Decision Trees

Practical Machine Learning (with R)

UC Berkeley

Spring 2015

Topics

- Administrativa
  - Role Call
  - Assignments due to github
  - Class Google Group (All joined)
- Expectations (Review)
- New Topics
  - R Meetup

REVIEW

Need a tool …

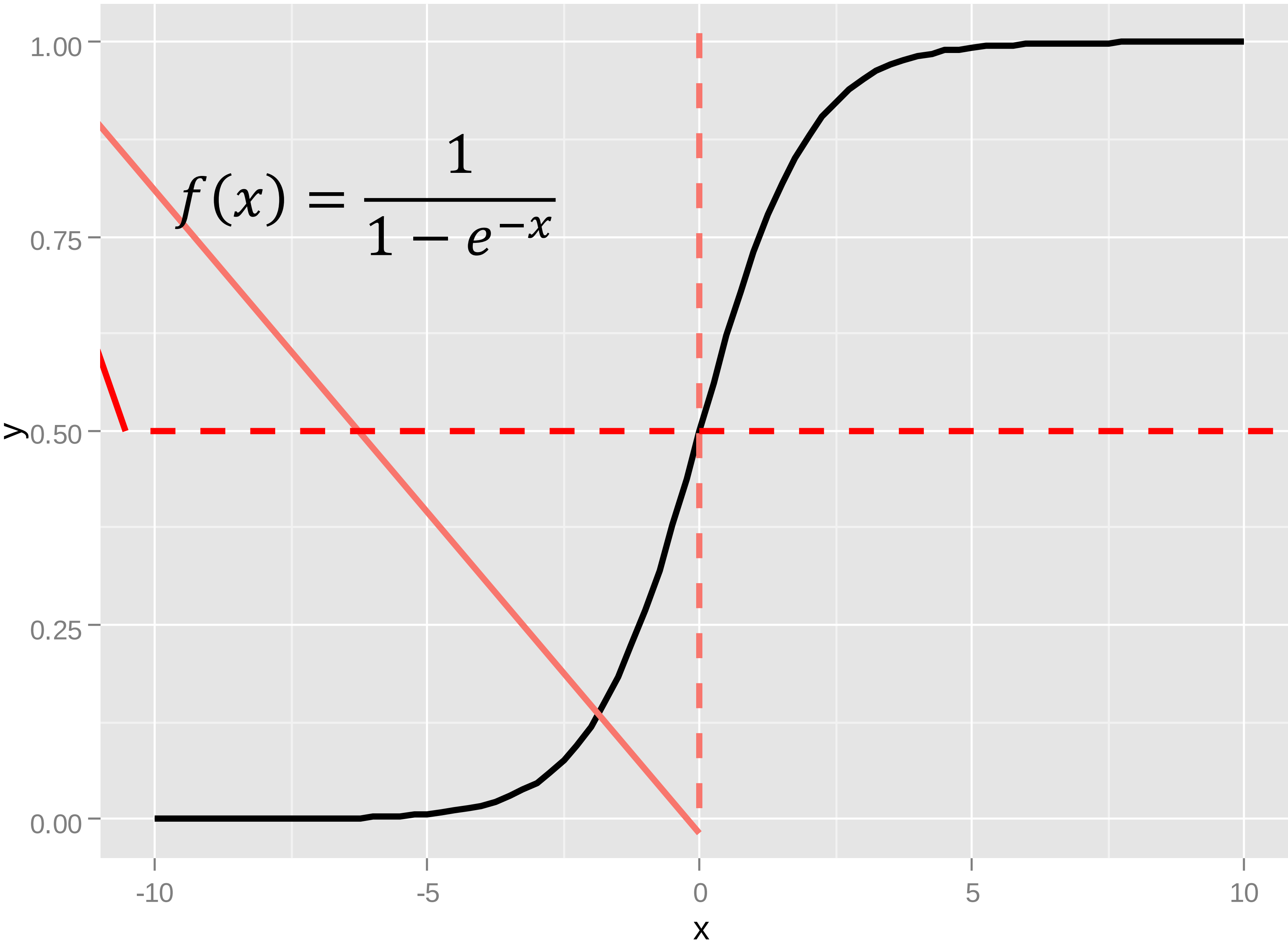
Inputs ? Outputs

`(-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:

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

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?

