Overview of ML

- Train,
- Test,
- Performance

$n = 1000$, $y = 1$ D x_1 - Age

OR x_2 - $

(target)

y - true party

$\hat{y} = \hat{f}(x_1, x_2)$

$\hat{y}$ - predicted party

$\hat{f}(\cdot)$

$y_i = \hat{f}(x_1, x_2)$

$E(y_i - \hat{f}(\cdot))^2 \Rightarrow \text{argmin}$

Γ

$$\underset{\hat{f}(x_1,x_2)}{\operatorname{argmin}} \; \mathbb{E}\,(y_i - \hat{y}_i)^2$$

$$\Rightarrow \; \mathbb{E}\,(y_i - \hat{f}(x_1, x_2))^2$$

① Training

② Performance

81 Trainings

$N = 1000$ $Y, Age, Income$

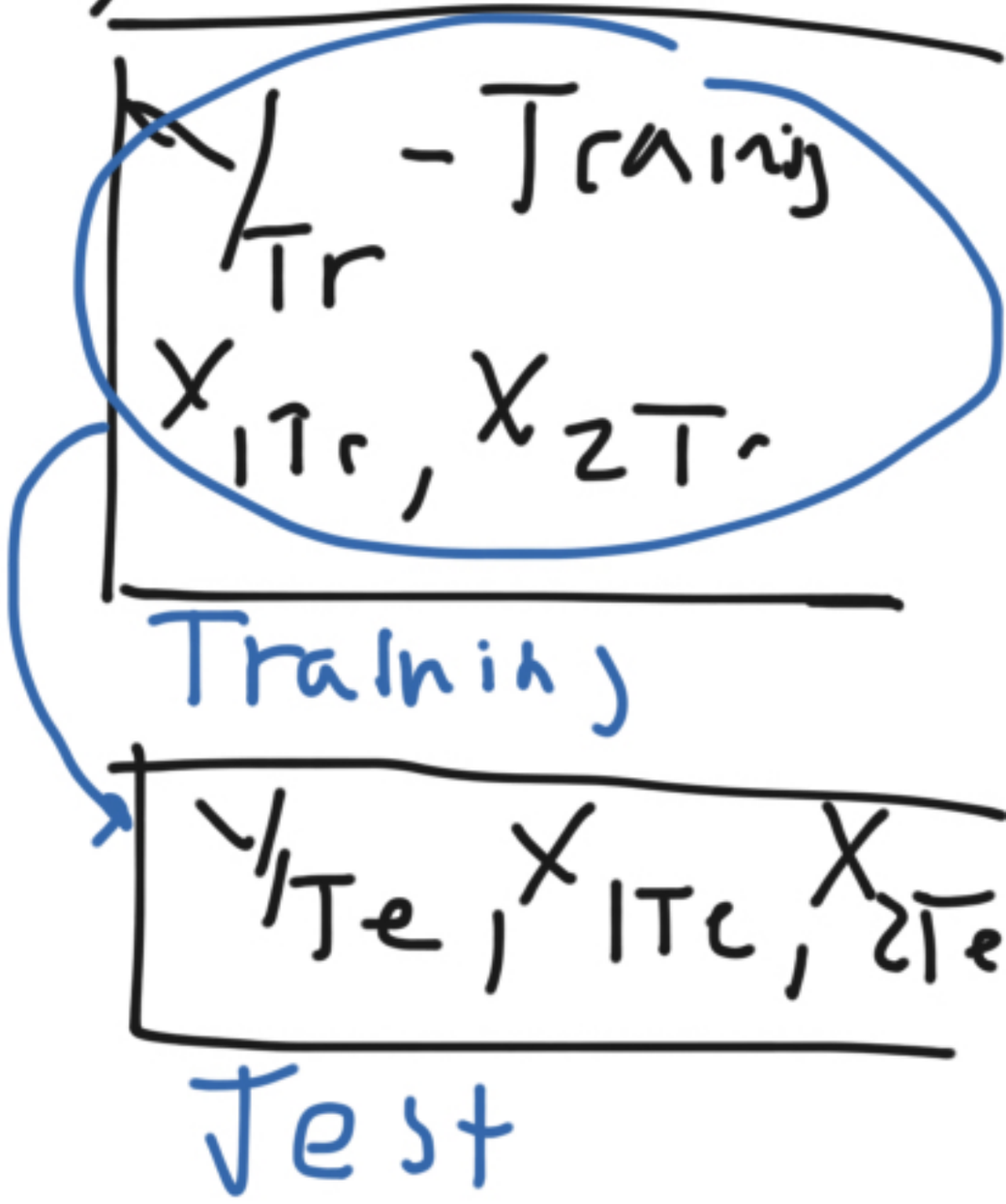
Training / Test

80% 20%

(800) (200)

$Y_{Tr} = \hat{f}(X_{1Tr}, X_{2Tr})$ ①

$\hat{Y}_{Te} = \hat{f}(X_{1Te}, X_{2Te})$



$y_{Te} - \hat{y}_{Te}$

Divide the data? 80/20

omin bias/max representativeness of the training data

Person	Party	Age	$	PersonR
1	1			3
2	0			101
3	0			120
.	1			.
.	.			.
.	.			.
.	.			.
.	.			.
.	.			.
800				
000	.			.
.	.			.
.	.			.

Trenton

Princeton

Randomly select 800 obs for training

① $\hat{y}_{Tr} = \hat{f}(X_{1Tr}, X_{2Tr})$ $N_{Tr} = 800$

② $\hat{y}_{Te} = \hat{f}(X_{1Te}, X_{2Te})$ $N_{Te} = 200$

③ Compare $\hat{y}_{Te}$, y_{Te}

Pred, True

Step 2 - Performance Metrics

→ classification - categorical labels

→ regression - continuous value (price, % pem vote)

discrete { 2: I, 1 - b, 0 - R }

discrete { 2, 1, 0 } # of people in a room

% den (0 - 100%)

Continuous

Acc = 80%

Precision = 90%

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{90}{120} = 75\%$$

$$\text{Recall (Sensitivity)} = \frac{TP}{TP+FN}$$

$$F_1 = 2 \times \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\right) = 2\left(\frac{0.90 \cdot 0.75}{0.9 + 0.75}\right)$$

$$= 0.81$$

$$\underline{\text{Specificity}} = \frac{TN}{TN+FP}$$

True negative rate

Performance (regression)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2}{N}}$$

$N = 1000$ cities

$Y =$ % voting for HRC in each city

$X_1 =$ Age (avg for each city)

$X_2 =$ $ (avg for each city)

80/20 Training / Test

Train $y_{Tr} = \alpha + \beta_1 x_{1Tr} + \beta_2 x_{2Tr}$

Test $\hat{y}_{Te} = \alpha + \beta_1 x_{1Te} + \beta_2 x_{2Te}$

y_{Te} - true % votes for HRC in test data

$\hat{y}_{Te}$ - predicted % " " " "

$$RMSE_{T_e} = \sqrt{\frac{\sum_{i=1}^{N_{T_e}} (y_i - \hat{y}_i)^2}{N_{T_e}}}$$

$N_{T_e} = 200$

Bias-Variance Tradeoff

$$E(y - \hat{y}) = \text{Var}(\hat{f}(x)) + \text{Bias}(\hat{f}(x))$$

$\hat{y} = \hat{f}(x)$

How well the algorithm predicts on other datasets

$(RMSE_{Te})$

Predicting on training data

$(RMSE_{Tr})$