Warm-up Exercises

The 100-Datasets Bootcamp

Galton Dataset

Regression towards the mean

Explore the Galton dataset

It was in investigating this historical dataset that Galton made the crucial observation of “regression towards the mean”

CONCEPTUAL GOAL

- Explain the relation between correlation and linear regression
- When would regression towards the mean not happen?
- When would regression towards the mean be extreme?
- Can you relate correlation to notions of dimensionality reduction geometrically?

LAB GOAL: Create a narrative from your analysis of the Galton dataset, where you mention the following:

- Why does the “regression effect” take place?
- Visualization of the data to bring out the facts pertaining to the predictors of height
- A regression model for the offspring height based on parents’ heights
- Feature-engineering for a better model
- Which regression algorithm works best
- Can you predict the father’s and mother’s height from the height of children?
- What do you observe about the parents’ heights relative to children?

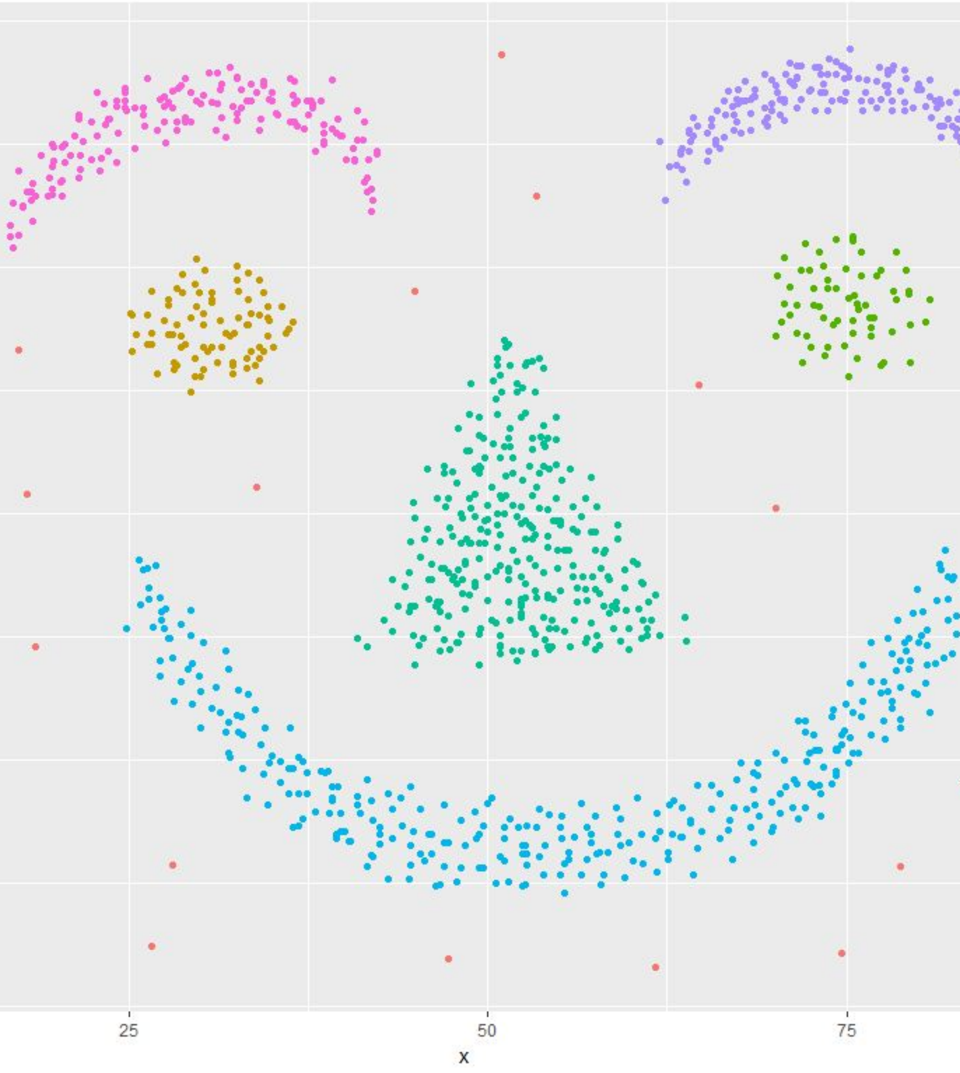
Dataset

Data is in the file: `galton/galton-families.csv`

Hint for R-users: you can directly access the data as `GaltonFamilies` in the `HistData` library.

Smiley Dataset

The deceptively simple smiley!



Smiley dataset

Recall that the deceptively simple smiley dataset created enough mischief to wipe the smile off the common clusterers. Here, the goal is classification, and to see how the different classifiers fare.

CONCEPTUAL GOAL

- Why do some classifiers work while others don't?