MODEL PERFORMANCE

Model Performance (thus far)

- Determine performance metric:
  - **RMSE (regression)**
  - **Accuracy (classification)**
- Fit Model
- Calculate statistic ("metric") on Training Data

"*training*" or "*apparent*" performance will:

- over-fit to training data
- predict very well, unbelievably well
- Not generalize to *new data*.

Cardinal Rule

DO NOT ESTIMATE PERFORMANCE ON YOUR TRAINING DATA

→ Need tool for unbiased estimate for calculating performance

Measurements and Statistics

Measurement

Quantification of a phenomena

Deterministic
≠
Stochastic

Statistic

measurement of a *stochastic* phenomena

Examples

- `mean(x)    <- x is generated by a stochastic process`
- `sd(x)`

Statistics

- "True" value unknown →uncertainty
- Uncertainty can be measured
  - Variance
  - Standard deviation
  - Confidence Interval
  - …
- Repeated measurements decrease the uncertainty

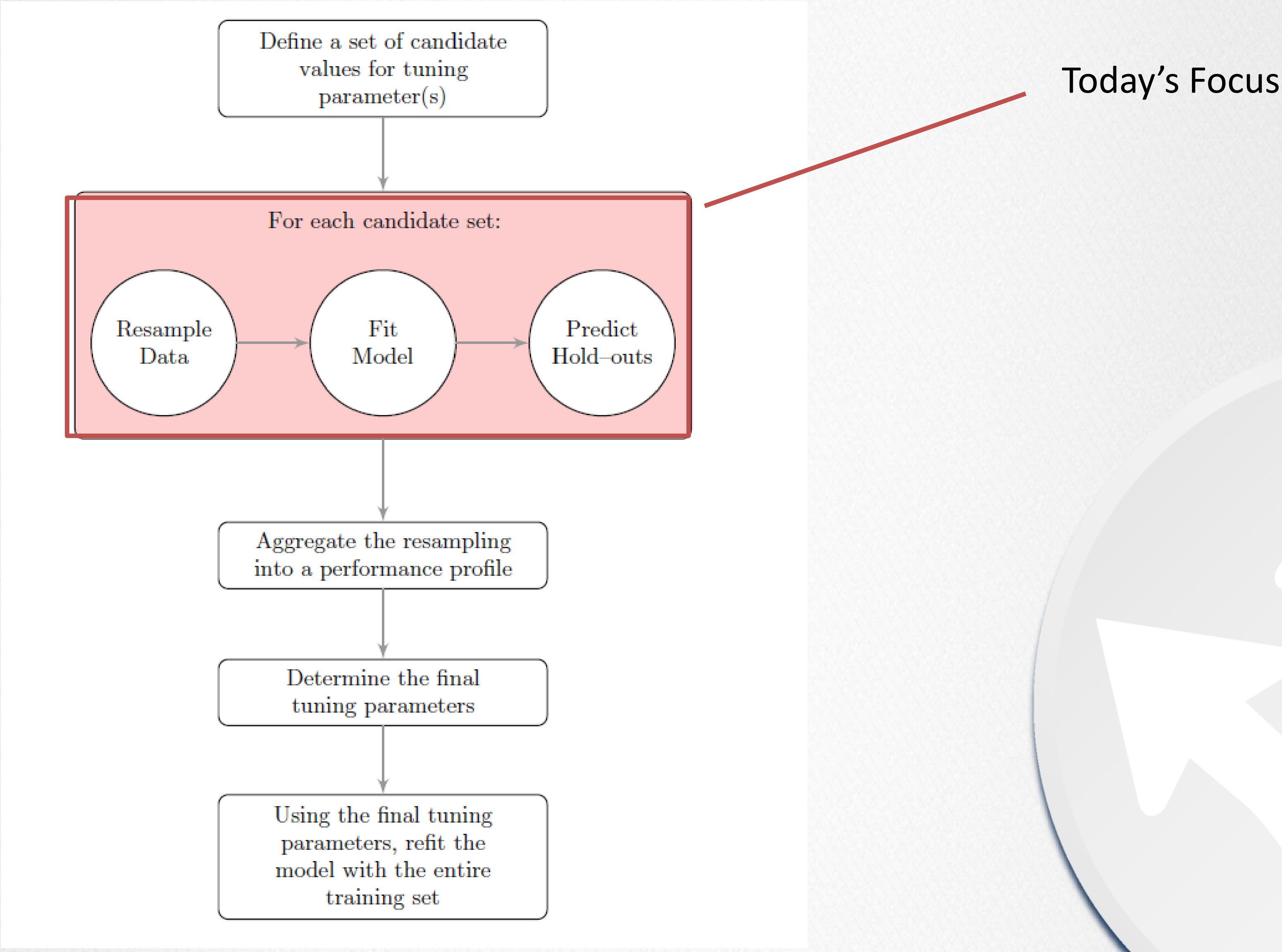
EXERCISE 1: CALCULATE `sd(mean(x))`

RESAMPLING

Kuhn benefits of resampling:

- Selection of optimal tuning parameter(s) "With so many choices how do we
- Unbiased estimate of model performance

Kuhn's Resampling Process



RESAMPLING

- Best Solution (n-permitting)
  - split data into training and test data
  - and do what Kuhn says.

Why(?)

