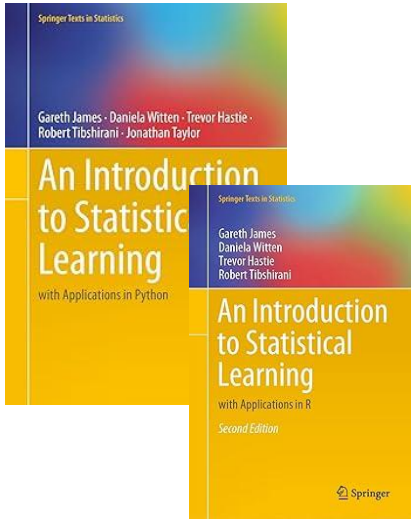
Data Analytics for Operations

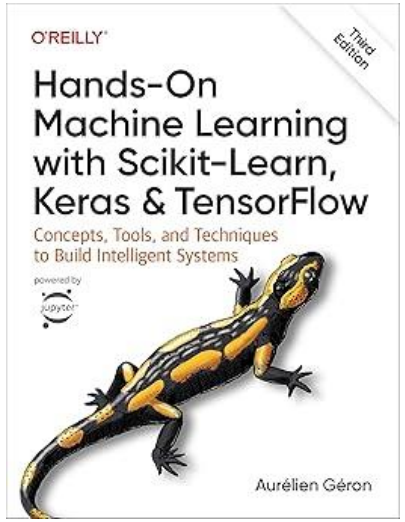
INSY 7120

Lecture 10b – Intro to Classification

Dan O'Leary // dan.oleary@auburn.edu

Today





- start the classification unit!
- Logistic Regression, KNN
- Binary, multiclass, multilabel classification problems
- Gradient descent solver
- Assessment – confusion matrix, precision, recall, ROC AUC
- Data prep considerations, e.g., stratified split, feature scaling
- Imbalanced data
- SKL grid search and pipelines

Classification (ISL Ch 4)

- Regression – predicting numeric response (quantitative)
- Classification – predicting categorical response (qualitative)
  - binary (two categories aka classes)
  - nominal (unordered)
  - ordinal (ordered, even / uneven interval)
- Often predicting the probability that an observation belongs to each categories, assigning the most likely – quantitative basis
- Many models suitable for classification; we will focus on logistic regression, and K-nearest neighbors for now, more in next unit

Examples of Classification

- Image recognition: cat, breed of cat, breeds of cats and dogs
  - binary; multi-class; multi-class, multi-label
- Person in emergency room; symptoms → condition
  - multi-class
- Credit card charge → fraudulent / not
  - binary classification
- DNA sequence → disease
  - multi-class; can be used to infer what differences in the sequences may contribute to the disease
- Wilderness audio recording → species identification
  - multi-class; can be used to study the population of species in an area

