Decision Trees

Practical Machine Learning (with R)

UC Berkeley

Fall 2015

Topics

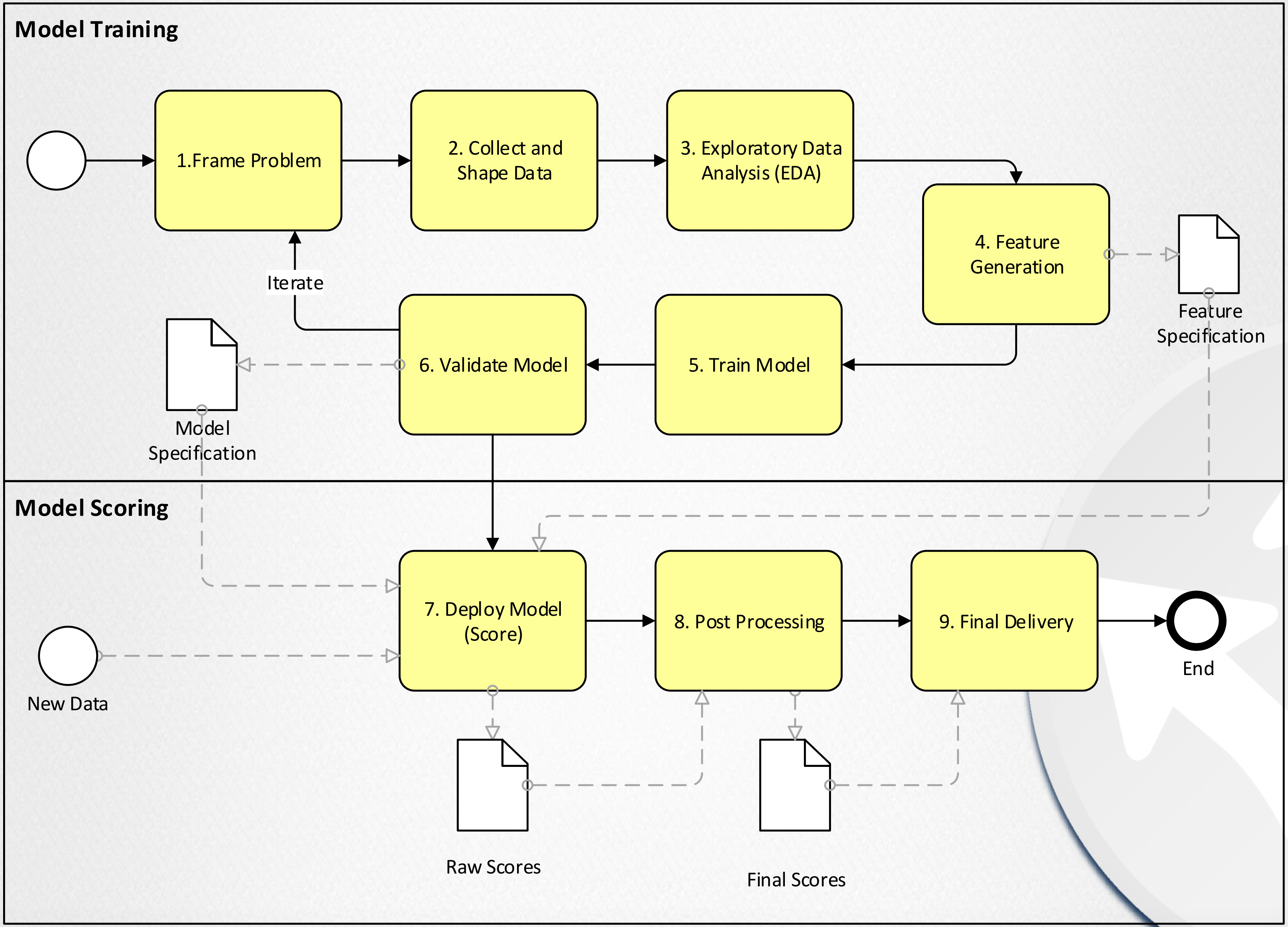
- Administrativa
  - Github
    - Reorganization → one repository
    - Please put
- Review and Expectations
- In-Class Assignment
- New Topics

REVIEW AND EXPECTATIONS

Review and Expectation

- Github
- Transformations
- Logistic Regression ~ Linear Regression with the **logit** link function
- `glm( …, family=" ")`

