Logistic Regression

Max Turgeon

DATA 2010–Tools and Techniques in Data Science

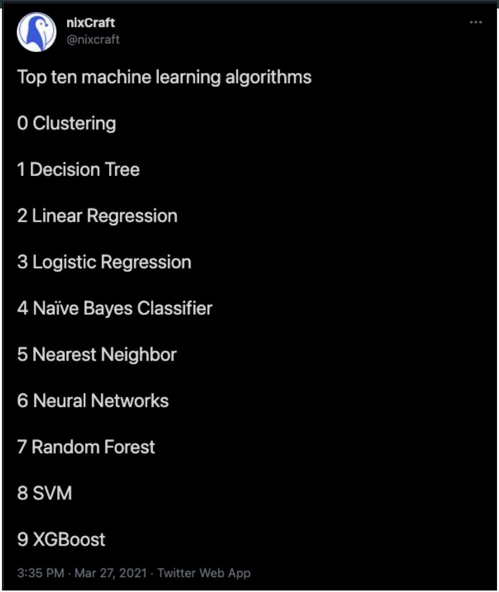
Lecture Objectives

- Fit logistic regression models using R and Python.
- Understand the output and interpret the coefficients.
- Evaluate the model using different metrics.

Motivation i

- In the last module, we discussed linear regression.
  - Measure differences in **averages** between different subgroups.
  - For continuous outcome variables.
- **Logistic regression** is a way to model the relationship between a binary outcome variable and a set of covariates.
  - It's still a regression model, but it can be turned into a classifier.

Motivation ii



nixCraft
@nixcraft

Top ten machine learning algorithms

0 Clustering

1 Decision Tree

2 Linear Regression

3 Logistic Regression

4 Naïve Bayes Classifier

5 Nearest Neighbor

6 Neural Networks

7 Random Forest

8 SVM

9 XGBoost

3:35 PM · Mar 27, 2021 · Twitter Web App

Main definitions

- Y is a binary outcome variable (i.e. $Y = 0$ or $Y = 1$).

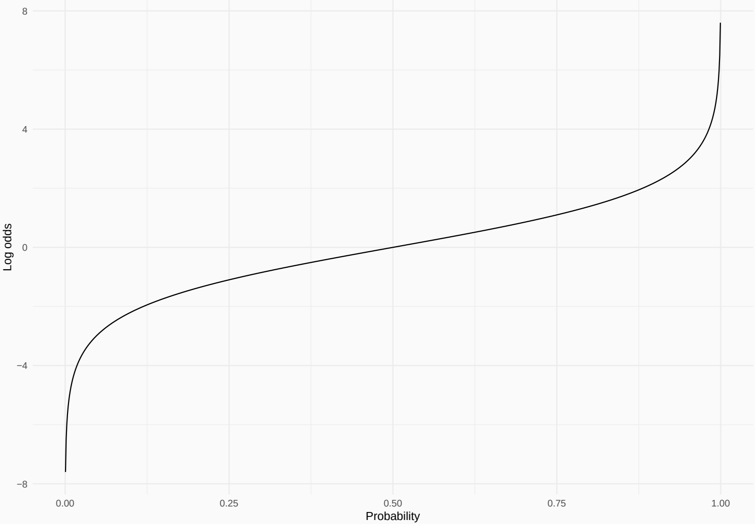
$$\operatorname{logit}\left(E(Y \mid X_1, \ldots, X_p)\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

- Note: $\operatorname{logit}(t) = \log(t/(1-t))$.
- The coefficients β_i represent comparisons of **log odds** for different values of the covariates (i.e. for different subgroups).

Comments i

- If Y is a binary random variable, then $E(Y) = P(Y = 1)$.
- The **odds** is the ratio $P(Y = 1)/P(Y = 0)$.
    - E.g. if the odds is 2, then $Y = 1$ is twice as likely than $Y = 0$.
    - In other words, $P(Y = 1) = 0.66$.
- The **logit function** takes probabilities (which are between 0 and 1) and transforms them to a real number (from $-\infty$ to ∞)