- 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

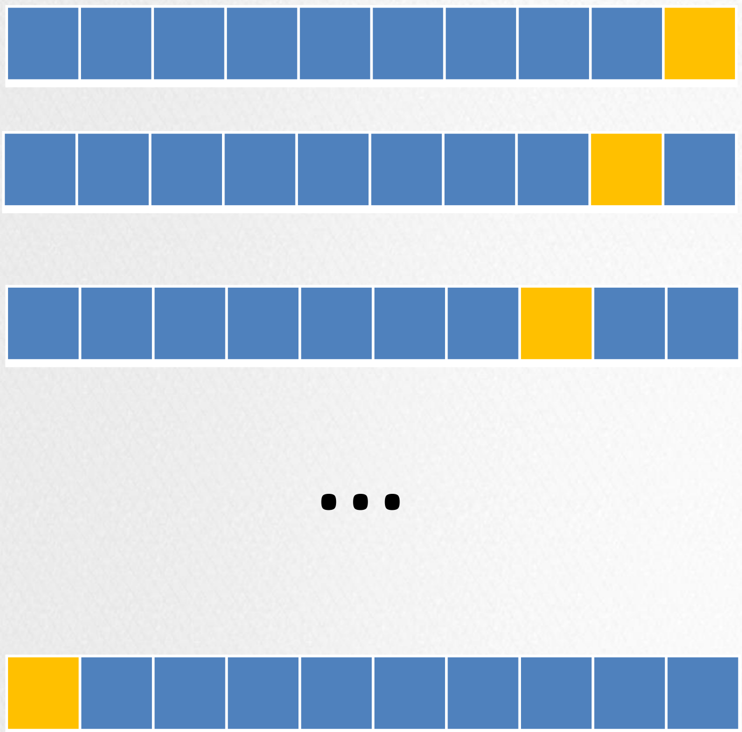
AKA Monte Carlo Splitting

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

- Calculate Metric

$$Metric = AVG_i(metric)$$

10-Fold Cross Validation



