Machine Learning Fundamentals

Practical Machine Learning (with R)

UC Berkeley

Spring 2016

Agenda

- Administrativa
  - Role Call
  - Missing data from class-list.xlsx
  - Images
  - Assignments due to github
  - Class Google Group (All joined)
- Expectations (Review)
- New Topics
  - R Meetup

REVIEW

Git

- Pulled changes from class Git Hub repository as of last Wednesday

- Attempted/Completed `02-exercises.Rmd`

- Added
- Committed
- Pushed to *your* Git Hub repository

R Skills

- You have tried
  - ***dplyr/tidyr*** and/or
  - ***data.table***

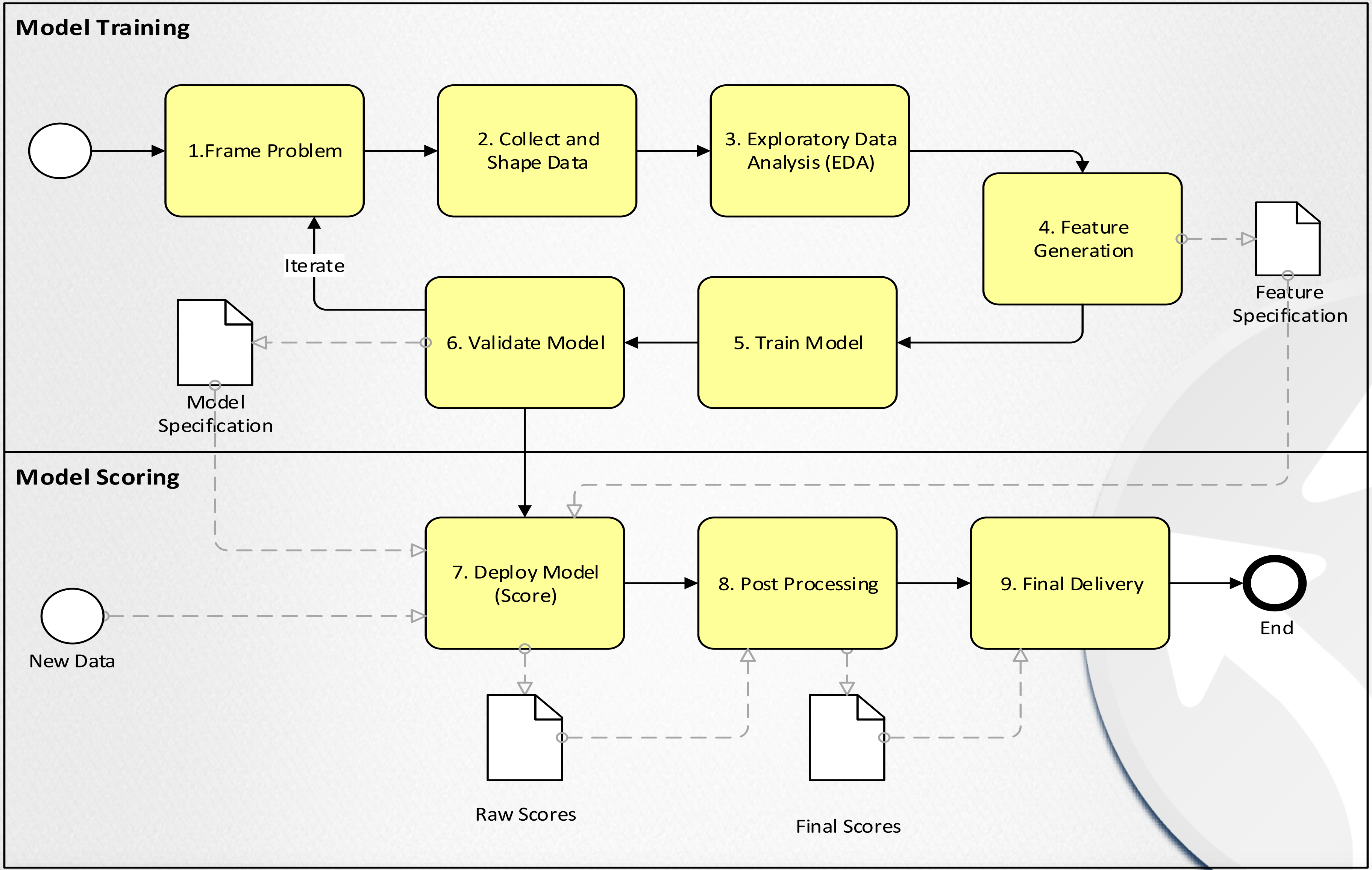
- You know what `%>%` does and love it

- Comfortable plotting
  - Feature vs. response
  - Estimate vs. actual
  - Add *lines* and *trend lines* to plot