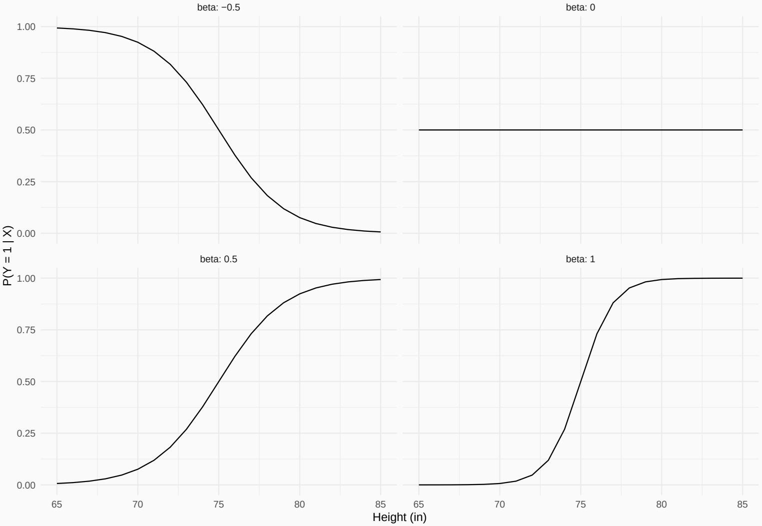
Comments ii



Example i

- Assume we have one covariate X: height in inches.
- The covariate Y: whether someone is a good basketball player (or not).
- Let's look at the effect of β on the relationship between X and $P(Y = 1 \mid X)$.

Example ii



beta: −0.5
beta: 0
beta: 0.5
beta: 1
1.00
0.75
0.50
0.25
0.00
P(Y = 1 | X)
65
70
75
80
85
Height (in)

Example i

- Consider the following 2x2 table:

	Right-handed	Left-handed	Total
Male	43	9	52
Female	44	4	48
Total	87	13	100

- Let Y be handedness, and let X be sex.
- **Note**: The odds for female is $(44/48)/(4/48) = 11$; the odds for male is $(43/52)/(9/52) = 4.78$.

Example ii

```r
library(tidyverse)
# Create dataset
dataset <- bind_rows(
  data.frame(Y = rep("right", 43),
             X = rep("male", 43)),
  data.frame(Y = rep("right", 44),
             X = rep("female", 44)),
  data.frame(Y = rep("left", 9),
             X = rep("male", 9)),
  data.frame(Y = rep("left", 4),
             X = rep("female", 4)))
```

Example iii

```
glimpse(dataset)
```

```
## Rows: 100
## Columns: 2
## $ Y <chr> "right", "right", "right", "right",
"right", "right", "right", "righ~
## $ X <chr> "male", "male", "male", "male",
"male", "male", "male", "male", "mal~
```

Example iv

```r
# Outcome must be 0 or 1
dataset <- mutate(dataset, Y = as.numeric(Y=="right"))

glm(Y ~ X, data = dataset,
    family = "binomial")
```

```
##
## Call: glm(formula = Y ~ X, family =
"binomial", data = dataset)
##
## Coefficients:
```

Example v

```
## (Intercept) Xmale
## 2.3979 -0.8339
##
## Degrees of Freedom: 99 Total (i.e. Null); 98
Residual
## Null Deviance: 77.28
## Residual Deviance: 75.45 AIC: 79.45
```

```r
# Relationship with odds?
log(11)
```

```
## [1] 2.397895
```

Example vi

```r
log(4.78/11)
```

```
## [1] -0.8334547
```

Interpreting coeffficients i

- The regression coefficients in logistic regression measure differences in **log odds**.
    - Or put another way: they measure ratios of odds on the log scale.
    - Very common to take the exponential of coefficients (and confidence intervals).
- Let's start with the example of a single binary covariate X.

Interpreting coeffficients ii

- If $X = 0$, we have

$$\log \frac{P(Y = 1 \mid X = 0)}{P(Y = 0 \mid X = 0)} = \beta_0.$$

- In other words, the intercept term β_0 corresponds to the log-odds when all covariates are equal to zero.

Interpreting coeffficients iii

- Now, let's look at $X = 1$

$$\log \frac{P(Y = 1 \mid X = 1)}{P(Y = 0 \mid X = 1)} = \beta_0 + \beta_1.$$

- Therefore, β_1 is the difference in log-odds between $X = 1$ and $X = 0$.
- Using logarithm rules, the difference in log-odds is the same as the log of the odds ratio.

Exercise

The dataset `case2001` from the `Sleuth3` package contains information about members of the Donner party who got trapped by snow on their way to California.

Using logistic regression, fit a model predicting the survival probability as a function of age. *Hint*: You'll need to transform the variable `Status` from `Died/Survived` to `0/1` (with 1 corresponding to survival).

