typically predict $Y = 1$ if $\hat{p}(x) \geq 0.5$; $Y = 0$ otherwise
- Threshold 0.5 can be changed for a different value $p^* \in [0, 1]$

Multiple logistic regression ($p > 1$)

Model:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

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

Interpretation of coefficients:

- β_i: the effect of X_i on log-odds of $Y = 1$, holding other predictors fixed
  - a 1-unit increase in X_i multiplies the odds by e^{β_i}, when other predictors are controlled

Decision boundary:

- The hyperplane $\{\mathbf{x} \mid \beta_0 + \sum_{i=1}^{p} \beta_i x_i = 0\}$; a point if $p = 1$, a line if $p = 2$
- Linear boundary in $(x_1, \ldots, x_p)$

Example: The Default data set mystery

	Coefficient	Std. error	z-statistic	p-value
`Intercept`	−10.6513	0.3612	−29.5	<0.0001
`balance`	0.0055	0.0002	24.9	<0.0001

	Coefficient	Std. error	z-statistic	p-value
`Intercept`	−3.5041	0.0707	−49.55	<0.0001
`student[Yes]`	0.4049	0.1150	3.52	0.0004

	Coefficient	Std. error	z-statistic	p-value
`Intercept`	−10.8690	0.4923	−22.08	<0.0001
`balance`	0.0057	0.0002	24.74	<0.0001
`income`	0.0030	0.0082	0.37	0.7115
`student[Yes]`	−0.6468	0.2362	−2.74	0.0062

Figure: In the `Default` dataset, simple logistic regression shows a significantly *positive* association between `student` and `default`, whereas multiple logistic regression yields a significantly *negative* association [JWHT21, Tables 4.1 - 4.3].

Explanation: Confounding by `balance`

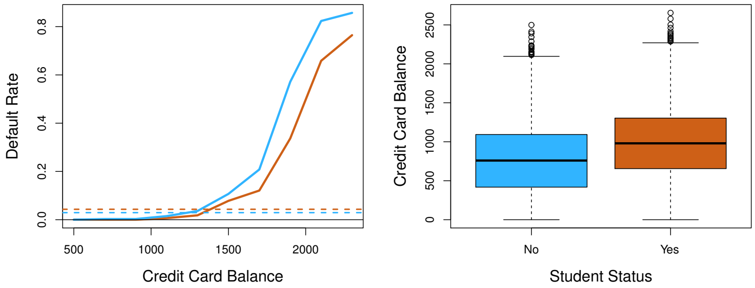

Figure: Confounding in the `Default` dataset. **Left:** default rates for students (orange) vs. non-students (blue). **Right:** boxplots of balance distribution [JWHT21, Tables 4.1 - 4.3].

- Simple logistic: `student` seems positively related to default due to higher overall `default` rate
- Once `balance` is accounted for, students are less likely to default
- Contradiction arises from *confounding* by `balance`; students tend to carry higher balance

Multinomial logistic regression ($K > 2$)

Idea: Use class K as baseline, and model

$$\log\left(\frac{p_k(x)}{p_K(x)}\right) = \beta_{k,0} + \beta_{k,1}X_1 + \cdots + \beta_{k,p}X_p \quad \text{for } k = 1, \ldots, K-1$$

$$\Rightarrow \quad \Pr(Y = k \mid X = x) = \begin{cases} \frac{\exp(\beta_{k,0}+\beta_{k,1}X_1+\cdots+\beta_{k,p}X_p)}{1+\sum_{k'=1}^{K-1}\exp\left(\beta_{k',0}+\beta_{k',1}X_1+\cdots+\beta_{k',p}X_p\right)}, & \text{if } k = 1, \ldots, K-1, \\ \frac{1}{1+\sum_{k'=1}^{K-1}\exp\left(\beta_{k',0}+\beta_{k',1}X_1+\cdots+\beta_{k',p}X_p\right)}, & \text{if } k = K \end{cases}$$

- Each class probability arises from exponentiating its own linear form
- Changing the baseline only alters coefficient representation & its interpretation, not the predicted probabilities