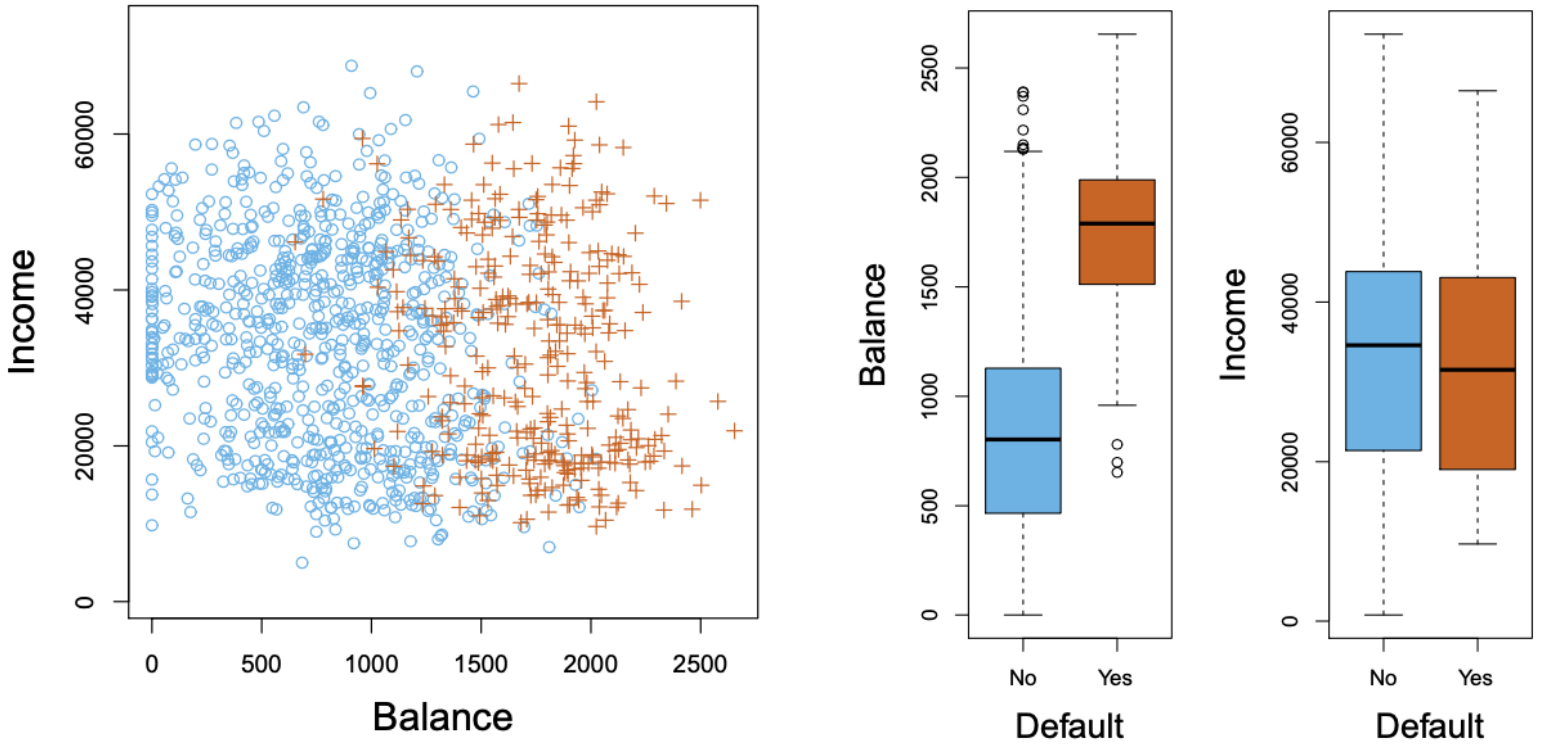


FIGURE 4.1. *The* `Default` *data set.* Left: *The annual incomes and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue.* Center: *Boxplots of* `balance` *as a function of* `default` *status.* Right: *Boxplots of* `income` *as a function of* `default` *status.*

Logistic Regression (4.3)

- Obvious (exaggerated) relationship between balance and default; not typically the case – for illustration purposes
- Given balance (X_1) and income (X_2), predict default (Y)
- Logistic regression models the *probability* that Y belongs to default or not-default classes

$$\Pr(default = Yes \mid balance) = p(balance) \in [0,1]$$

- Then you might predict default for any observation where $p(balance) > 0.5$, a 50% probability
- More conservative: $p(balance) > 0.1$, only 10% chance
- The decision / classification threshold is based on business rules and model tuning

The Logistic Model (4.3.1)

- Must model $p(X)$ using a function that gives outputs between 0 and 1 for all values of $X \rightarrow$ *logistic function*

$$p(x) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

- $p(X)$ has an S-shaped curve that reaches 0 and 1 at the limits
- Why not linear regression?
  - cannot account for classification with more than two classes (order dependence)
  - does not provide meaningful estimates of $\Pr(Y|X)$, even for 2 classes (some predictions outside [0,1] interval
  - for more detail, see section 4.2