Solution i

```r
library(Sleuth3)
library(tidyverse)

# First transform outcome to 0/1
dataset <- mutate(case2001,
                  Y = as.numeric(Status == ”Survived”))

fit <- glm(Y ~ Age, data = dataset,
           family = ”binomial”)
```

Solution ii

```
coef(fit)
```

```
## (Intercept)         Age
##  1.81851831 -0.06647028
```

- We can't interpret the intercept, as it would correspond to age 0.
- The coefficient for age is -0.07, which means for two groups whose age differ by 1 year, the log odds differ by -0.07.
- Alternatively, the odds ratio is $\exp(-0.07) = 0.94$.
    - Sometimes you'll see "odds decreased by 6%".

Logistic regression and splines i

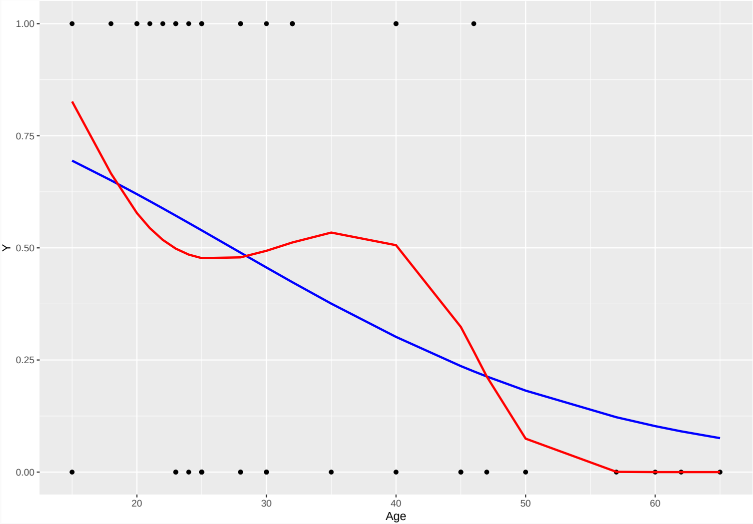
- We can also use splines with logistic regression!

```r
library(splines)
fit2 <- glm(Y ~ bs(Age), data = dataset,
            family = "binomial")
```

Logistic regression and splines ii

```
dataset |>
    mutate(.fitted = fitted(fit),
           .fitted2 = fitted(fit2)) |>
    ggplot(aes(x = Age)) +
    geom_point(aes(y = Y)) +
    geom_line(aes(y = .fitted),
              col = "blue", size = 1) +
    geom_line(aes(y = .fitted2),
              col = "red", size = 1)
```

Logistic regression and splines iii



Evaluating logistic regression models

- We can use the same metrics as for linear regression, except for MAPE (why?).
    - But MSE is instead called the **Brier score**.
- Logistic regression models can also be turned into a classifier.
    - Choose a threshold t.
    - If the fitted value (i.e. estimated probability) is greater than t, classify as positive. Otherwise, classify as negative.
- For a fixed t, we can compute accuracy, precision, etc.
- Alternatively, we can compute these metrics for **all** thresholds t. This is typically summarized using:
    - Receiver Operating Characteristic (ROC) curve.
    - Precision-Recall curve.

ROC curve i

- The ROC curve is defined by plotting the **true positive rate** (TPR) against the **false positive rate** (FPR) for each value of t.
    - TPR is also called the recall
    - FPR is 1 - Specificity.
- When $t = 0$, every observation is called positive. We get $TPR = FPR = 1$.
- When $t = 1$, every observation is called negative. We get $TPR = FPR = 0$.
- As t changes, it draws a curve from the lower-left to the upper-right corner of the unit square.

ROC curve ii

- The closer the curve to the upper-left corner, the better the model.

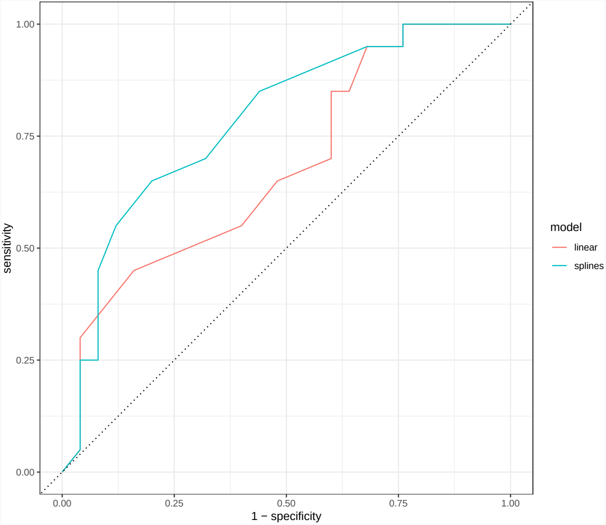
```
# First create dataset with predictions
data_pred <- bind_rows(
    tibble(truth = factor(dataset$Y),
           estimate = fitted(fit),
           model = "linear"),
    tibble(truth = factor(dataset$Y),
           estimate = fitted(fit2),
           model = "splines")
)
```