LOOCV : K→n

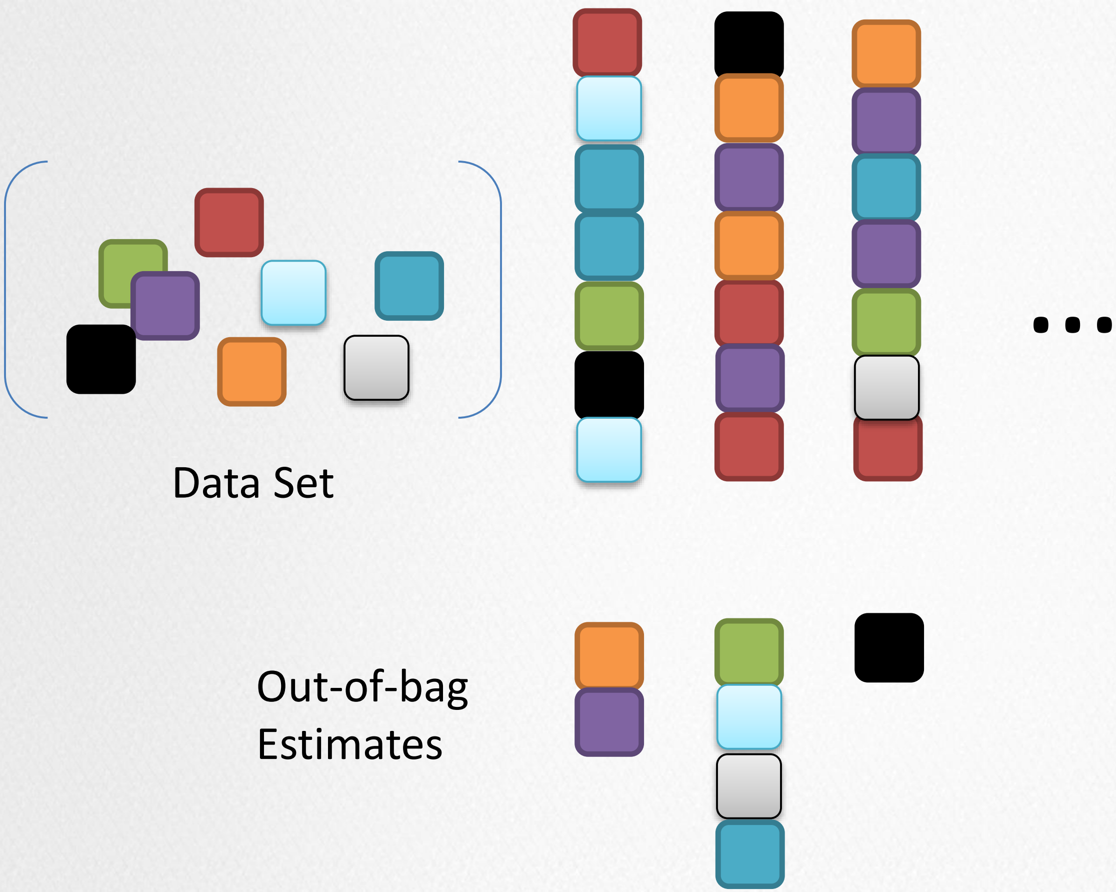
- Split the data set into 10 equal sized samples.
- Leave one sample out (fold)
  - Fit the model
  - calculate the metric on the fold
  - Repeat choosing another sample until done
- Calculate Metric