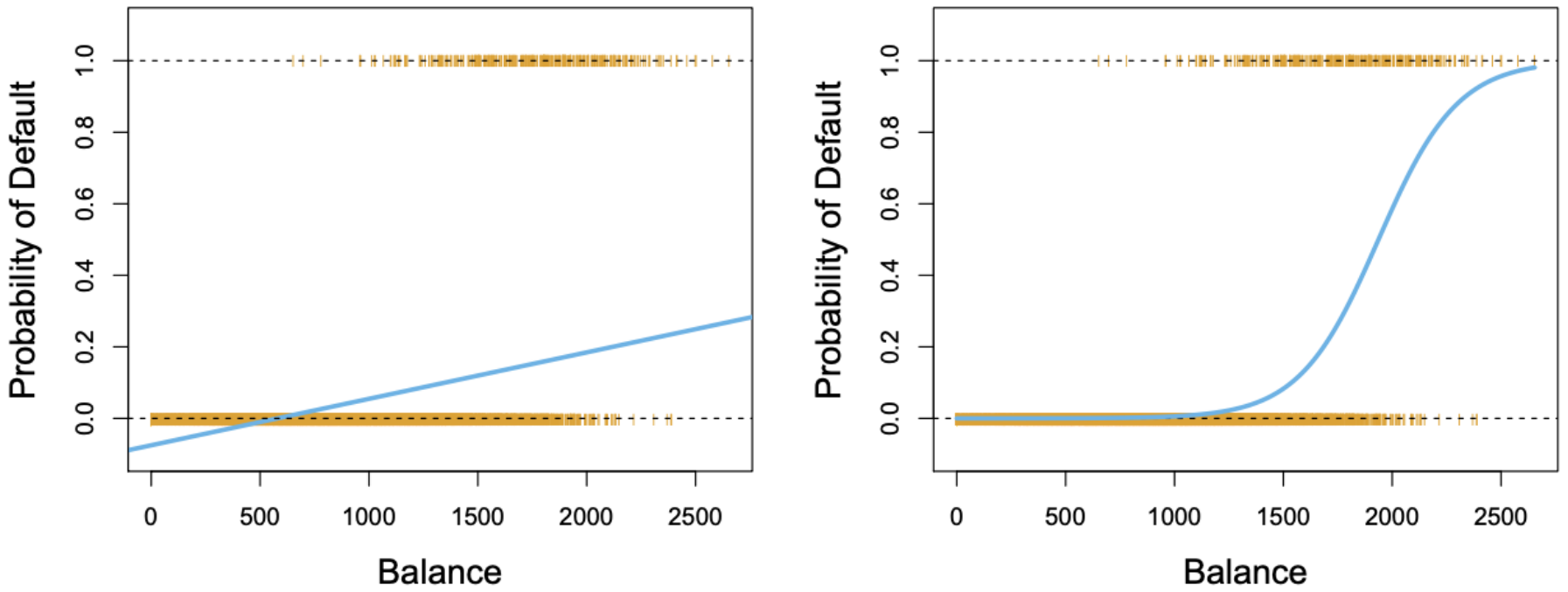


FIGURE 4.2. *Classification using the* `Default` *data.* Left: *Estimated probability of* `default` *using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for* `default` *(*`No` *or* `Yes`*).* Right: *Predicted probabilities of* `default` *using logistic regression. All probabilities lie between* 0 *and* 1.

Assume for a moment that we CAN solve for the coefficients in the case of simple logistic regression…

	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

TABLE 4.1. *For the* Default *data, estimated coefficients of the logistic regression model that predicts the probability of* default *using* balance*. A one-unit increase in* balance *is associated with an increase in the log odds of* default *by* 0.0055 *units.*

for an individual with a balance of \$1,000 is

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1{,}000}}{1+e^{-10.6513+0.0055\times 1{,}000}} = 0.00576,$$

which is below 1 %. In contrast, the predicted probability of default for an individual with a balance of \$2,000 is much higher, and equals 0.586 or 58.6 %.

