MACHINE LEARNING WORKFLOW

OUTLIER REMOVAL

BE SURE WHAT YOU'RE DOING!!!

STRATEGIES FOR FINDING 1-D OUTLIERS:

- 3-SIGMA RULE
- CHEBYSHEV INEQUALITY
- 3-IQR RULE

- SCALING

VARIABLES ARE MEASURED IN SPECIFIC UNITS: MEANING ERRORS IN VARIABLES ARE MEASURED IN SPECIFIC UNITS AS WELL!

- MinMax Scaler
- Standard Scaler
- Robust Scaler

CATEGORICAL ENCODING

MATH MODELS CONSUME MATH! HOW CAN WE TREAT CATEGORICAL VARIABLES?

- One Hot Encoder
- Ordinal Encoder

TRAIN/TEST/VALIDATE FRAMEWORK

WE MUST AVOID OVERFITTING!

- train_test_split()
- CROSS-VALIDATION

lower limit $= \mu - 3\sigma$
upper limit $= \mu + 3\sigma$

$\mu \pm 3\sigma \approx 88\%$ of data
$\mu \pm 5\sigma \simeq 96\%$ of data

$IQR = Q75 - Q25$
lower limit $= Q25 - IQR$
upper limit $= Q75 + IQR$

SUFFER FROM THE PRESENCE OF OUTLIERS

MORE COMPLEX, HARD TO INTERPRET

MISSING VALUES

- DROP UNECESSARY COLUMNS
- DO MISSING VALUES HAVE MEANING?

Simple Imputer ?