Comprehensive ML Process



Goal:

BUILD UP A TOOL BOX OF SKILLS

Worked Example: Boston Housing Transformations and Stepwise

NEW TOPICS

RMARKDOWN:DEMOSTRATION

MODEL PERFORMANCE

Model Performance

- Determine relevant metric, e.g. **RMSE**, **FPR**
- Calculate statistic ("metric")

- On training data
  Training or *apparent* performance → bias → over-fitting

Need unbiased estimate for calculating performance

Resampling

- Best Solution: Data Splitting
  Split data into training and test data
  - Easy to interpret defend
  - Requires data not be consumed by model
  - Computationally easy
  - Is generally not (by itself) the most accurate → no confidence

- Resampling Strategies
  - Repeated Splitting
  - K-Fold Cross Validation
  - Bootstrap

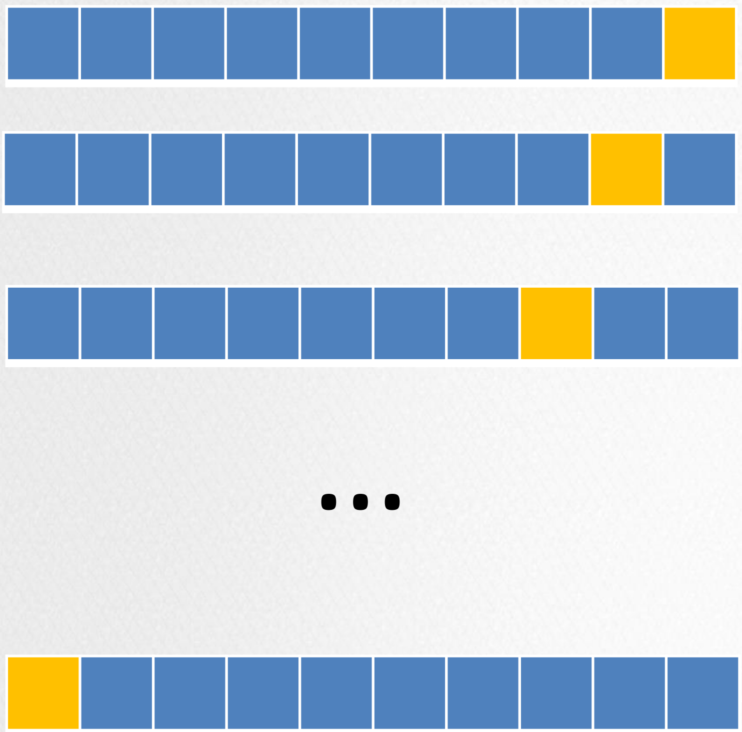
Repeated Splitting

AKA Monte Carlo Splitting

- Split data 75%-25%
  - Fit Model
  - Calculate Metric
  - Repeat with Different Split(30+ times)
- Calculate Metric

