K-Means Clustering

Forrest Davis

September 4, 2025

Fall 2025 COSC410A: Applied Machine Learning
Colgate University

Small Groups

G1	G2	G3	G4	G5	G6	G7
Jack	Tina	Julia	Jude	Paige	Andrew	Ashley
Matthew	Sylvia	Cathy	Jacob	Grace	Aayusha	Angie
Dilni	William	Morgan	Vincent	Luca	Hugo	

Example Data

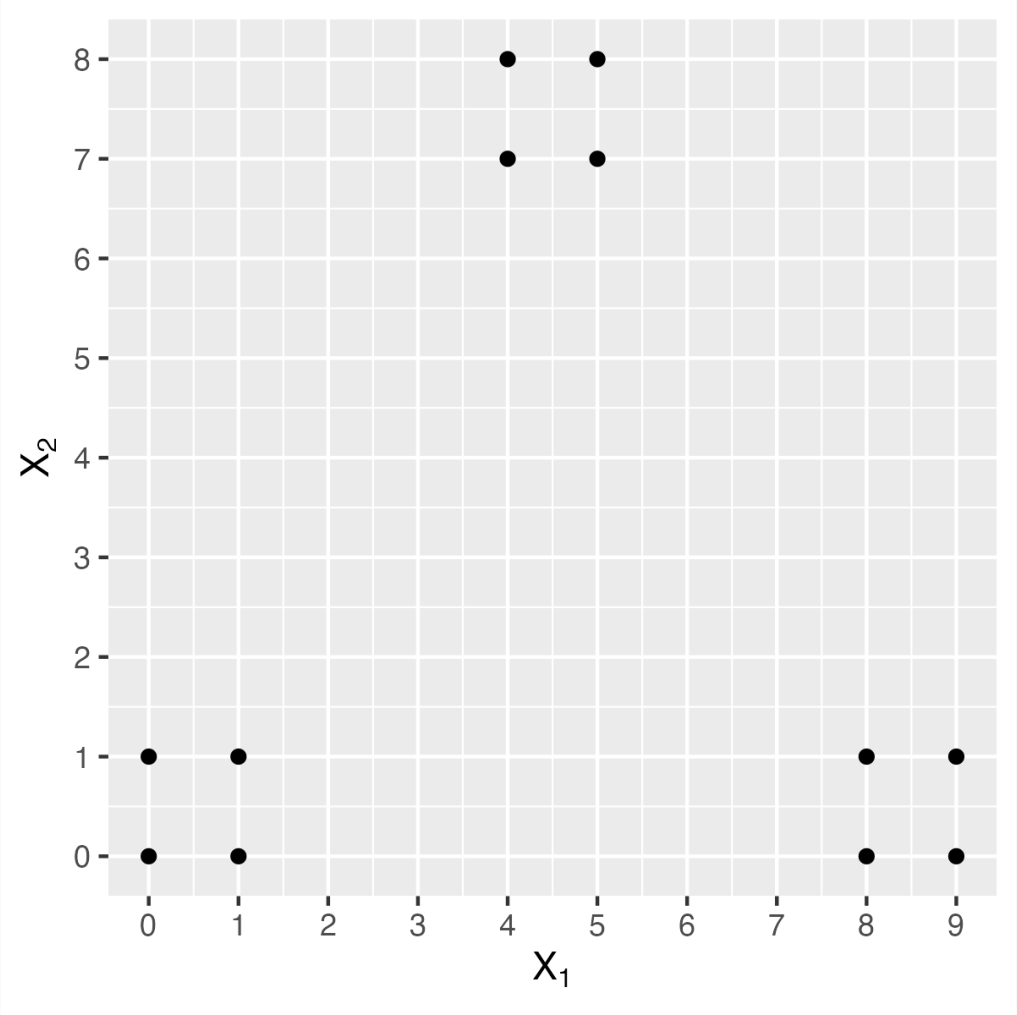
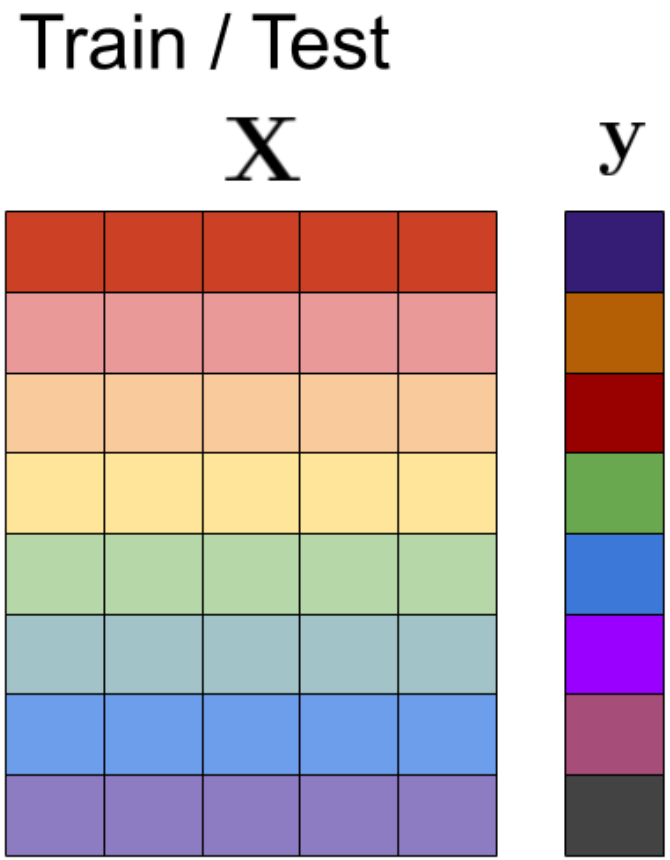
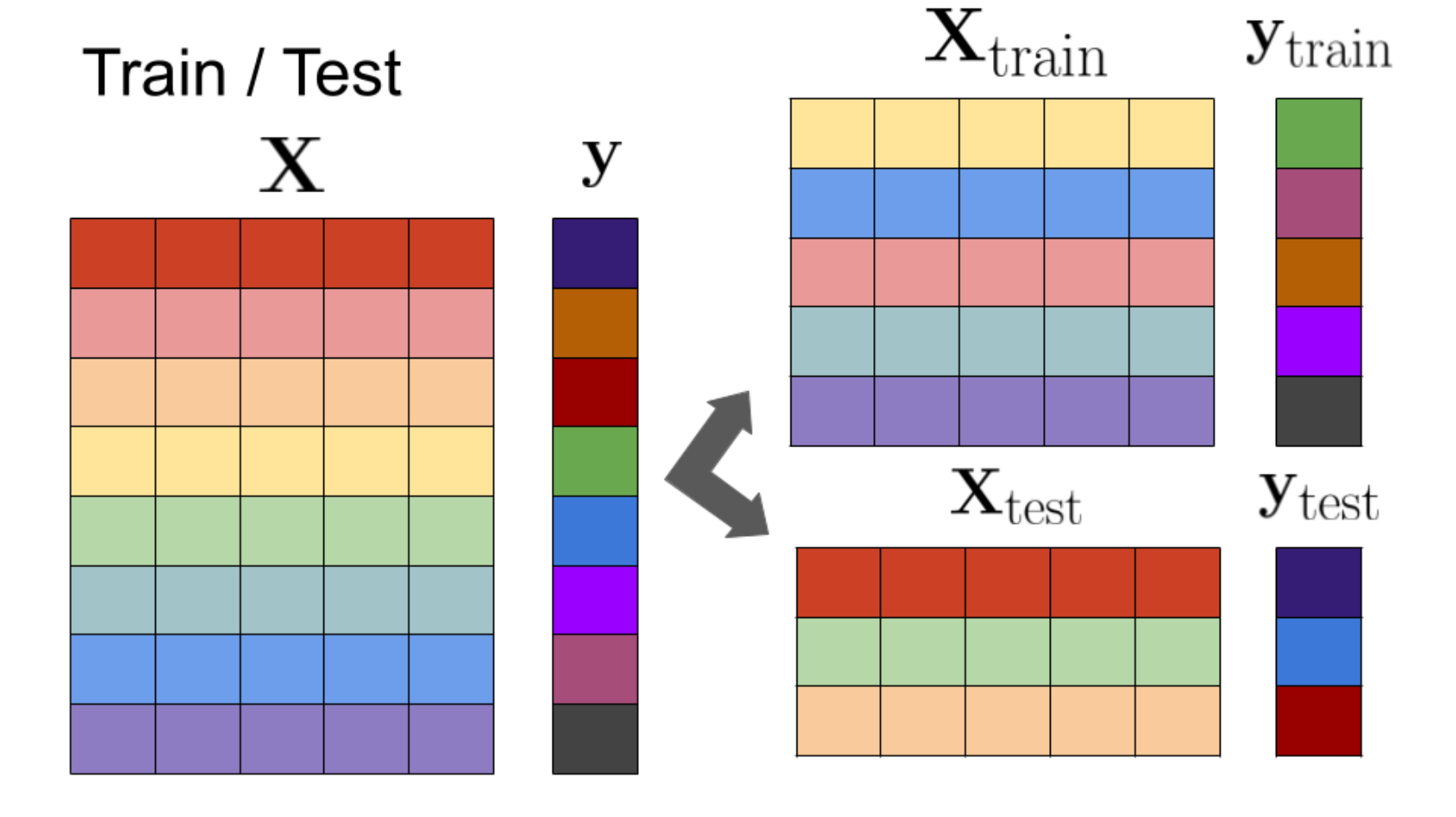