Concepts

- Difference between
  - supervised and unsupervised models
  - Semi-supervised
  - Adaptive learning
- Difference between classification and regression
- Three components for ML algorithms …

Expectations: Process



3 Requirement for Algorithm

- A method for evaluating how well the algorithm performs (**ERRORS**)
- A restricted class of function (**MODEL**)
- A process for proceeding through the restricted class of functions to identify the functions (**SEARCH/OPTIMIZATION**)

Reading

- Chapters 3.2-3.7, skim 3.8 "Transformations"
  - Centering and scaling `?scale`
  - Skewness: log, sqrt, inverse, box-cox `E1071:skewness MASS::boxcox`
  - Missing values
    - Remove
    - Impute
  - Feature/Predictor remove: irrelevance, p>n
  - Collinearty of Predictors: ?cor
    - PCA,
    - Iterative feature removal
  - Binning predictors (problems – loss of precision)
  - Dummy variables
    - Loss of precision → increase in error

READING

- Chapters 6.2 and 6.3

Linear Regression model

→ Abstract to multiple dimensions

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \dots$$

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

Mathy-r !!!

LINEAR REGRESSION

You should be able to:

- Extract the coefficients
- Express the models as an ***equation***
- Use the model to ***predict*** responses for new data

Linear Regression

- train a linear regression model
- Interpret linear regression model
  - "stars" (significance), `Estimate`, `Std. Error`, `R-squared`, `Pr(>|t|)`

```
Call:
lm(formula = FE ~ EngDispl, data = cars2010)

Residuals:
    Min      1Q  Median      3Q     Max 
-14.486  -3.192  -0.365   2.671  27.215 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  50.5632     0.3985  126.89   <2e-16 ***
EngDispl     -4.5209     0.1065  -42.46   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.624 on 1105 degrees of freedom
Multiple R-squared:   0.62,	Adjusted R-squared:  0.6196 
F-statistic:  1803 on 1 and 1105 DF,  p-value: < 2.2e-16
```

LINEAR REGRESSION (PREDICTOR SIGNIFICANCE)

`Pr(>|t|)`

Linear regression t-statistic is the probability that the "true value" of the statistic falls outside the student t-distribution.

- Is expressed as a probability.
- Lower is "better" i.e. more significant

Think of it (loosely) as the probability of the coefficient being wrong. It's an estimate after-all.

Indication of Bad Model Fit

These are signs of a bad model fit:

- No significant coefficients / predictors
- Many insignificant predictors
- Coefficients … too large or too small
- Low R-squared
- Skewed or non-zero centered residuals

Errata: Linear Regression Errors

- Two different types of errors measured
  - For ***fitting*** models
  - For **comparing** models

- Minimize square error loss (SSE) ***sum of squared errrors***

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

- *choose Beta such that the sum of squared errors is minimized.*
- Solved by Direct Solution or Numerical Optimization

LINEAR REGRESSION (INTUITION)

- Which is the more important variable?