$$Metric = AVG_i(metric)$$

10-Fold Cross Validation



LOOCV : K→n

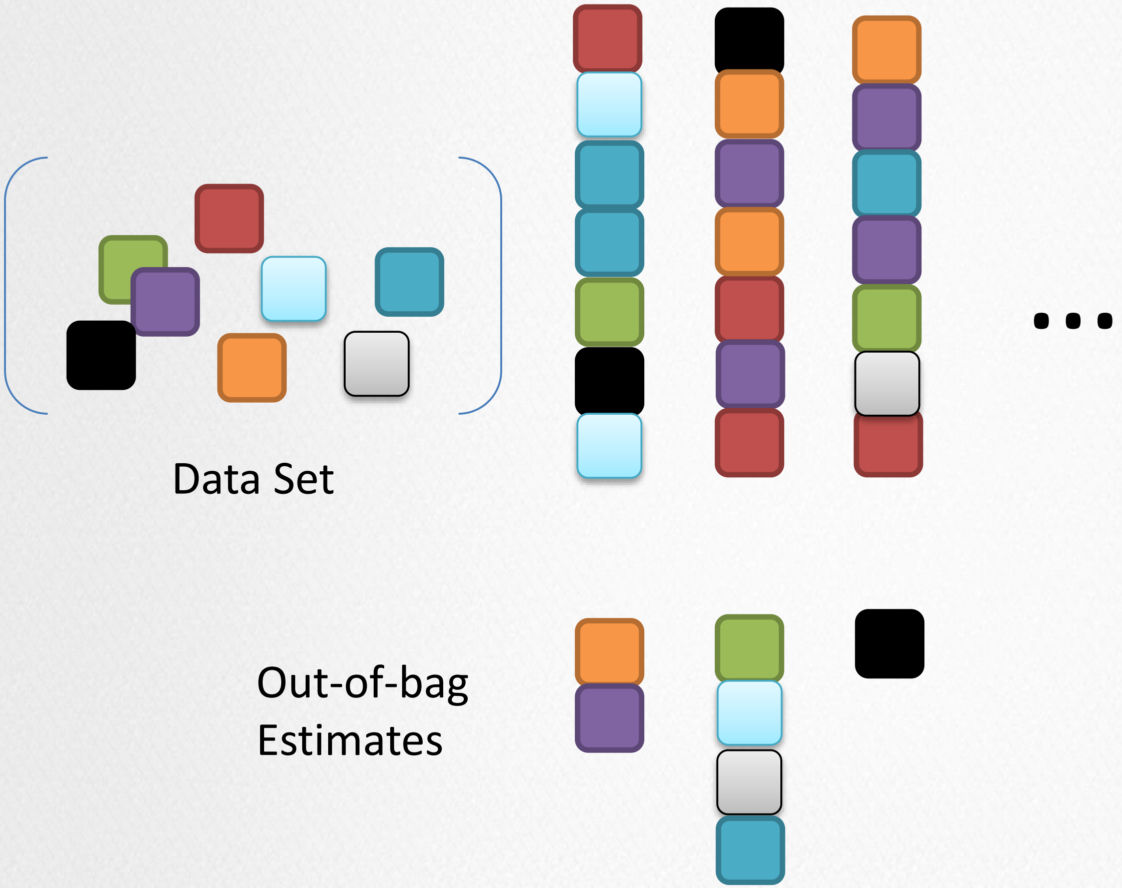
- Split the data set into 10 equal sized samples.
- Leave one sample out (fold)
  - Fit the model
  - calculate the metric on the fold
  - Repeat choosing another sample until
- Calculate Metric