	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

TABLE 4.2. *For the* `Default` *data, estimated coefficients of the logistic regression model that predicts the probability of* `default` *using student status. Student status is encoded as a dummy variable, with a value of* 1 *for a student and a value of* 0 *for a non-student, and represented by the variable* `student[Yes]` *in the table.*

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=Yes}) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431,$$

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=No}) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292.$$

The Logistic Model (4.3.1)

$$p(x) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

- $p(X)$ has an S-shaped curve that reaches 0 and 1 at the limits
- To interpret the coefficients, some manipulation is required

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

- $p(x)/\big(1-p(x)\big)$ is called the *odds*, can take on any value 0 to ∞
  - 1 in 5 people with an odds of ¼ will default: $p(X) = 0.2 \rightarrow \frac{0.2}{1-0.2} = 1/4$

The Logistic Model (4.3.1)

- The log of the odds term is called the *log odds* or *logit*
- The log odds term is is linear in X:

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

- But there is not a straight-line relationship between X and $p(X)$
- Increasing X by one unit changes the *log odds* by β_1 or you could say that it multiplies the *odds* by e^{β_1}

Estimating the Coefficients (4.3.2)

- β_0 and β_1 are unknown, must be estimated, typically with *maximum likelihood method*
- find coefficients that the predicted probability of the label assigned to each each observation corresponds as closely as possible to the observed label
- for the default example, we're looking for probabilities close to zero for those that did not default and close to 1 for those that did

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))$$ *likelihood function*

Gradient Descent

- No closed-form analytical solution for maximizing likelihood → an iterative optimization algorithm is typically used
- Gradient Descent
  - Start with initial values for β_0 and β_1 (typically random)
  - Compute the gradient (direction of steepest increase) of the log-likelihood function, i.e., the partial derivative of log likelihood with respect to β_0 and β_1
  - Move β_0 and β_1 in the direction of the gradient (for maximizing) or opposite to it (for minimizing the negative log-likelihood)
  - Repeat until convergence
- Analogy: How would a blind person get to the bottom of a hill?

Gradient Descent (HOML Ch4)

