Introduction to Logistic Regression

Ivan Corneillet

Data Scientist

Learning Objectives

After this lesson, you should be able to:

- Build a logistic regression classification model using *scikit-learn*
- Describe the logit and sigmoid functions, odds, and odds ratios as well as how they relate to logistic regression
- Evaluate a model using metrics such as classification accuracy/error

Outline

- Review (train, validation, and test sets; (k-fold) cross-validation)
- Logistic regression and linear regression
- Logit and sigmoid functions; odds and odd ratios
- Putting everything together
- Interpreting the coefficients
- Codealong on the Iris dataset
- Lab
- Review
- In-flight
  - **Unit Project 3 (extended; due next session on 3/29)**
  - Final Project 2 (due in 2.5 weeks)

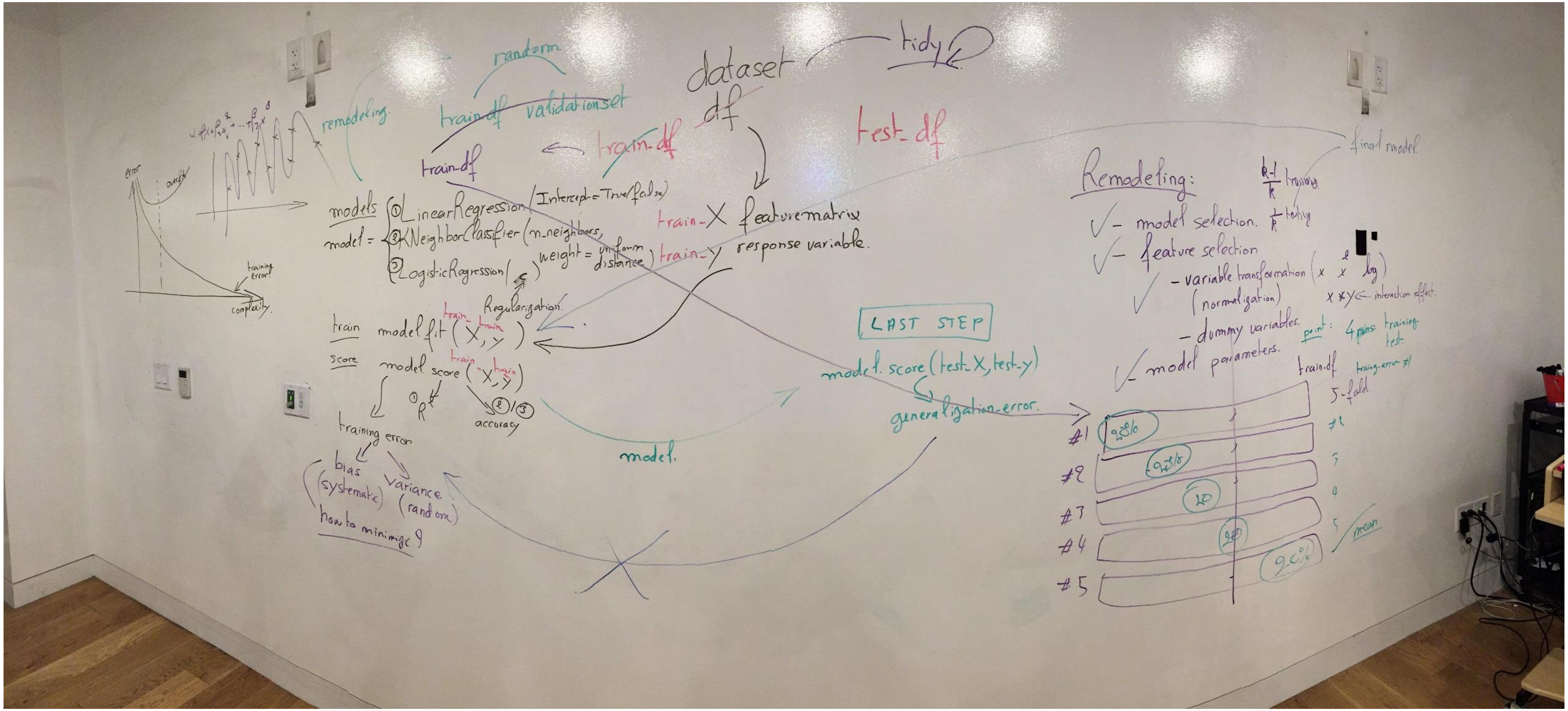
Pre-Work

Pre-Work

Before this lesson, you should already be able to:

- Implement a linear model (`LinearRegression`) with *scikit-learn*
- Define the concept of coefficients
- Recall metrics for accuracy and misclassification
- Recall the differences between L1 and L2 regularization

Review – Train, Validation, and Test Sets; (k-fold) Cross-Validation



dataset
df
tidy
random
train-df validationset
remodeling
train-df
test-df
train-df
models
model = ① LinearRegression (Intercept = True/false)
② KNeighborClassifier (n_neighbors, weight = uniform / distance)
③ LogisticRegression (c)
Regularization
train-X feature matrix
train-y response variable
train model.fit (X, y)
score model.score (X, y)
training error
bias (systematic)
variance (random)
how to minimize ?
accuracy
model.
LAST STEP
model. score (test-X, test-y)
generalization-error.
final model
Remodeling:
model selection.
feature selection
- variable transformation (normalization)
- dummy variables.
model parameters.
interaction effect.
train.df
5-fold
mean
error
complexity
training error
#1
#2
#3
#4
#5

Logistic regression and linear regression

Logistic regression is a generalization of the linear regression model to classification problems

- The name is somewhat misleading
  - "Regression" comes from fact that we fit a linear model to the feature space
  - But it is really a technique for classification, not regression

- We use a linear model, similar to linear regression, in order to solve if an item *belongs* or *does not* belong to a class model
  - It is a binary classification technique: $y = \{0, 1\}$
  - Our goal is to classify correctly two types of examples:
    - Class 0, labeled as 0, e.g., *"belongs"*
    - Class 1, labeled as 1, e.g., *"does not belong"*

Why is logistic regression so valuable to know?

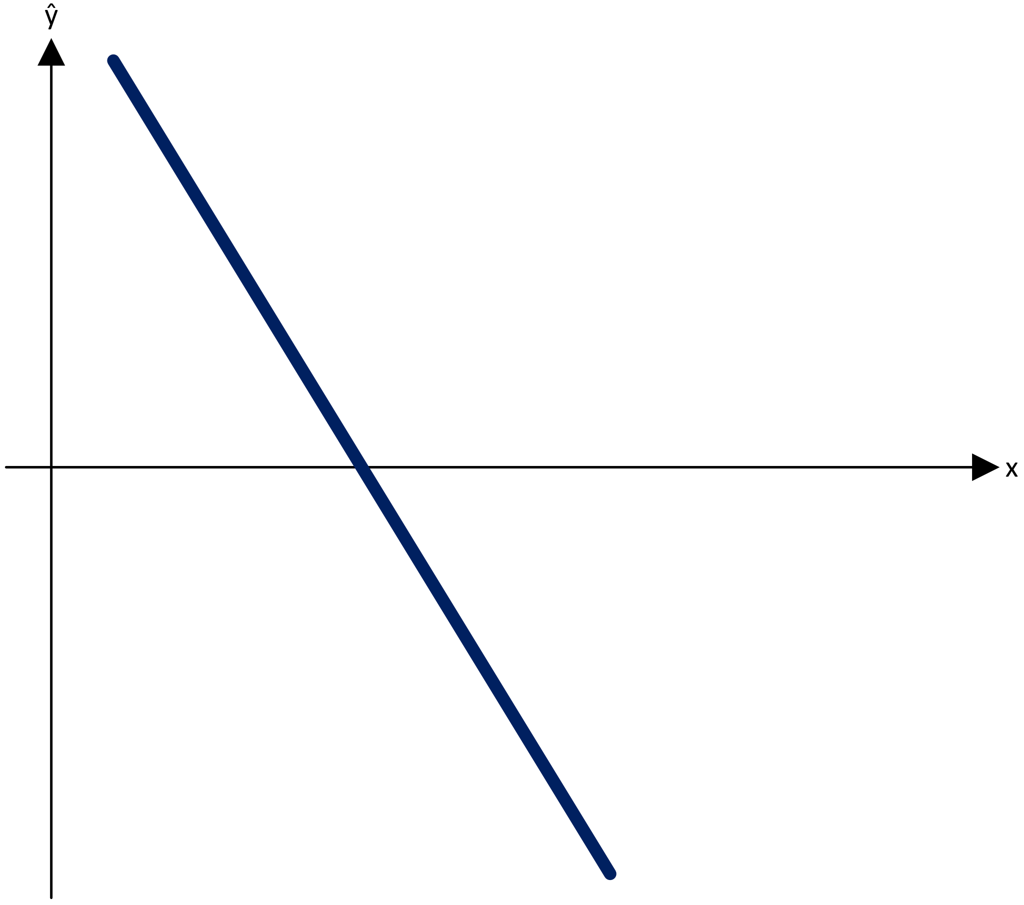
- It addresses many commercially valuable classification problems, such as:
- Fraud detection (e.g., payments, e-commerce)
- Churn prediction (marketing)
- Medical diagnoses (e.g., is the test positive or negative?)
- and many, many others…