$$Metric = AVG_i(metric)$$

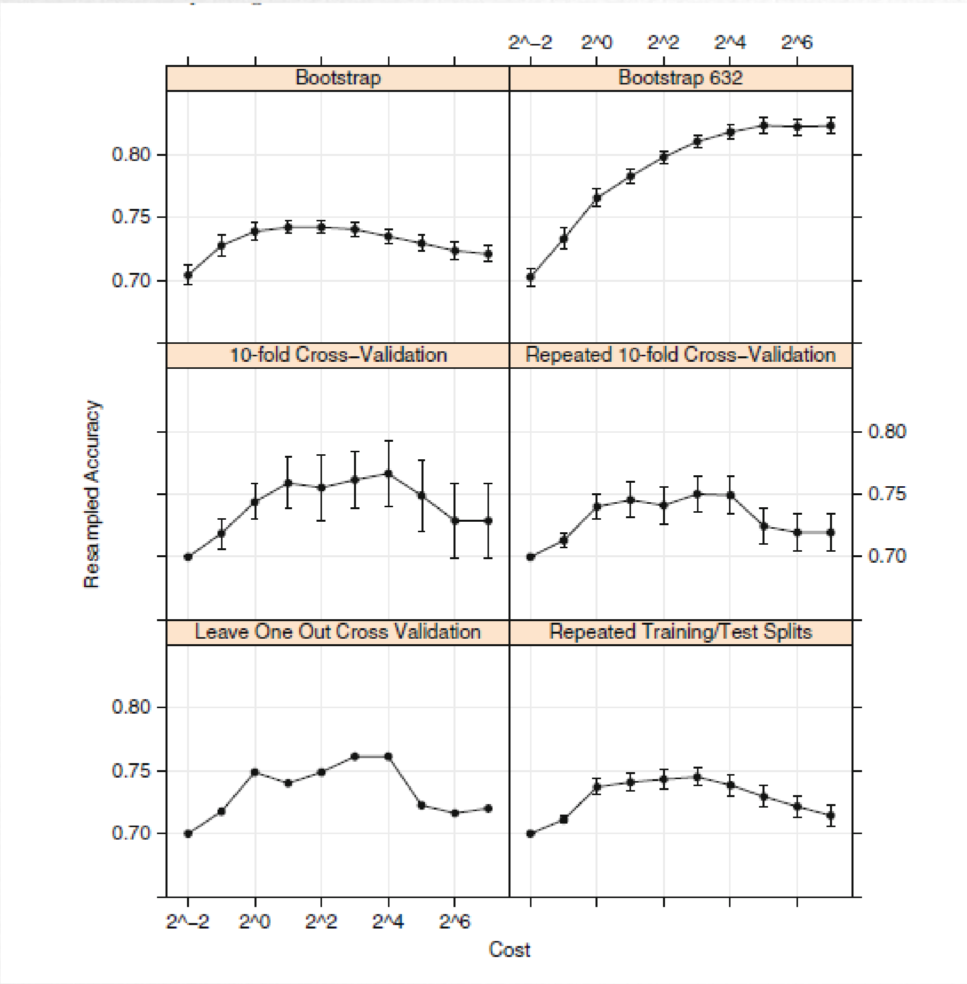
- 5 or 10-fold common

Bootstrap

- "Sampling with Replacement"



Which Is Best?



- There isn't one.

K-fold cross validation

Higher Variance
Lower Bias

Bootstrap

Lower Variance
Higher Bias

MODEL PERFORMANCE IS <u>*NOT*</u> *TRAINING* PERFORMANCE

CLASSIFICATION PERFORMANCE

METRICS FOR BI-NOMIAL CLASSIFICATION

Total Population

Total Population

Observed	
Positive	Negative

Total Population

Predicted

Positive

Negative

Total Population

Observed

Positive

Negative

Predicted

Positive

Negative

Accuracy

"How many did I get right"

Error Rate

or Misclassification Rate

"How many did I get wrong"

Total Population		Observed	
		Positive	Negative
Predicted	Positive		
	Negative		

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

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

Alternatives: Norm by Observed

Total Population		**Observed**	
		Positive	**Negative**
Predicted	Positive	True Positive Rate (TPR), **Sensitivity**, Recall True Positives **Observed Positives**	False Positive Rate (FPR), Fall-Out False Positives **Observed Negatives**
	Negative	False Neg. Rate (FNR), Miss rate False Negatives **Observed Positives**	True Neg. Rate (TNR), **Specificity** (SPC) True Negatives **Observed Negatives**