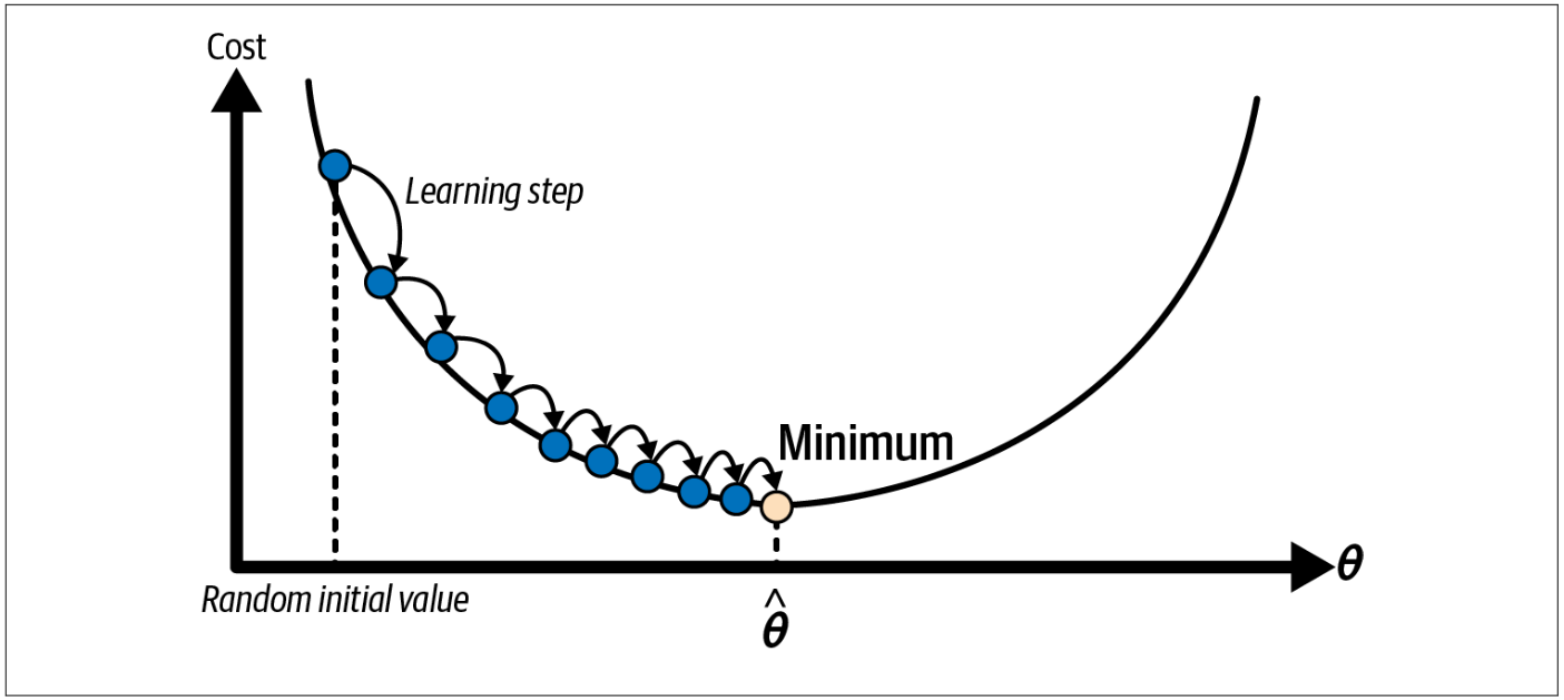


Figure 4-3. In this depiction of gradient descent, the model parameters are initialized randomly and get tweaked repeatedly to minimize the cost function; the learning step size is proportional to the slope of the cost function, so the steps gradually get smaller as the cost approaches the minimum

Gradient Descent (HOML Ch4)

- Important parameter in GD = step size, aka *learning rate*
  - Small learning rate → can be slow to converge
  - To large → can create erratic, divergent behavior