With linear regression, $\hat{y}$ is in $]-\infty;\ +\infty[$, not $[0;1]$. How do we fix this for logistic regression?

- The key variable in any regression problem is the outcome variable $\hat{y}$ given the covariate x

$$\hat{y} = \hat{\beta}x$$

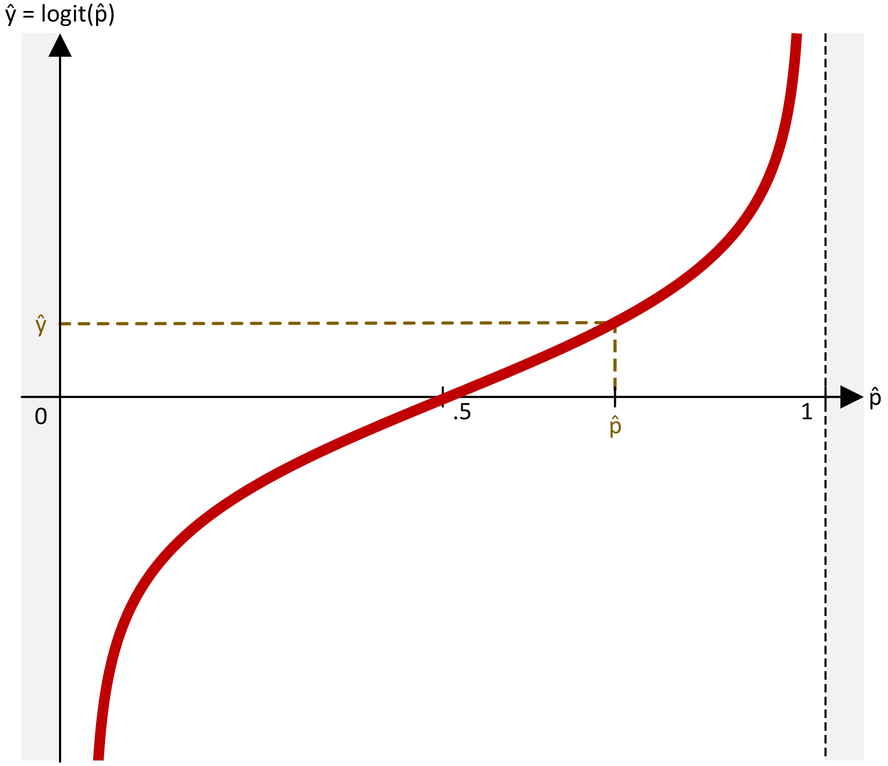
- With linear regression, $\hat{y}$ takes values in $]-\infty;\ +\infty[$
- However, with logistic regression, $\hat{y}$ takes values in the unit interval $[0;1]$