$$Metric = AVG_i(metric)$$

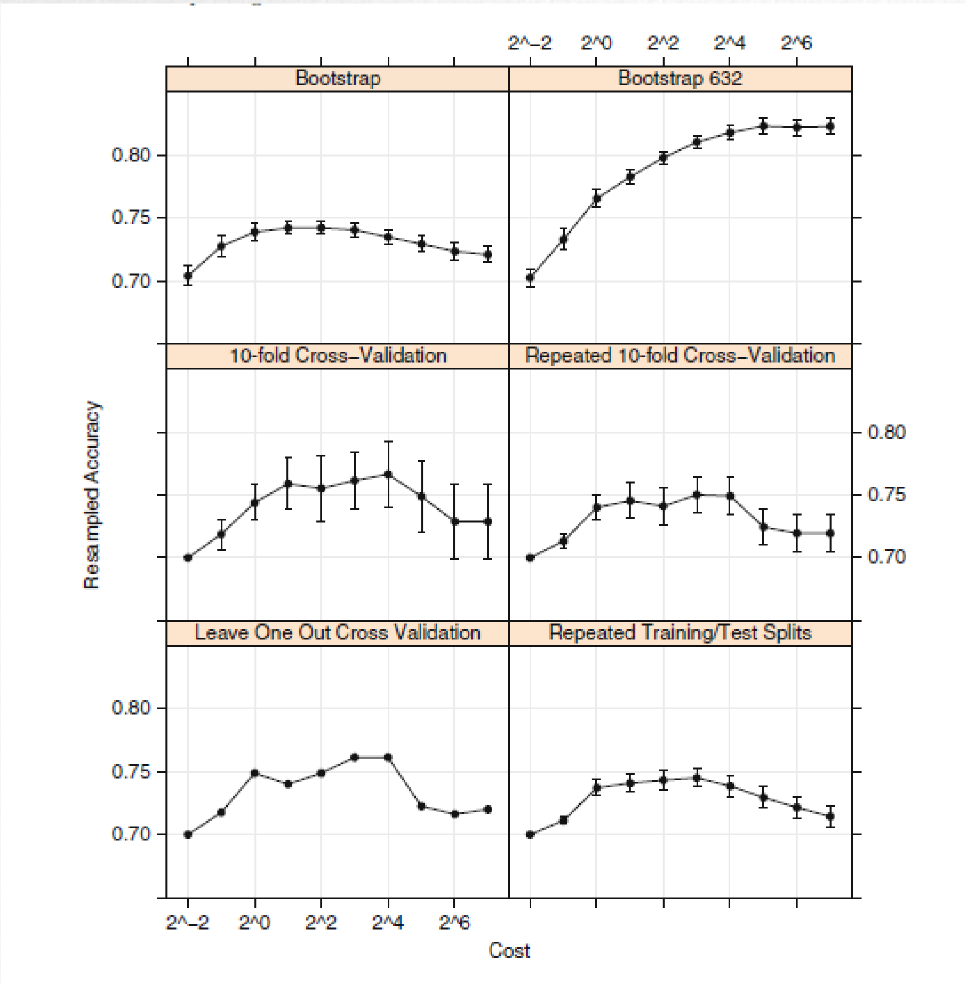
- 5 or 10-fold common

Bootstrap

- "Sampling with Replacement"



Which Is Best?



- There isn't one.

K-fold cross validation

Higher Variance
Lower Bias

Bootstrap

Lower Variance
Higher Bias

CALCULATING PERFORMANCE IS *NOT* THE SAME AS FITTING THE MODEL

EXERCISE: LECTURES/04-DECISION-TREES/RESAMPLING.RMD

Model Formula (higher Order Terms)

- Model Formula …

CLASSIFICATION PERFORMANCE

METRICS FOR BI-NOMIAL CLASSIFICATION

Total Population

Total Population

Observed	
Positive	**Negative**

prevalence

Total Population

Predicted	Positive
	Negative

Total Population

Observed

Positive

Negative

Predicted

Positive

Negative

Accuracy

Error Rate

or Misclassification Rate

Total Population		Observed	
		Positive	Negative
Predicted	Positive		
	Negative		

- https://en.wikipedia.org/wiki/Sensitivity_and_specificity

Total Population		Observed: Positive	Observed: Negative
Predicted	Positive	True Positive	False Positive (Type I Error)
Predicted	Negative	False Negative (Type II Error)	True Negative

Alternatives: Norm by Observed

Total Population		Observed: Positive	Observed: Negative
Predicted	Positive	True Positive Rate (TPR), **Sensitivity**, Recall $\frac{\text{True Positives}}{\textbf{Observed Positives}}$	False Positive Rate (FPR), Fall-Out $\frac{\text{False Positives}}{\textbf{Observed Negatives}}$
Predicted	Negative	False Neg. Rate (FNR), Miss rate $\frac{\text{False Negatives}}{\textbf{Observed Positives}}$	True Neg. Rate (TNR), **Specificity** (SPC) $\frac{\text{True Negatives}}{\textbf{Observed Negatives}}$

Alternatives: Norm by Predicted

Total Population		Observed	
		Positive	Negative
Predicted	**Positive**	Pos. Predictive Value (PPV), **Precision** True Positives / **Predicted Positives**	False Discovery Rate (FDR) False Positives / **Predicted Positives**
	Negative	False Omission Rate(FOR) False Negatives / **Predicted Negatives**	Negative Predictive Value (NPV) True Negatives / **Predicted Negatives**

- https://en.wikipedia.org/wiki/Sensitivity_and_specificity

MORE FUN …

https://en.wikipedia.org/wiki/Sensitivity_and_specificity

EXERCISE BINOMIAL METRICS: SKIN-NON SKIN

Even More Complication

- Not all errors need count "*equivocal zone*" or "*intermediate zone*"