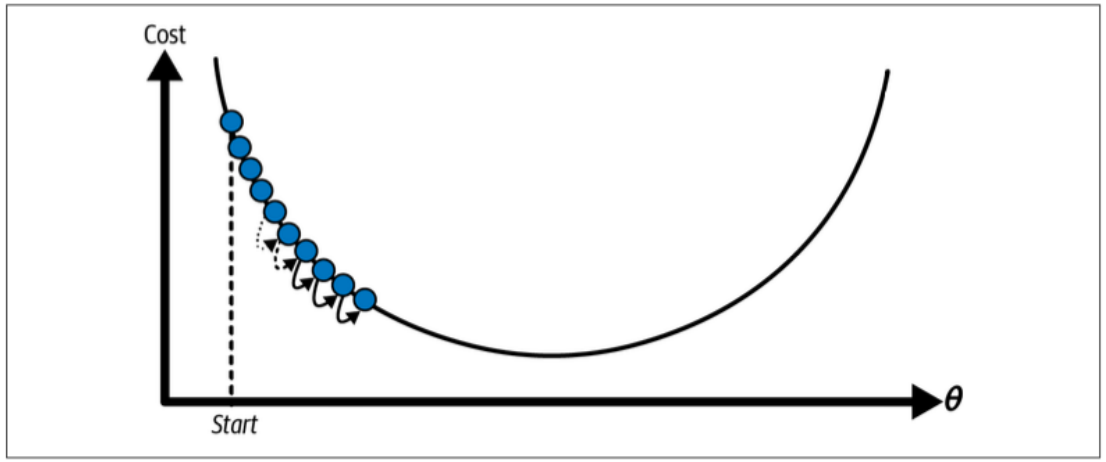


Figure 4-4. Learning rate too small

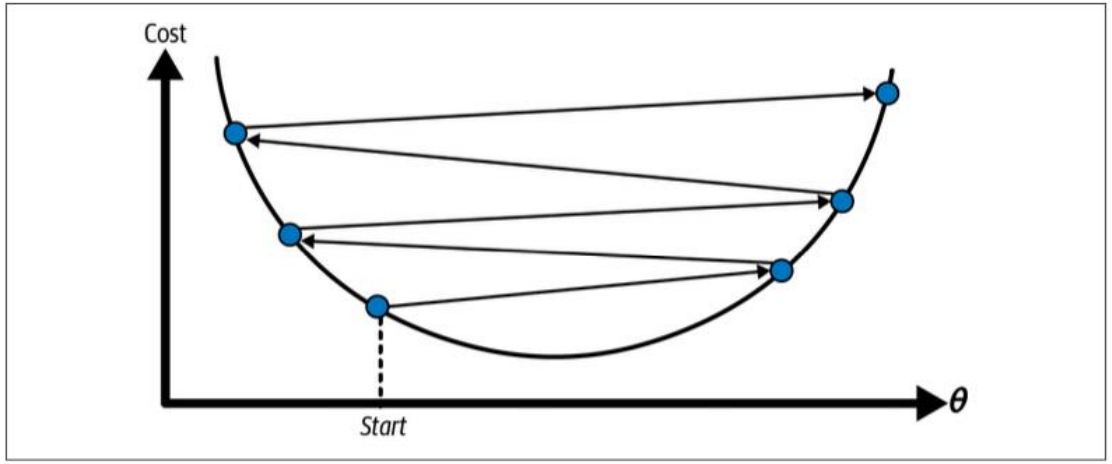
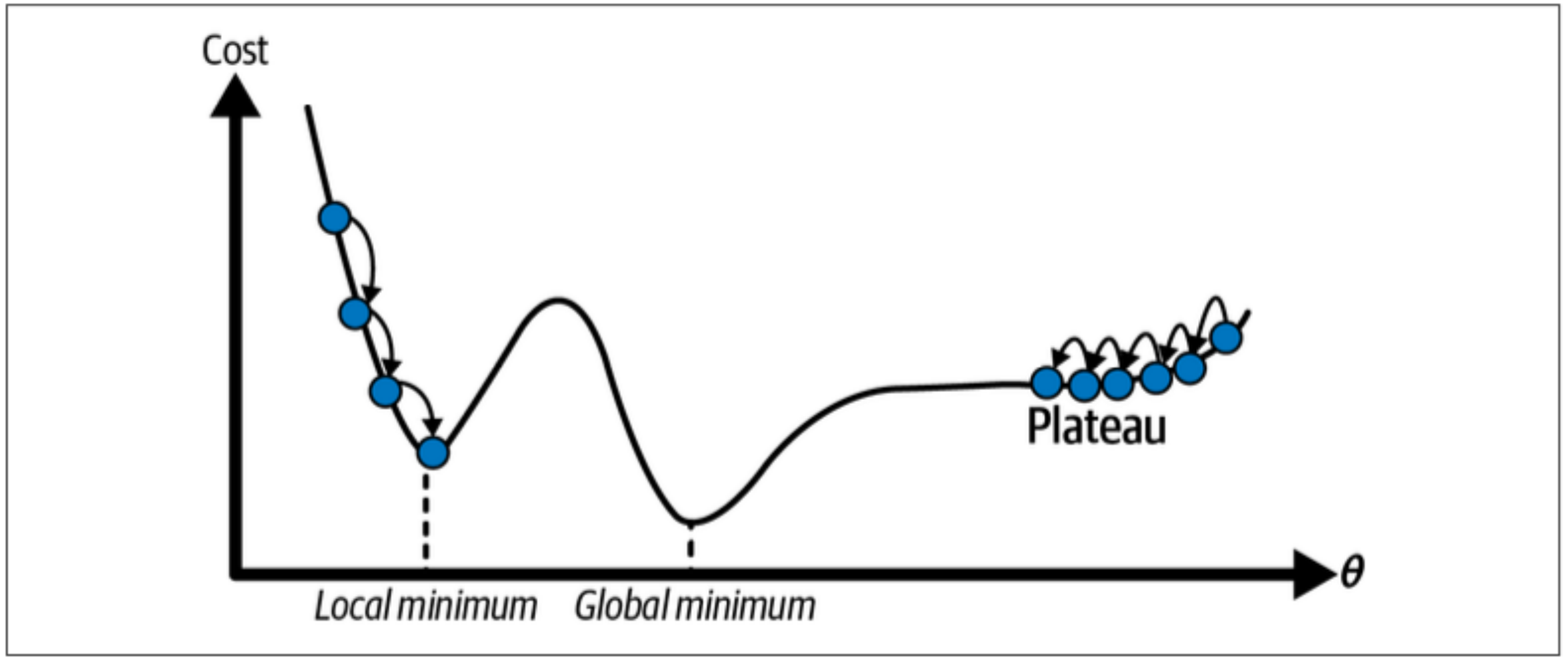