With transformations called the *logit* function (a.k.a., the *log-odds* function) and its inverse, the *logistic* function (a.k.a., *sigmoid* function)

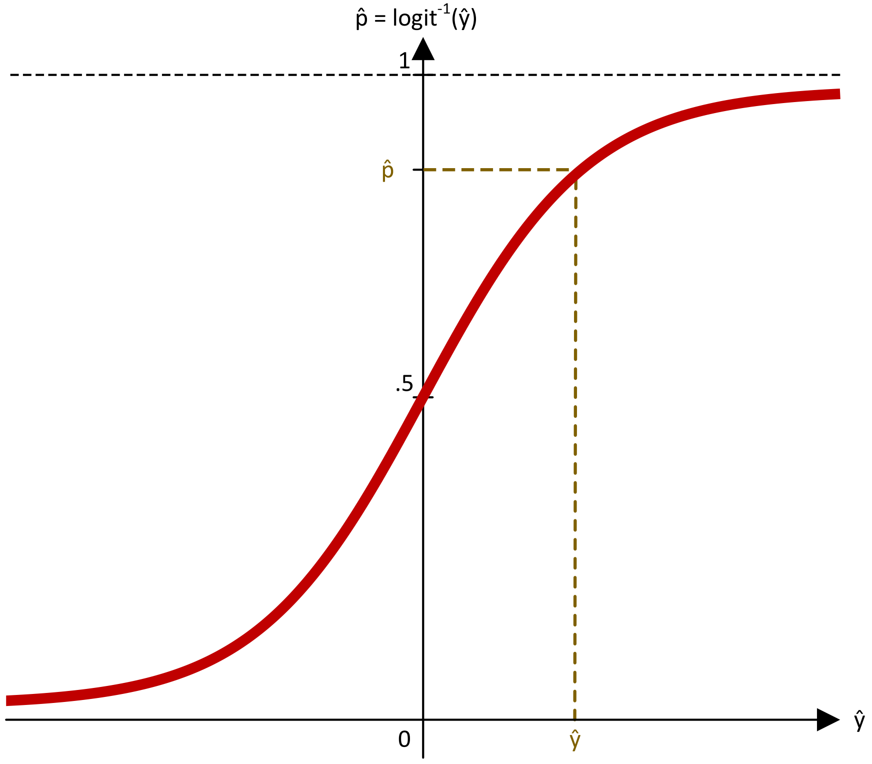
$logit$ maps $\hat{p}$ ([0; 1]) to $\hat{y}$ (]−∞; +∞[)

$$logit(\hat{p}) = ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{y}$$



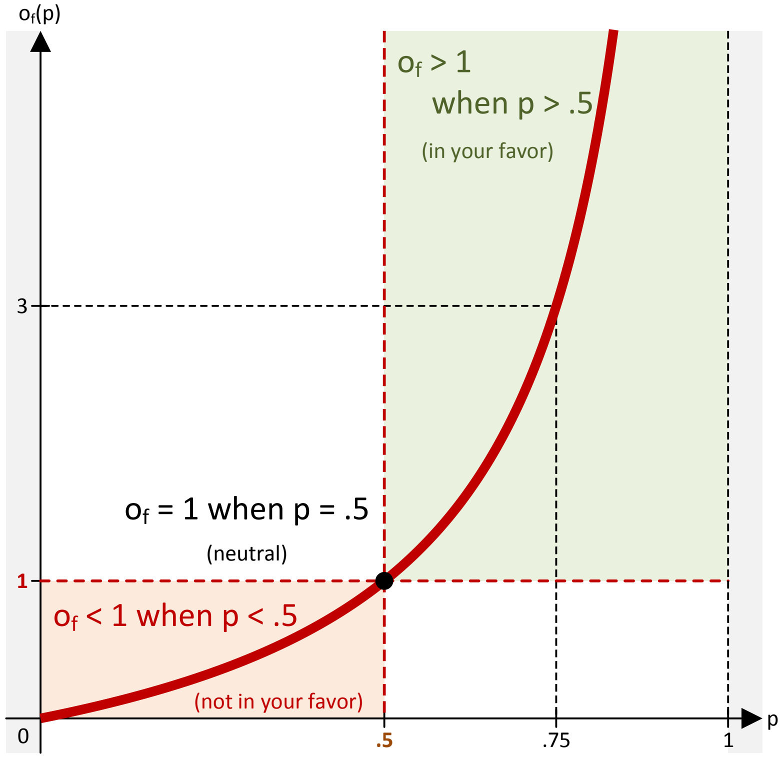
$\pi = logit^{-1}$ maps $\hat{y}$ (]−∞; +∞[) to $\hat{p}$ ([0; 1])