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  51.3541     0.4593 111.814  < 2e-16 ***
EngDispl     -3.7454     0.2507 -14.941  < 2e-16 ***
NumCyl       -0.5880     0.1722  -3.414 0.000664 ***
```

- Coefficients … multiply then sum
- Number Line (in units of the response)
  - Start at intercept
  - Multiple term by value of the variable
  - Move those number of units of y.

Linear Regression (Intuition)

- Data is generated by an unknown stochastic process that the model creates the data, i.e. x's

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots$$

- Deterministic : always produces the same answer
- Stochastic: non-deterministic, contains some element of randomness, but not entirely random.

Linear Regression Limitations

Limitation	Solution
Linear Response Does not fit higher order functions or interactions	• Transform data • Express in Model Formula
Insignificant Predictors Left in the Model	• Use model variant that does feature selection • Use Recursive Feature Elimination (RFE) routines
Sensitive to inputs: Outliers give out-sized influence on model fit	• Remove outliers • Transform Predictors • Use Robust Regression
Highly correlated predictors yield non-sensical models	• Use Regularization • RFE
Comparatively not sensitive	• ???

TRANSFORMATIONS

- Centering and Scaling: `scale`*
- Resolve skewness: `log, sqrt, inv`
- Resolve outliers: spatial sign, `PCA`

Some algorithms require scaling

Some are insensitive

Time consuming

Somewhat of an art

- Genetic algorithms (GA)

Add complexity

Contribute to loss of interpretability

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 …

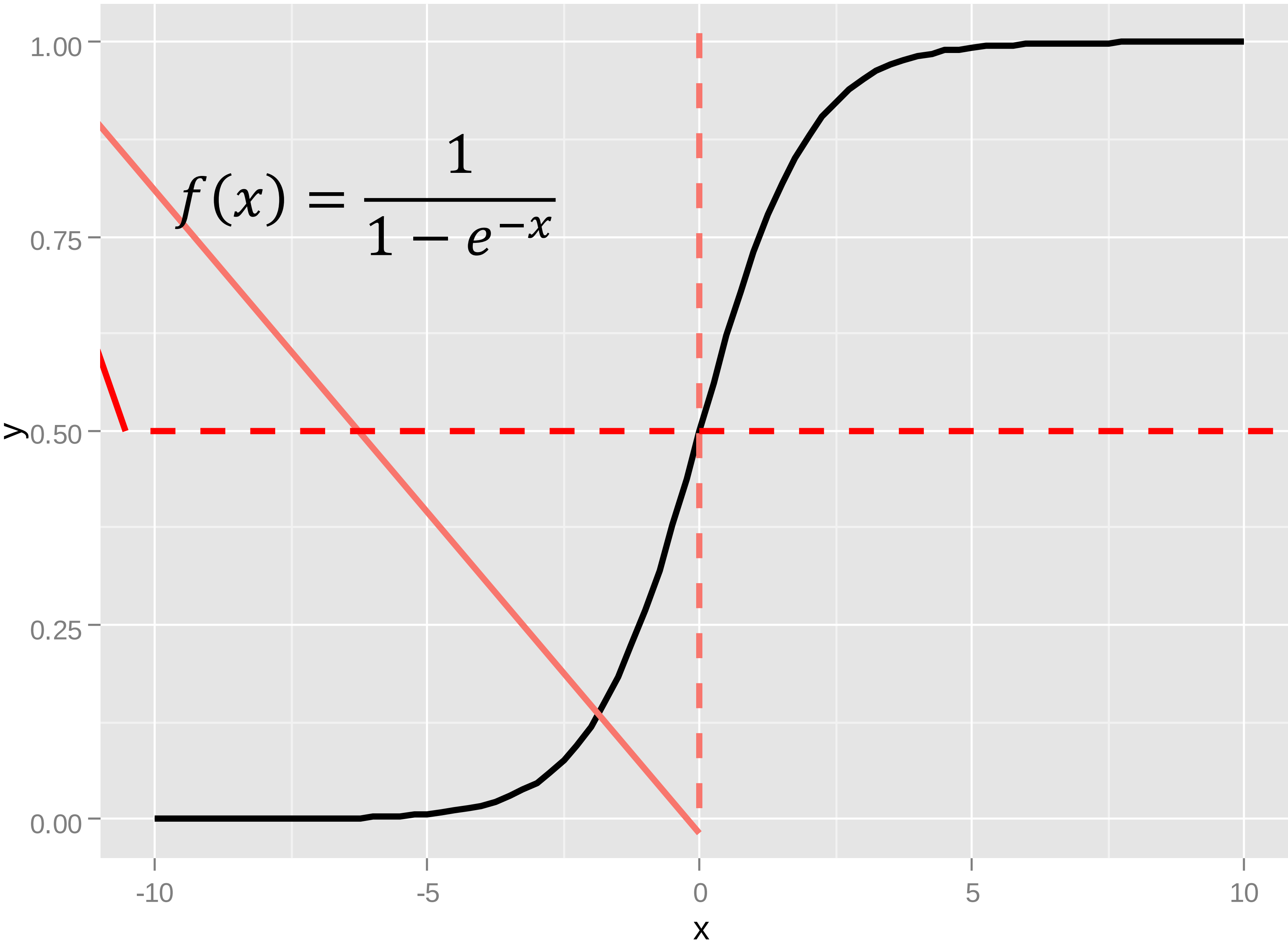
Inputs ? Outputs

`(-Inf, Inf)` → [0,1]

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

Logistic function

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



$f(x) = \frac{1}{1 - e^{-x}}$
1.00
0.75
0.50
0.25
0.00
y
-10
-5
0
5
10
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:

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

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.

Worked Example: GermanCredit

APPENDIX