Figure 4-5. Learning rate too high

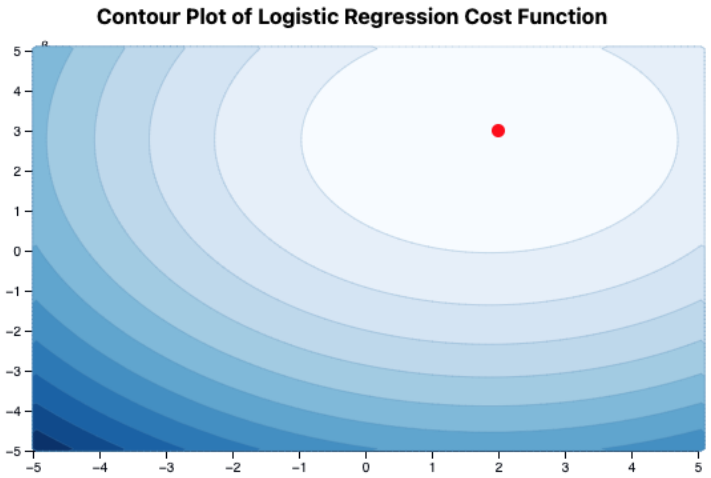
Gradient Descent (HOML Ch4)

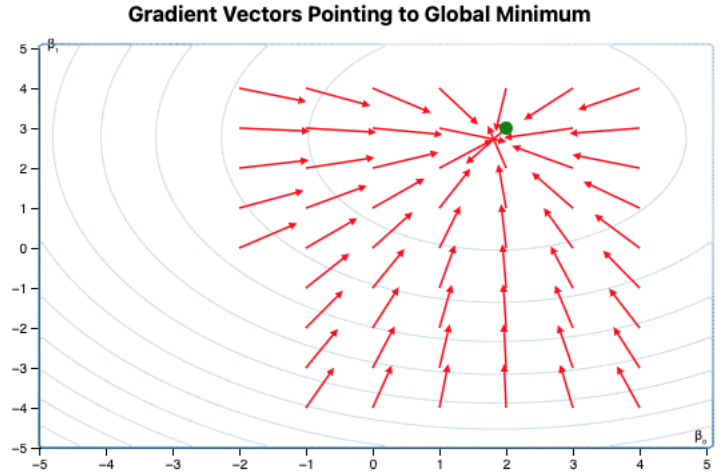
- Not all cost functions have a regular “bowl” shape with a single **global minimum**; may have holes, ridges, plateaus, other irregular shapes
- Result: local minima that may “trick” the algorithm

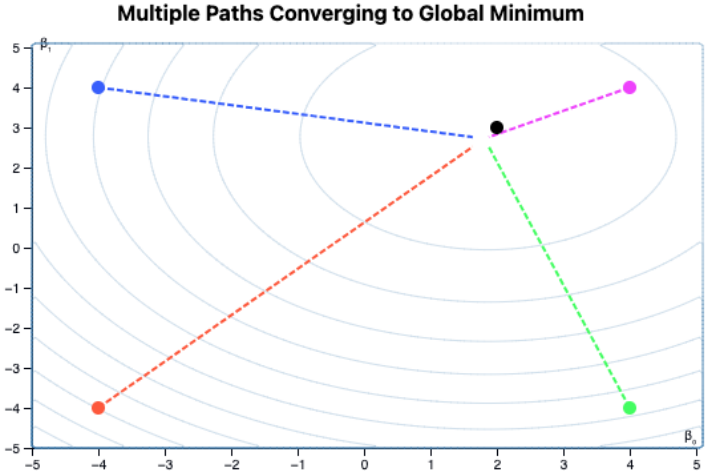


Gradient Descent (HOML Ch4)

- Luckily, the cost function for logistic regression (negative log-likelihood) is convex, which guarantees that:
  - there is only one global minimum (no local minima)
  - gradient descent will converge on the optimal solution for any initialization
  - there is a single, unique solution