$$\pi(\hat{y}) = \frac{e^{\hat{y}}}{e^{\hat{y}}+1} = \hat{p}$$

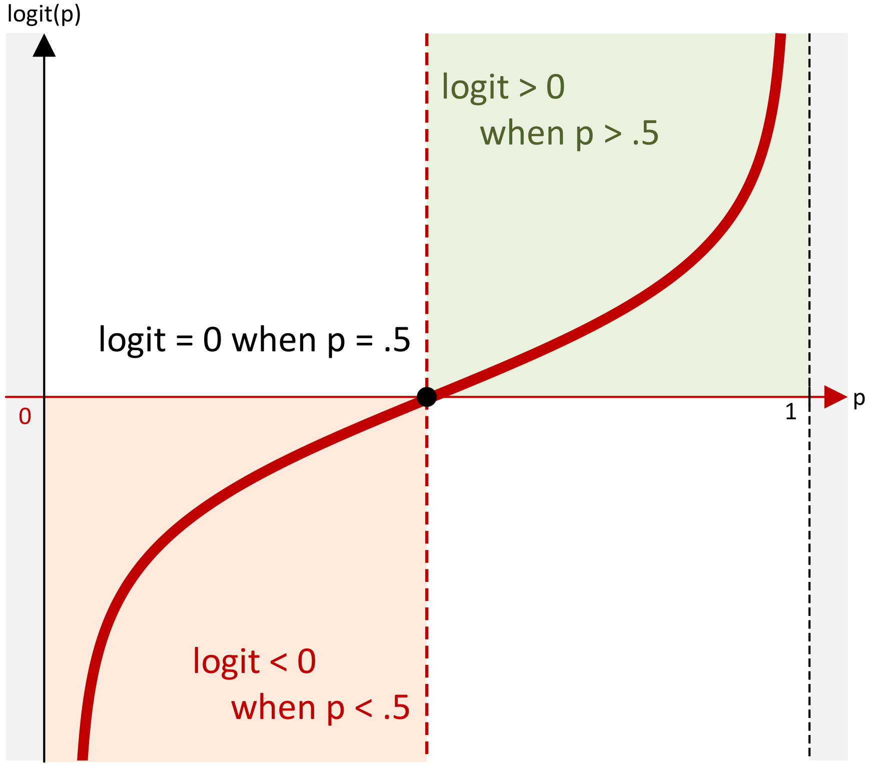


Why is the *logit* function also called the *log-odds* function?

$$\underbrace{o_f}_{odds\ (in\ favor)} = \frac{\overbrace{p}^{probability\ that\ the\ event\ (with\ probability\ p)\ happens}}{\underbrace{1-p}_{probability\ that\ the\ event\ does\ not\ happen}}$$