Alternatively, an equivalent *softmax* formulation treats all K classes symmetrically:

$$\Pr(Y = k \mid X = x) = \frac{\exp(\beta_{k,0} + \beta_{k,1}X_1 + \cdots + \beta_{k,p}X_p)}{1 + \sum_{k'=1}^{K}\exp\left(\beta_{k',0} + \beta_{k',1}X_1 + \cdots + \beta_{k',p}X_p\right)}$$

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:

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

- Optimal classifier that minimizes error rate *in theory*
- Usually impossible to compute *in practice*, since $\Pr(Y \mid X)$ is unknown
- **Question:** Even if we could, is the error rate always 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	No	9432	138	9570
default status	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

		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	

Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1−Specificity
True Pos. rate	TP/P	1−Type II error, power, sensitivity, recall
Pos. Pred. value	TP/P*	Precision, 1−false discovery proportion
Neg. Pred. value	TN/N*	

Figure: **Top:** Possible classification outcomes in a population. **Bottom:** Important measures for classification, derived from the confusion matrix [JWHT21, Tables 4.6 & 4.7].

Pop-up quiz: Error metrics

Question: In a binary classification with many more negatives than positives, why might we prefer measures like precision (TP/P*) and sensitivity (TP/P) over overall error rate?

A) Because error rate is always 50% in such cases, regardless of the classifier.

B) Because false positives and false negatives are equally bad in all scenarios.

C) Because error rate can be misleading when one class is rare, while precision/recall better capture performance on the minority class.

D) Because if we have more negatives, the classifier rarely needs to predict $Y = 1$.

Answer: (C) is correct: precision/recall focus on performance for the minority class, which error rate can obscure.

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
- Changing p^* alters false positives and false negatives

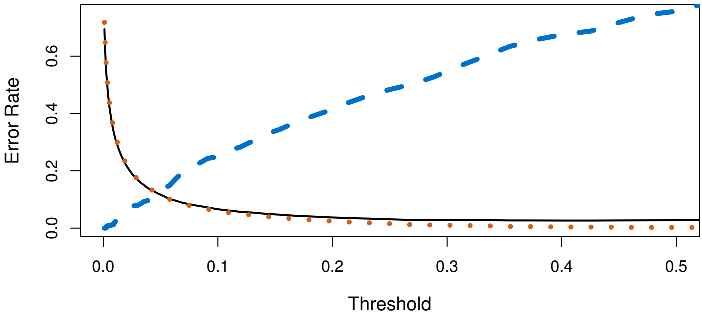

Figure: False positive (orange dotted) and false negative (blue dashed) error rates as a function of the threshold value p^* for the `Default` dataset [JWHT21, Figure 4.7].

Receiver operating characteristic (ROC) curve

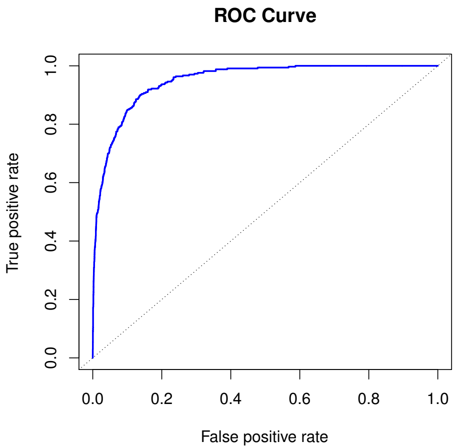

Figure: An example ROC curve, with AUC [JWHT21, Figure 4.8].

ROC curve

- Plot TPR vs. FPR as p^* moves from 0 to 1
- Summarize the performance via area under curve (AUC)

Area under curve (AUC)

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

Wrap-up

Logistic regression:

- Extension to multiple predictors ($p > 1$)
  - Interpretation of coefficients
  - Linear decision boundary
- Extension to $K > 2$ classes (multinomial logistic)
  - Coefficients may differ if baseline class is changed, but predictions remain the same

Assessing classification:

- Error rate & the Bayes classifier
- Confusion matrix, FP/FN & threshold selection
- ROC curve, AUC

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.