Multiple Logistic Regression (4.3.4)

We now consider the problem of predicting a binary response using multiple predictors. By analogy with the extension from simple to multiple linear regression in Chapter 3, we can generalize (4.4) as follows:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p, \tag{4.6}$$

where $X = (X_1, \ldots, X_p)$ are p predictors. Equation 4.6 can be rewritten as

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}. \tag{4.7}$$

As with simple logistic regression, maximum likelihood method is used to estimate the p coefficients

	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

TABLE 4.3. *For the* `Default` *data, estimated coefficients of the logistic regression model that predicts the probability of* `default` *using* `balance`, `income`, *and student status. Student status is encoded as a dummy variable* `student[Yes]`, *with a value of* 1 *for a student and a value of* 0 *for a non-student. In fitting this model,* `income` *was measured in thousands of dollars.*

By substituting estimates for the regression coefficients from Table 4.3 into (4.7), we can make predictions. For example, a student with a credit card balance of \$1,500 and an income of \$40,000 has an estimated probability of default of

$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times1{,}500+0.003\times40-0.6468\times1}}{1+e^{-10.869+0.00574\times1{,}500+0.003\times40-0.6468\times1}} = 0.058. \qquad (4.8)$$

A non-student with the same balance and income has an estimated probability of default of

$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times1{,}500+0.003\times40-0.6468\times0}}{1+e^{-10.869+0.00574\times1{,}500+0.003\times40-0.6468\times0}} = 0.105. \qquad (4.9)$$

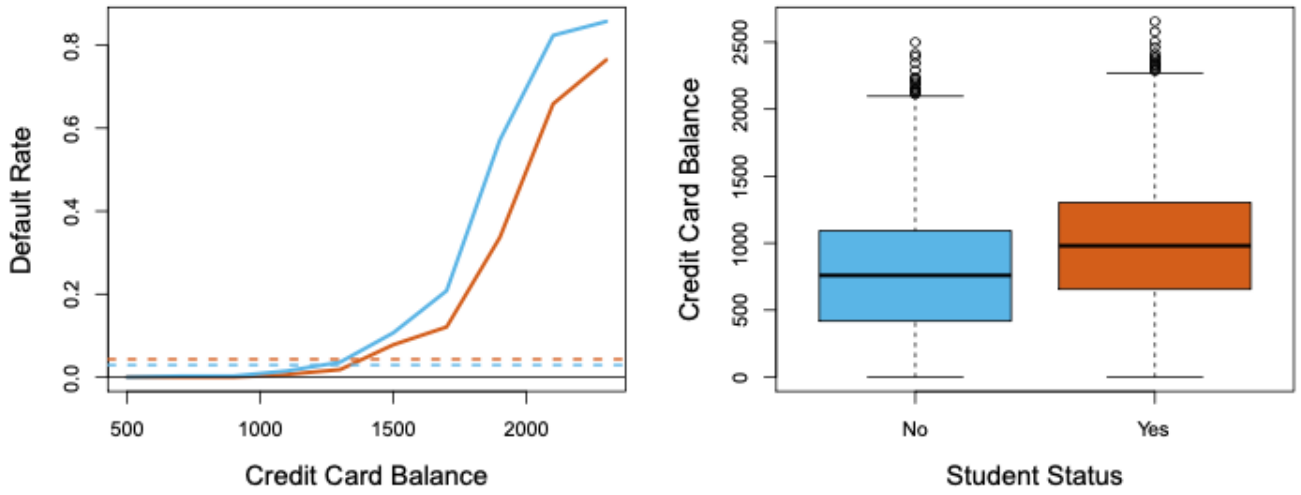


FIGURE 4.3. *Confounding in the* `Default` *data.* Left: *Default rates are shown for students (orange) and non-students (blue). The solid lines display default rate as a function of* `balance`*, while the horizontal broken lines display the overall default rates.* Right: *Boxplots of* `balance` *for students (orange) and non-students (blue) are shown.*

What is the apparent contradiction here and what causes it?