$$logit(p) = ln(o_f) = ln\left(\frac{p}{1-p}\right)$$



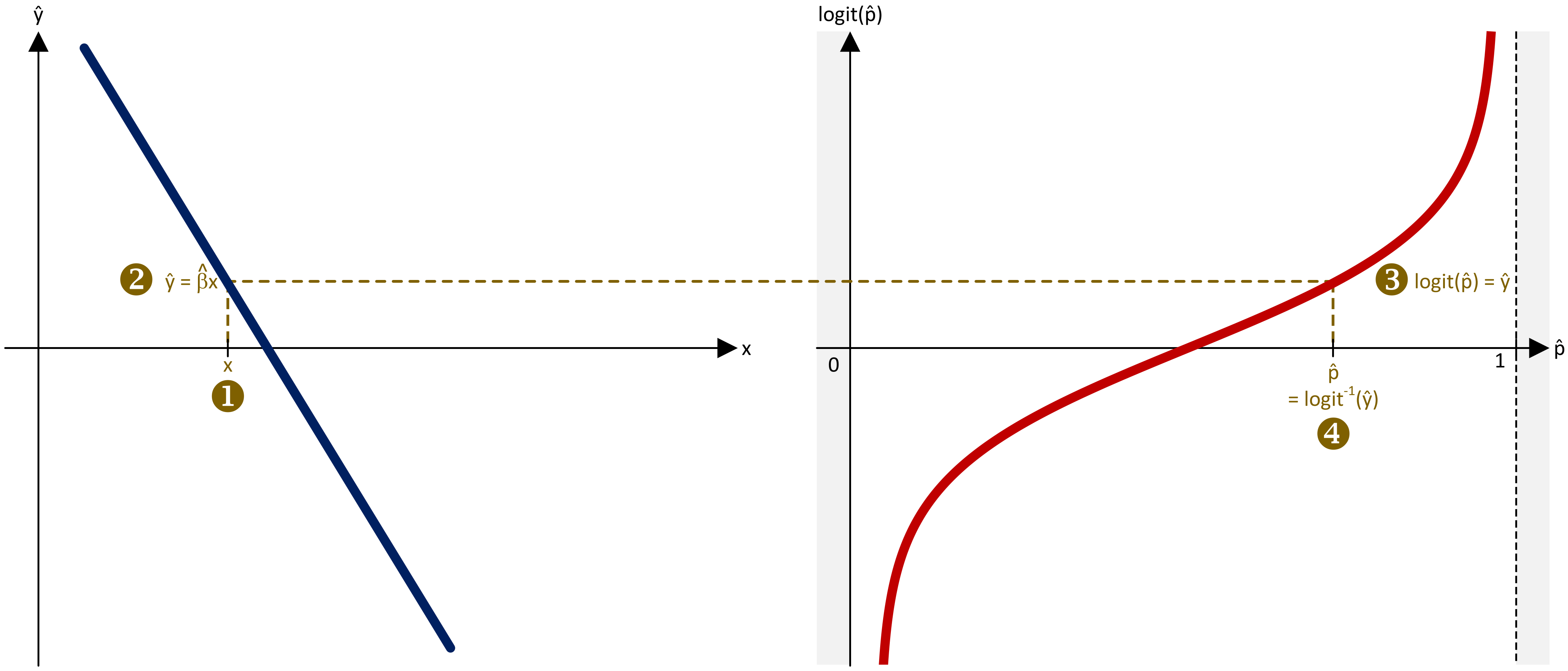
Putting everything together

Logistic Regression

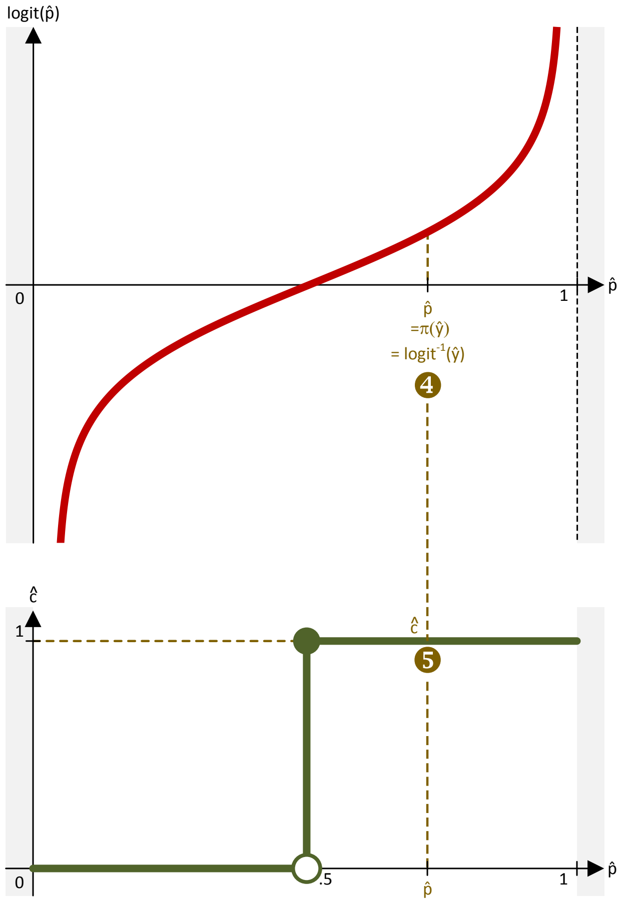
- Putting together $\hat{y} = \hat{\beta} x$ and $\hat{p} = \pi(\hat{y})$ (mapping $\hat{y}$ back to $\hat{p}$), we get

$$\hat{p} = \pi(\hat{\beta} x) = \frac{e^{\hat{\beta} x}}{e^{\hat{\beta} x} + 1} = \frac{1}{1 + e^{-\hat{\beta} x}}$$

$$\hat{p} = logit^{-1}(\hat{y}) = logit^{-1}(\hat{\beta}x) = \frac{1}{1 + e^{-\hat{\beta}x}}$$



Finally, probabilities are "snapped" to class labels (e.g., by thresholding at 50%)



Interpreting the coefficients

Interpreting the coefficients

- With linear regressions, $\widehat{\beta_j}$ represents the change in y for a change in unit of x_j

$$ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\beta}x = \widehat{\beta_0} + \widehat{\beta_1} \cdot x_1 + \cdots + \widehat{\beta_k} \cdot x_k$$

- With logistic regressions, $\widehat{\beta_j}$ represents the **log-odds** change in y for a change in unit of x_j

- This also means that $e^{\widehat{\beta_j}}$ represents the multiplier change in **odds** in y for a change in unit of x_j