ROC curve iii

```
library(yardstick)

data_pred |>
    group_by(model) |>
    roc_curve(truth, estimate,
              event_level = "second") |>
    autoplot()
```

ROC curve iv



sensitivity
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1 − specificity
model
linear
splines

ROC curve v

- The second model (with splines) is considered better because it gets closer to the upper-left corner.
    - **Careful**: This is on the training data.
- This can be summarized by a single number, called the **Area Under the Curve** (AUC).
- Perfect classification would give $AUC = 1$, so higher is better.
- We typically want $AUC > 0.5$; otherwise we can get a better classifier by flipping the predictions (positive to negative).

ROC curve vi

```r
data_pred |>
    group_by(model) |>
    roc_auc(truth, estimate,
            event_level = "second")
```

```
## # A tibble: 2 x 4
##   model   .metric .estimator .estimate
##   <chr>   <chr>   <chr>          <dbl>
## 1 linear  roc_auc binary         0.681
## 2 splines roc_auc binary         0.785
```

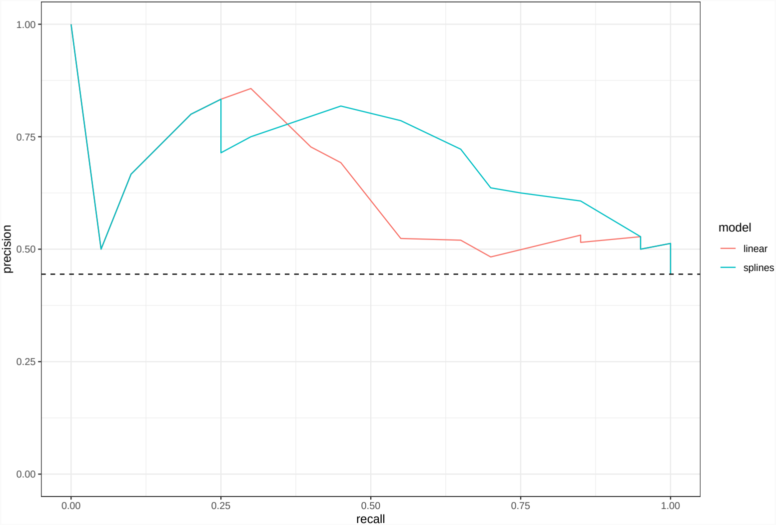
Precision-Recall curve i

- As the name suggests, it's a plot of the precision (on the y-axis) against the recall (on the x-axis), as we change the threshold t.
- When $t = 0$, every observation is called positive. We get Recall = 1, but Precision is the proportion of positive observations in the data.
- When $t = 1$, every observation is called negative. We get Recall = 0 and Precision = 1.
- The closer the curve to the horizontal line Precision = 1, the better the model.

Precision-Recall curve ii

```
data_pred |>
    group_by(model) |>
    pr_curve(truth, estimate,
             event_level = "second") |>
    autoplot() +
    geom_hline(yintercept = 20/45,
               linetype = "dashed")
```

Precision-Recall curve iii



Precision-Recall curve iv

```
# PR curve can also be summarized by the area
# under the curve
data_pred |>
    group_by(model) |>
    pr_auc(truth, estimate,
           event_level = "second")
```

```
## # A tibble: 2 x 4
##   model   .metric .estimator .estimate
##   <chr>   <chr>   <chr>          <dbl>
## 1 linear  pr_auc  binary         0.629
## 2 splines pr_auc  binary         0.700
```

Summary

- Logistic regression is an extension of linear regression for binary outcomes.
    - Easily extended to any binomial outcome.
- Instead of measuring differences in means, regression coefficients measure differences in log-odds.
    - But $\beta = 0$ still corresponds to no association!
- We can measure performance using either regression or classification metrics.
- You can also apply regularization to logistic regression.
- As a prediction model, logistic regression is surprisingly powerful.
    - Neural networks can be seen as a generalization.