DECISION PATTERNS

Machine Learning Fundamentals

Practical Machine Learning (with R)

UC Berkeley

Spring 2016

Topics

- Administrativa
  - Role Call
  - Missing data from class-list.xlsx
  - Images
  - Assignments due to github
  - Group (Joined)

REVIEW

R Primer

Help in R

`?,help,??,apropos`

Accessors

Access by name	`$name`
Access interpreted	`[[…]]` "reach" inside
Subset/slice	`[ …]`

Data Structures

- Vectors
- Lists
- data.frame — data.frame, data.table, tbl(dplyr)

Operators

Control Flow

`?Control`

Model Formula

`-tk`

USEFUL R PACKAGES

```
> install.packages("package-name")
```

- ***ML Framework:*** *caret* (Classification and Regression Training)
- **Pipe operators:** *magrittr (ctrl+shift+m), pipeR, backpipe*
- **Tables:** *data.tables, dplyr*
- **Visualization:** *ggvis, ggplot2*
- **Reporting:** *knitr, rmarkdown, shiny*
- ***Misc:*** *devtools (packages), hash*

DATA.FRAME VS. DATA.TABLE VS. DPLYR

EXPECTATIONS: R

- You have installed **R** and **Rstudio**
- If you are new to **R**, you will have checked out one of the resources and have started becoming familiar with syntax and functions.
- You have attempted the first assignment

EXPECTATIONS: GIT

- You have:
  - installed **git** and created a github account
  - **fork**ed the class repo(sitory)
  - **clone**d a local copy of the repo
  - **pulled** new changes
  - *completed the assignment*
  - **add**ed/**commit**ted your solutions
  - **push**ed the assignment back to your repo
- Now: **pull** changes from `csx460/csx460.git`

EXPECTATIONS: READING

- Read *APM* Chapters 1, 2.
- Understand terminology:
  - instance, observation, data point, sample
  - Training / test / validation set
  - Predictors, independent variables, attributes, descriptors, features
  - Outcome, dependent variable, target, class, response
  - Continuous / categorical / ordinal
  - Model building, training, parameter estimation
- Notation

EXPECTATION: READING 2

- **Type** of Response:
  - Continuous → **REGRESSION**
  - Categorical* → **CLASSIFICATION**

    *Binary is a special case

- **Availability** of "labelled" Responses
  - Available → **SUPERVISED**
  - Unavailable → **UNSUPERVISED**

  - Sometimes available/inferable → **SEMI-SUPERVISED**
  - Avail. as training progresses → **ADAPTIVE/REINFORCEMENT**