LAB GOAL: Create a narrative from your analysis of the Smiley dataset, where you mention the following:

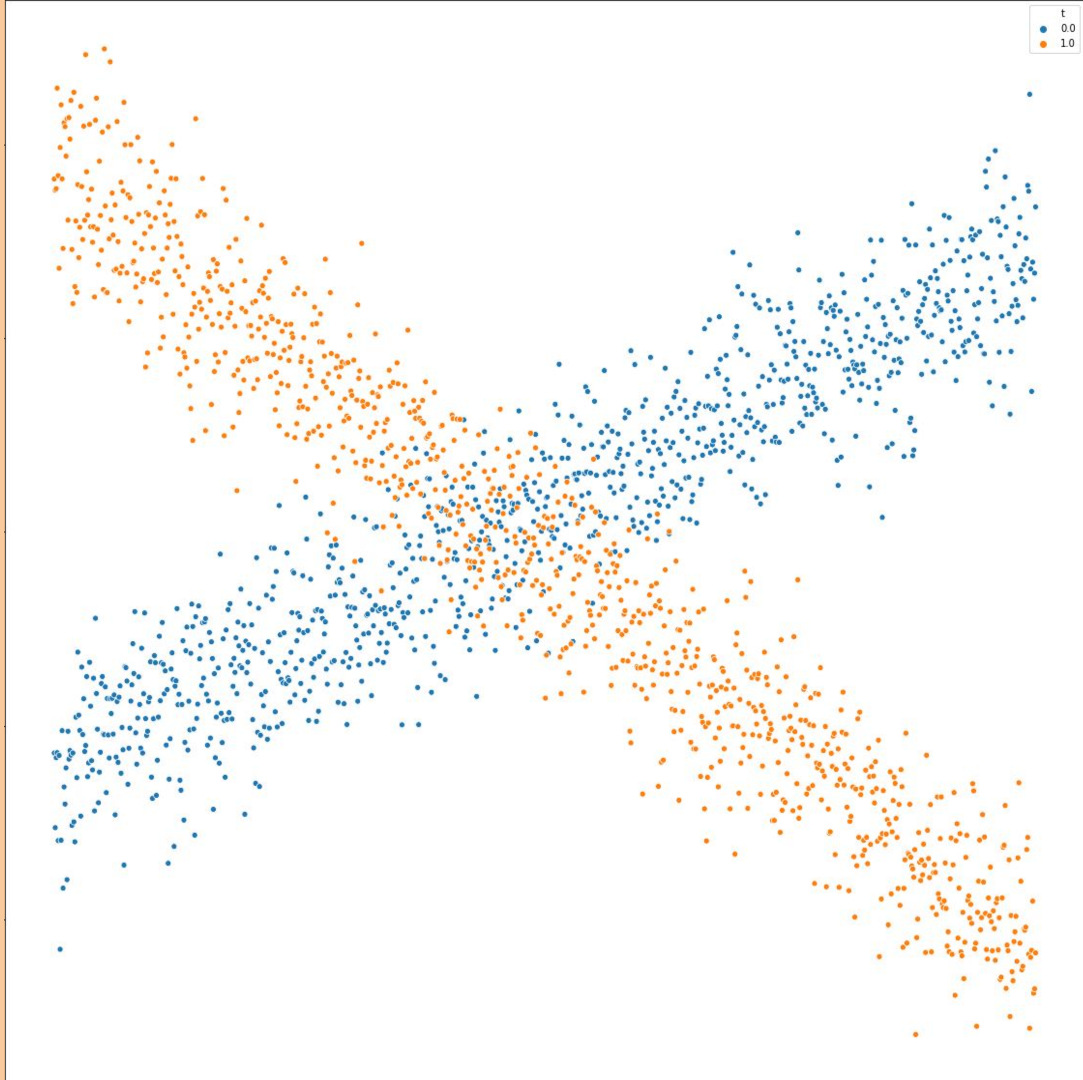
- Explore this dataset and note your observations
- Apply at least 10 different classifiers to this dataset
- Does hyperparameter tuning help much?
- Compare the results, and create justifications for the differential performances

Dataset

Data is in the file: `classifiers/smiley.csv`

Bow-Tie Dataset

Tie the bow with a good bow-tie classifier



Bowtie dataset

What is the simplest classifier you can create for this problem?

CONCEPTUAL GOAL

- Explain the virtues or necessity of using a simple classifier
- How does the performance compare to that of a SVM, RandomForest, XGBoost and a Deep-Neural network?

