UCI Echocardiogram Data

Dominick Rocco

February 15, 2016

- Echocardiogram data in UCI Machine Learning Repository
  - `http://archive.ics.uci.edu/ml/datasets/Echocardiogram`
- Downloaded data, worked in git repository
  - `https://github.com/rocconnick/uci_echocardiogram`
- Two notebooks, one module
  - First notebook, data exploration and cleaning (`initial_exploration.ipynb`)
  - Data loading/cleaning functionality moved to module (`ecg_tools.py`)
  - Second notebook for regression model (`survival_duration_regression.ipynb`)

- Worked in a notebook

  `https://github.com/rocconnick/uci_echocardiogram/blob/master/initial_exploration.ipynb`

- Problems with data:
  - One bad row
  - Many missing values
  - Missing values mostly consistent with dataset description, but not completely

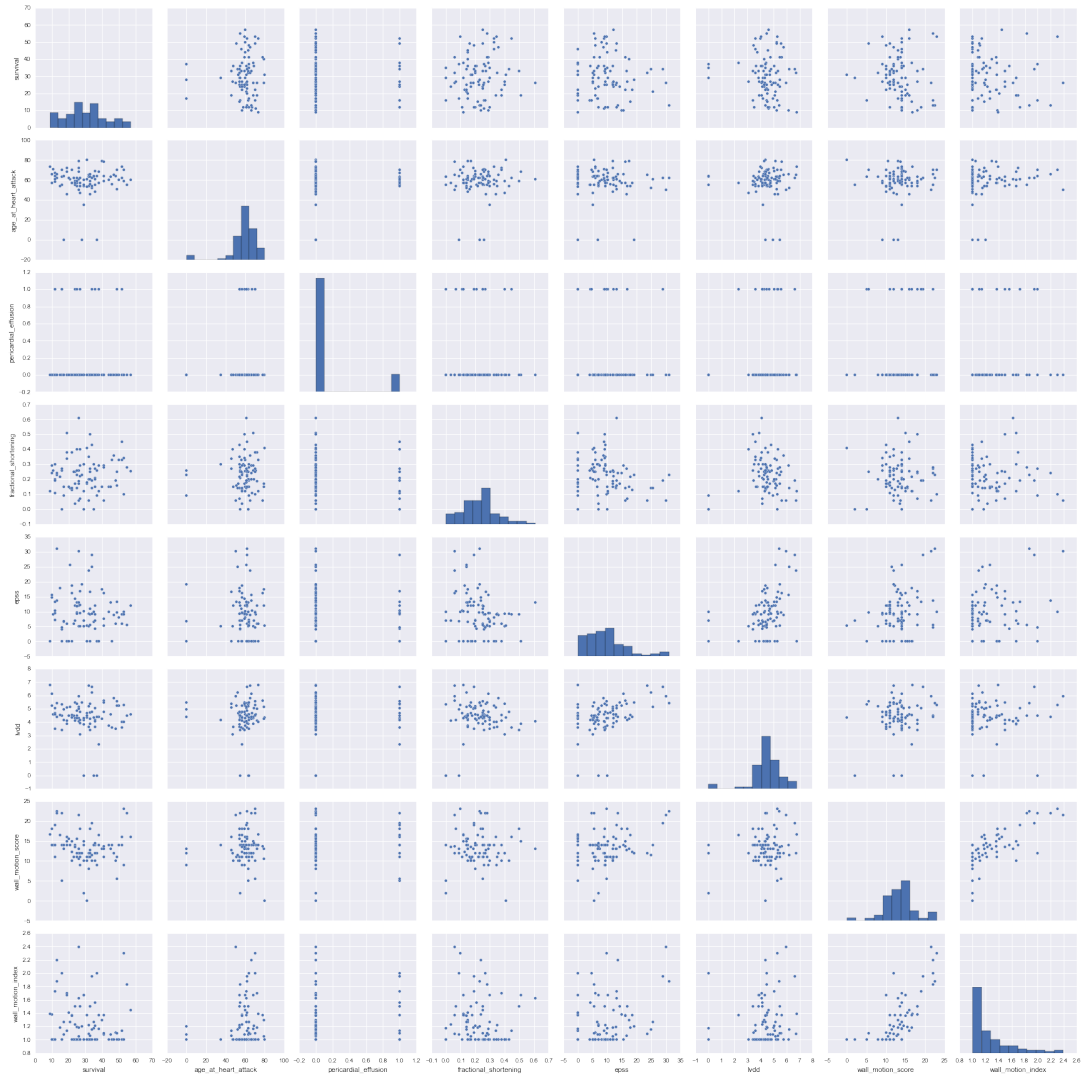
- Original dataset goal: predict which patients survive one year after heart attack
  - One quarter of patients enrolled in study less than a year ago
  - Only 4 of 92 remaining patients survived less than one year

- Instead, try to predict survival duration

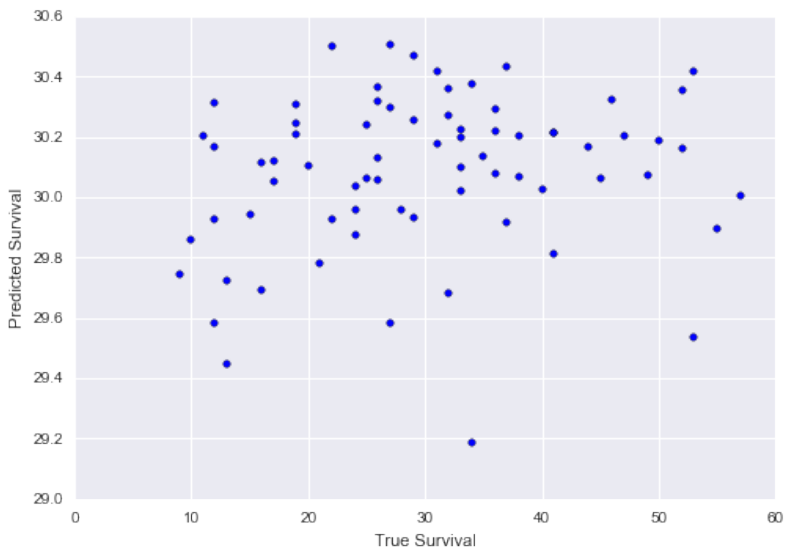
- Cleaning function ported from notebook to module

  `https://github.com/rocconnick/uci_echocardiogram/blob/master/ecg_tools.py`

- Top row: target variable vs. each other variable
- Not much correlation visible, anywhere really
- Originally imagined doing dimensionality reduction, but didn't seem worth it after seeing this
- Missing values replaced with median of distribution for that variable



- Split data 80/20 train test
- Tried several algorithms: least-squares, lasso, ridge, support vector, nearest-neighbor
- Grid search for hyperparameters, 5-fold cross validation at each grid point
  - Scikit-learn did the heavy lifting
- No algorithm successfully fit the *training* set, let alone the test set [1]



Ridge regression result on *training* set

[1] Nearest-neighbor just let every point be its own neighbor