DECISION PATTERNS

Logistic Regression

Practical Machine Learning (with R)

UC Berkeley

LOGISTIC REGRESSION

Background

Categorical Modeling:

$$\hat{y}_{cat} = f(\vec{x})$$

- Inputs
  - Categorical
  - Continuous variable can assume any value

Outputs:

How do we handle categories?

- same as linear regression?

Background

- Errors!

$$\hat{\boldsymbol{y}}^{\boldsymbol{cat}} \neq \boldsymbol{y}$$

  - Problem …

$$argmin_{\boldsymbol{\beta}} \sum \begin{Bmatrix} 1 \mid \hat{y} \neq y \\ 0 \mid \hat{y} = y \end{Bmatrix}$$

FUNCTION …

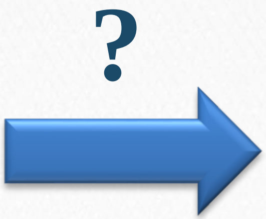
- Do the easiest thing first …
  Start with 2 categories "binomial dist"
  - A | B
  - TRUE | FALSE
  - 0 | 1

"Looks Math-y"

Need a tool ...

Inputs

`(-Inf, Inf)`



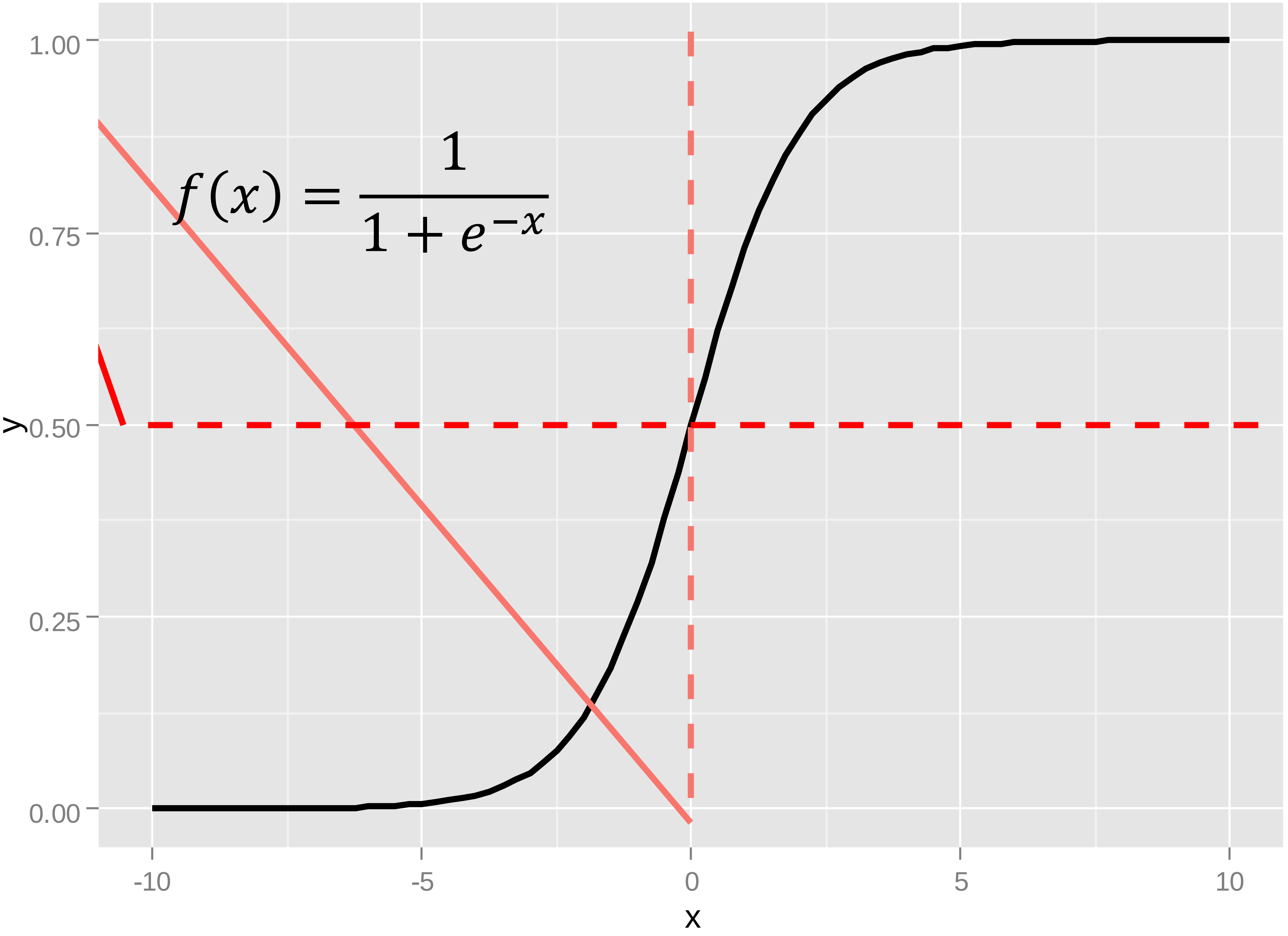
Outputs

[0,1]

$$f(x) = \frac{1}{1 + e^{-x}}$$

Logistic function

$$P(y) \sim \hat{y} = \frac{1}{1 + e^{-x}}$$



$$f(x) = \frac{1}{1 + e^{-x}}$$

Now What

- Proceed as we would with linear regression … and look for **β**'s

$$\hat{y} \sim \frac{1}{1+e^{-x}}$$

$$\hat{y} \sim \frac{1}{1+e^{-\beta_0+\sum_{i=1}^{p}\beta_i x_i}}$$

- Then solve as linear regression:

$$argmin_{\boldsymbol{\beta}}\left(\sum(\hat{y}-y)^2\right)$$

Logistic Regression Summary

```
Call:
glm(formula = Versicolor ~ . - Sepal.Length, family = binomial,
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1262  -0.7731  -0.3984   0.8063   2.1562

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    6.9506     2.2261   3.122  0.00179 **
Sepal.Width   -2.9565     0.6668  -4.434 9.26e-06 ***
Petal.Length   1.1252     0.4619   2.436  0.01484 *
Petal.Width   -2.6148     1.0815  -2.418  0.01562 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 190.95  on 149  degrees of freedom
Residual deviance: 145.21  on 146  degrees of freedom
AIC: 153.21

Number of Fisher Scoring iterations: 5
```

Log Odds

Variable
- Significance?
- Importance?

Not Done

- How do you go from [0,1] back to our binomial categories?

- Choice is somewhat arbitrary
  - `P=0.5`
  - Calibrate response

- Often don't care … you are interested in the probability anyway.

Questions

- Why not just use linear regression?
- What does a unit increase in x_1 correspond with?
- How are odds defined?
- What is the output of the logistic model? How is it interpreted?
- How do you get a class/label from the model?

APPENDIX

Worked Example: GermanCredit

Worked Example: NYC Flights