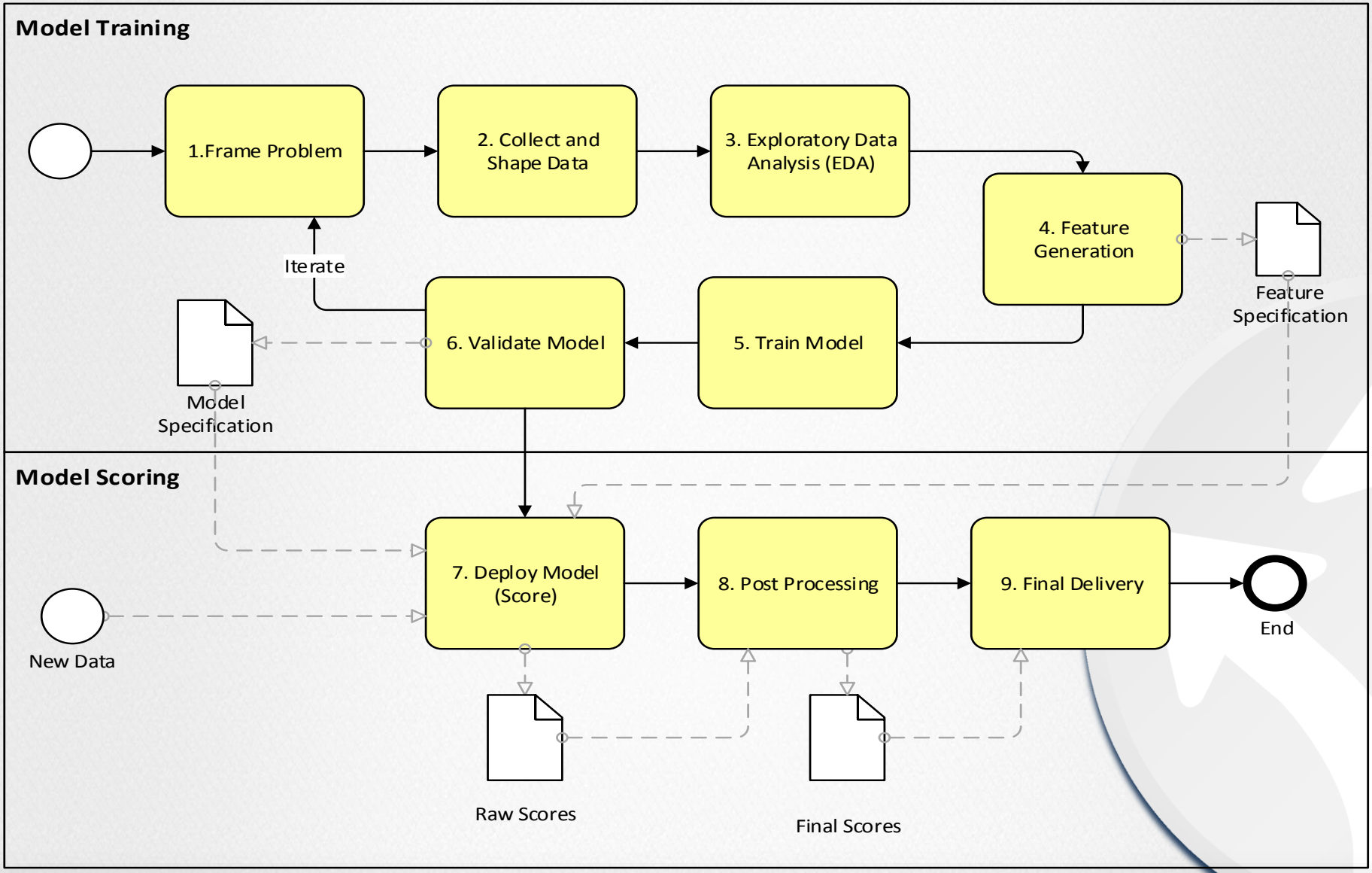
Goal Find A Function, f

- easy to evaluate
- Takes a one or more values of inputs
- yields a single output value for each input (row)

- Output, $\hat{\mathbf{y}}$, should be "close" to observed values, $\mathbf{y}$:

$$\hat{\mathbf{y}} \sim \mathbf{y}$$

Expectations: Process



REVIEW EXERCISES

Goal Find A Function, f

- easy to evaluate
- Takes a one or more values of inputs
- yields a single output value for each input (row)

- Output, $\hat{\boldsymbol{y}}$, should be "close" to observed values, $\mathbf{y}$:

$$\hat{\boldsymbol{y}} \sim \boldsymbol{y}$$

QUESTIONS:

- What do we mean by “close”?
- What functions are available to be used?

- How do we find one? The best one?

EXERCISE 1: FUEL ECONOMY

Our Model

Naïve Model

$$\hat{y} = mean(y)$$

Our Model, a linear model:

$$\hat{y} = \beta_0 + \beta_1 x_1$$

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**)

SEARCH

Find the parameters minimize that minimize the loss function …

SOLVE:

$$argmin_{\beta}\, L(\boldsymbol{y}, \hat{\boldsymbol{y}})$$

- Direct Solution (special case)

  linear regression $\hat{y} = \beta_0 + \boldsymbol{x}^T \boldsymbol{\beta}$

- Recursive Goal Seeking

LM / MODEL FORMULA

LINEAR REGRESSION MODEL

→ Abstract to multiple dimensions

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

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

Mathy-r !!!

APPENDIX