$$\frac{\widehat{odds}(x_j+1)}{\widehat{odds}(x_j)} = \frac{e^{\hat{y}(x_j+1)}}{e^{\hat{y}(x_j)}} = e^{\hat{y}(x_j+1)-\hat{y}(x_j)} = e^{(\blacksquare+\widehat{\beta_j}\cdot x_j+\blacksquare)-(\blacksquare+\widehat{\beta_j}\cdot(x_j+1)+\blacksquare)} = e^{\widehat{\beta_j}}$$

Activity: Interpreting the logistic regression coefficients

EXERCISE

ANSWER THE FOLLOWING QUESTIONS (5 minutes)

1. Suppose we are interested in mobile purchasing behavior. Let y be a class label denoting purchase/no purchase, and x denote whether a phone is an iPhone or not. After performing a logistic regression, we get $\beta_1 = .693$. What does this mean?
2. When finished, share your answers with your table

DELIVERABLE

Answers to the above question

Activity: Interpreting the logistic regression coefficients

EXERCISE

1. In this case, the odds ratio change is $e^{\beta_1} = e^{.693} = 2$, meaning the likelihood of purchase is twice as high if the phone is an iPhone

Codealong

The Iris dataset, Take 2

Iris Setosa **Iris Versicolor** **Iris Virginica**

Source: Flickr

Review

Review

You should now be able to:

- Build a Logistic regression classification model using *scikit-learn*
- Describe the logit and sigmoid functions, odds, and odds ratios as well as how they relate to logistic regression
- Evaluate a model using metrics such as classification accuracy/error

Q & A

Exit Ticket

Don’t forget to fill out your exit ticket here