Alternatives: Norm by Predicted

Total Population		Observed: Positive	Observed: Negative
Predicted	Positive	Pos. Predictive Value (PPV), **Precision** $\frac{\text{True Positives}}{\textbf{Predicted Positives}}$	False Discovery Rate (FDR) $\frac{\text{False Positives}}{\textbf{Predicted Positives}}$
	Negative	False Omission Rate(FOR) $\frac{\text{False Negatives}}{\textbf{Predicted Negatives}}$	Negative Predictive Value (NPV) $\frac{\text{True Negatives}}{\textbf{Predicted Negatives}}$

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

More Fun …

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

EXERCISE: BINOMIAL METRICS

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

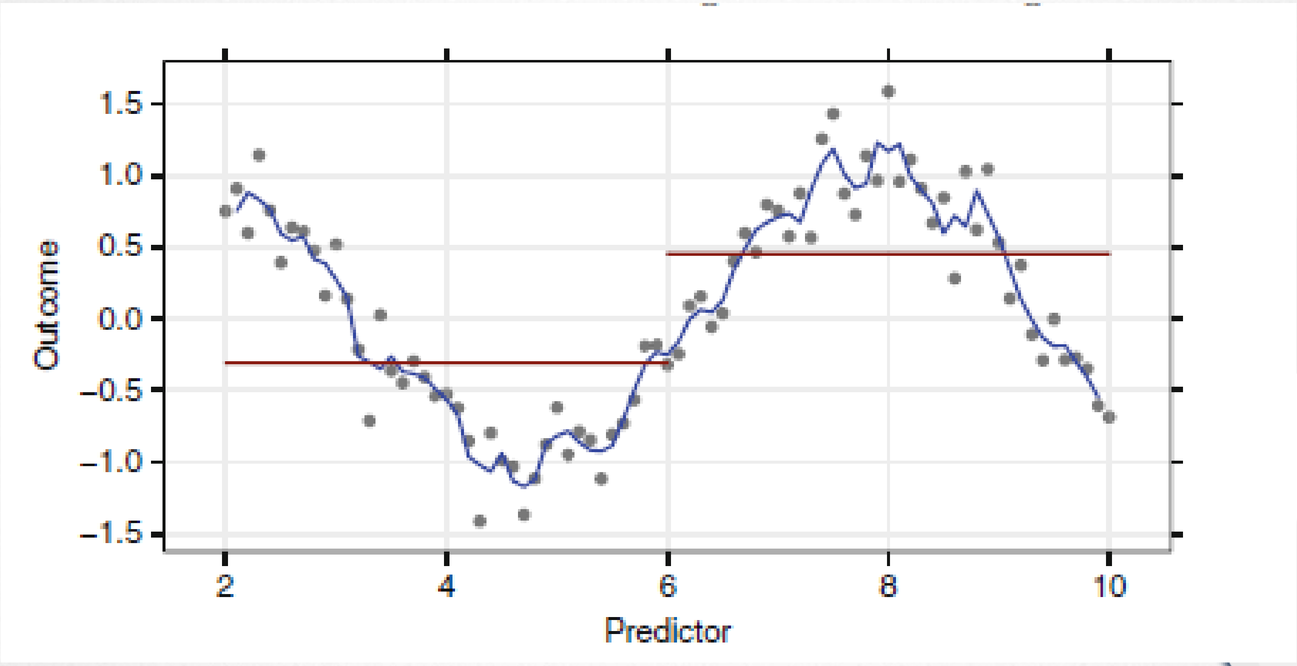
Terms

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

MULTICLASS CLASSIFICATION WITH LOGISTIC REGRESSION

Bias Variance Trade-OFF

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



		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	