Figure 1: Sample data for group work

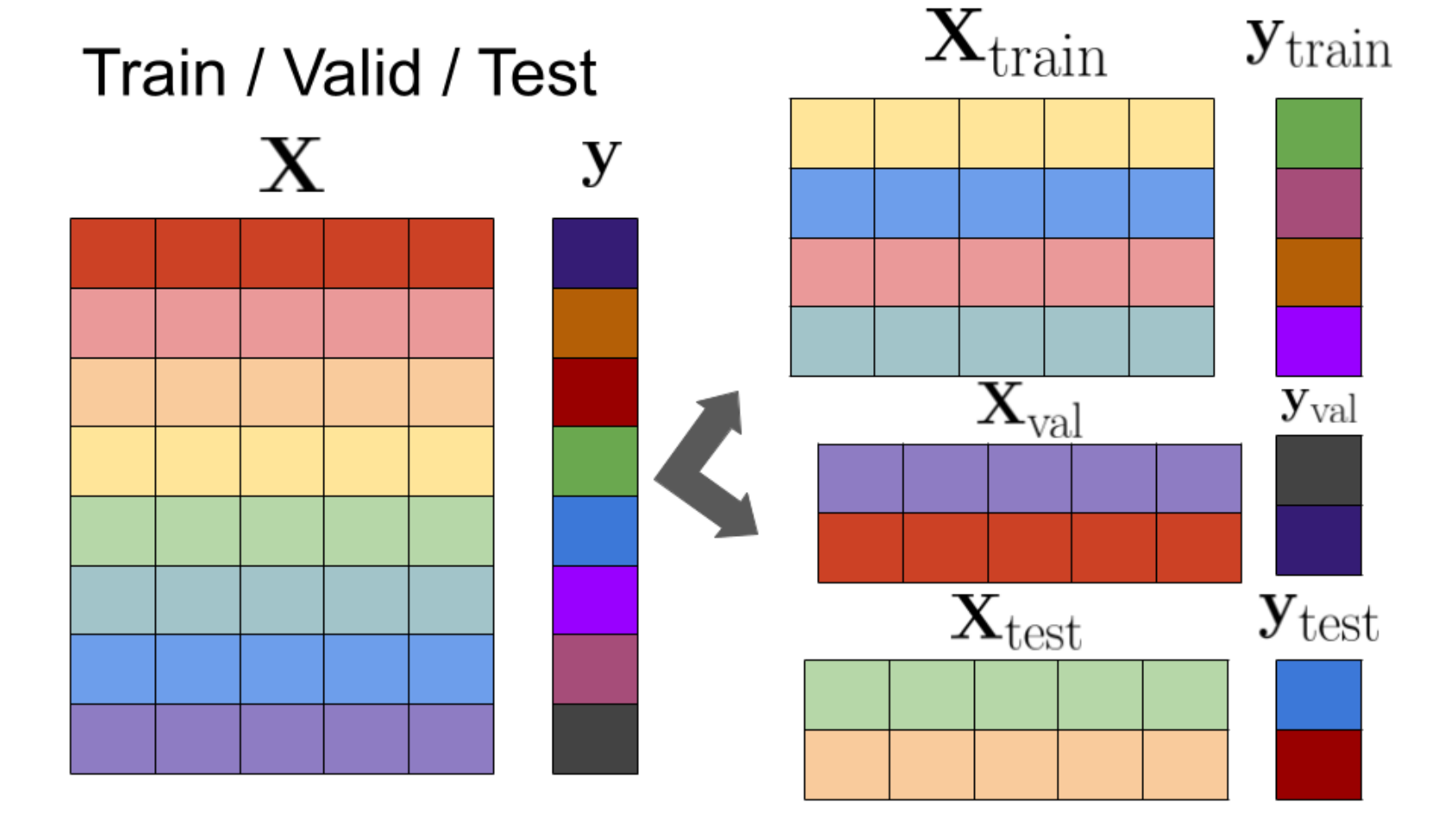
Data Splits



Data Splits


Train / Test
X
y
$\mathbf{X}_{\text{train}}$
$\mathbf{y}_{\text{train}}$
$\mathbf{X}_{\text{test}}$
$\mathbf{y}_{\text{test}}$

Data Splits


Train / Valid / Test
X
y
$\mathbf{X}_{\text{train}}$
$\mathbf{y}_{\text{train}}$
$\mathbf{X}_{\text{val}}$
$\mathbf{y}_{\text{val}}$
$\mathbf{X}_{\text{test}}$
$\mathbf{y}_{\text{test}}$

Data Splits

- Our guiding principle is that **test** data is used to assess the generalization ability of a model.
- That means **you should never train on or use your test data during model development**.
- You just use it as a final check!

K-Fold Cross Validation

- Train / validate / test split leaves much less data for training
  - Can reduce final performance (more data usually improves models)
- Solution? Cross-Validation

K-Fold Cross Validation