- *Prevalent when the model has three choices, e.g. A or B or Nothing.*

MUTLINOMIAL CLASSIFICATION

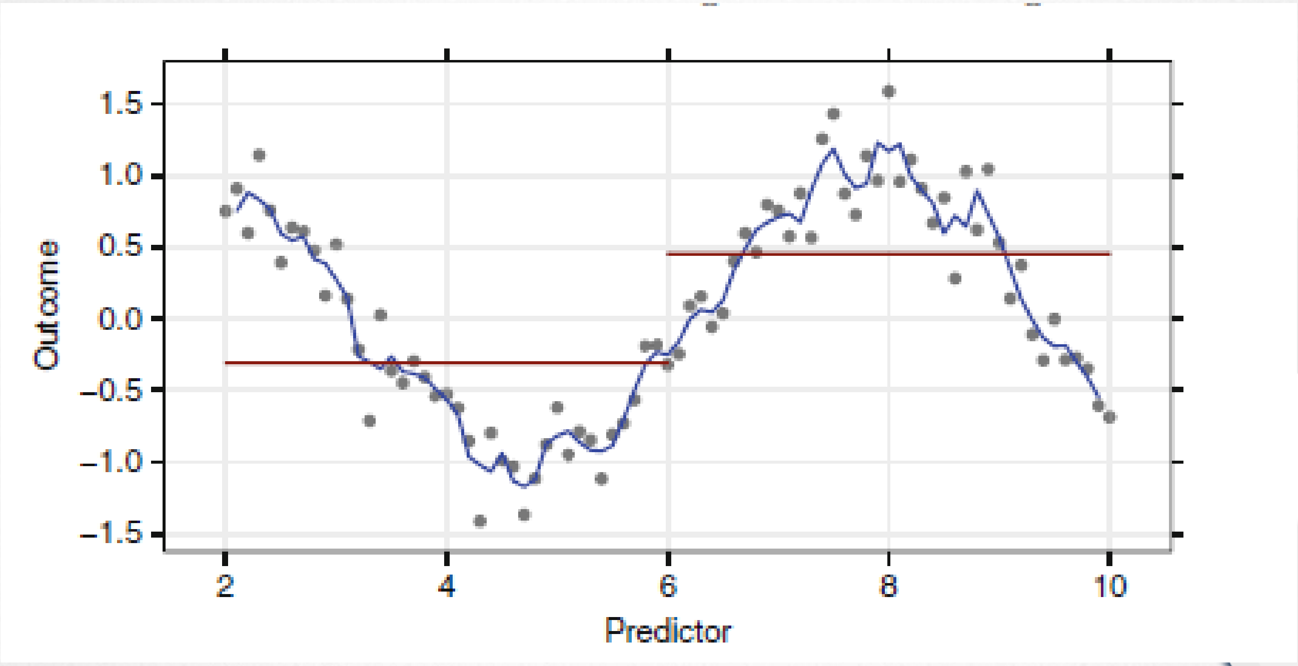
Terms

- Kappa Statistic,
- S-Statistics, F-Statistic

Multiclass Classification with Logistic Regression

Bias Variance Trade-OFF

$$E[MSE] = \sigma^2 + (model\ bias)^2 + model\ variance$$



Process

Example of ML algorithm(s)

- Spam Filter
- handwriting recognition (svm)
- Traffic engineering (lights)
- Weather prediction
- Sentiment analysis (social media)
- Netflix Recommender
- Fraud detection (Visa)
- Imaging processing
- (network) Intrution detection
- Self-driving cars

LOGISTIC REGRESSION

DECISION TREES

APPENDIX

Comparison of Models (Chart)

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)

		True condition			
	Total population	Condition positive	Condition negative	Prevalence = Σ Condition positive/Σ Total population	
Predicted condition	Predicted condition positive	**True positive**	**False positive** (Type I error)	Positive predictive value (PPV), Precision = Σ True positive/Σ Test outcome positive	False discovery rate (FDR) = Σ False positive/Σ Test outcome positive
	Predicted condition negative	**False negative** (Type II error)	**True negative**	False omission rate (FOR) = Σ False negative/Σ Test outcome negative	Negative predictive value (NPV) = Σ True negative/Σ Test outcome negative
	Accuracy (ACC) = Σ True positive + Σ True negative/Σ Total population	True positive rate (TPR), Sensitivity, Recall = Σ True positive/Σ Condition positive	False positive rate (FPR), Fall-out = Σ False positive/Σ Condition negative	Positive likelihood ratio (LR+) = TPR/FPR	Diagnostic odds ratio (DOR) = LR+/LR−
		False negative rate (FNR), Miss rate = Σ False negative/Σ Condition positive	True negative rate (TNR), Specificity (SPC) = Σ True negative/Σ Condition negative	Negative likelihood ratio (LR−) = FNR/TNR	