LAB GOAL: Create a narrative from your analysis of the Bowtie dataset, where you mention the following:

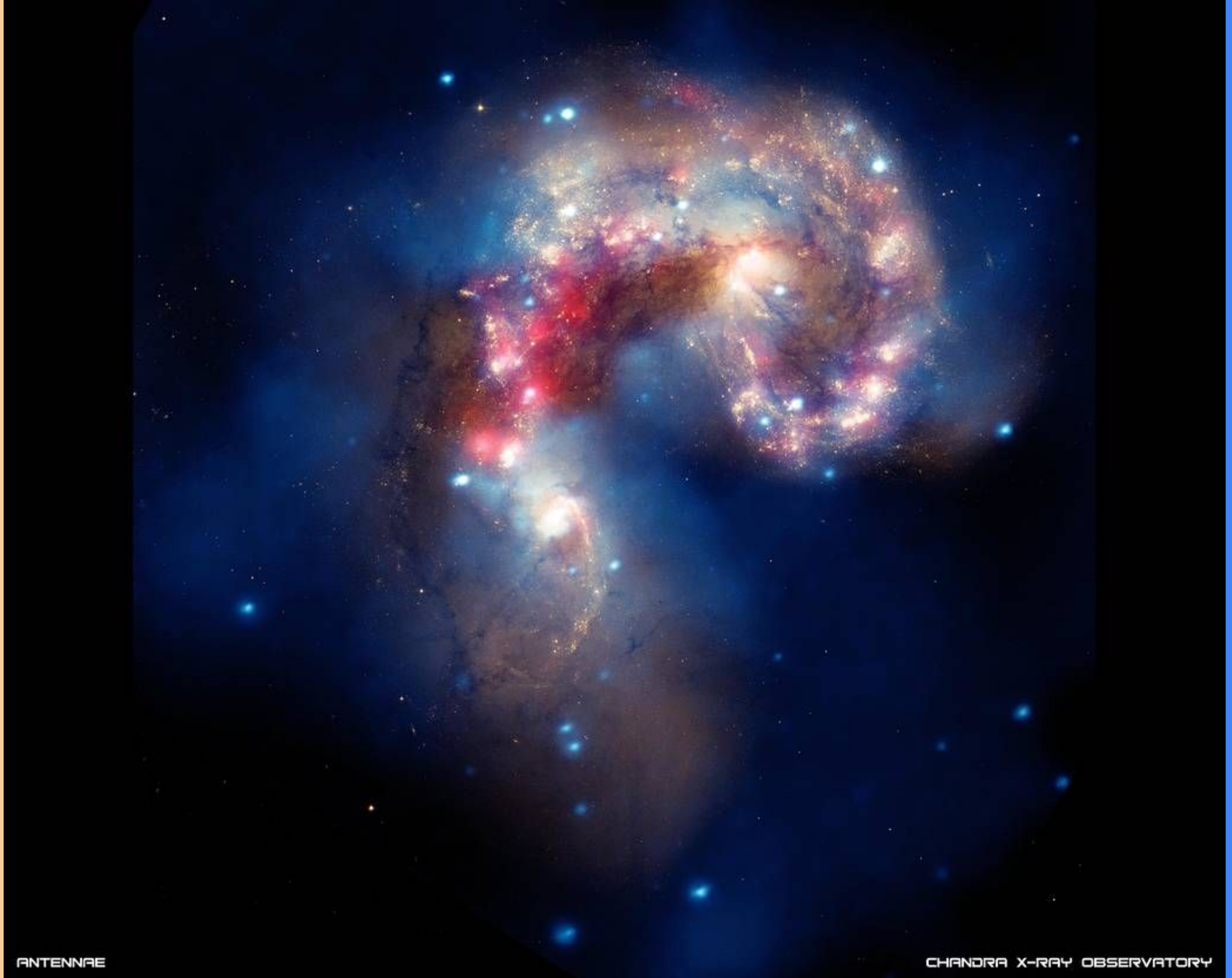
- Explore this dataset and note your observations
- Apply at least half a dozen different classifiers to this dataset
- Compare the results, and create justifications for the differential performances

Dataset

Data is in the file: `classifiers/bow-tie.csv`

Curves Dataset

Curves



Curves dataset

Solve this using only an interpretable classifier

CONCEPTUAL GOAL

- How does the performance compare to that of a SVM, RandomForest, XGBoost and a Deep-Neural network?

LAB GOAL: Create a narrative from your analysis of the Curves dataset, where you mention the following:

- Explore this dataset and note your observations
- First, model this with an easily interpretable classifier
- Apply at least half a dozen other different classifiers, that need not be interpretable
- Compare the results, and create justifications for the differential performances

Dataset

Data is in the file: `classifiers/curves.csv`

Ripples Dataset

Ripples in a pond

Ripples dataset

Solve this using only an interpretable regressor

CONCEPTUAL GOAL

- How does the performance compare to that of a SVM, RandomForest, XGBoost and a Deep-Neural network?

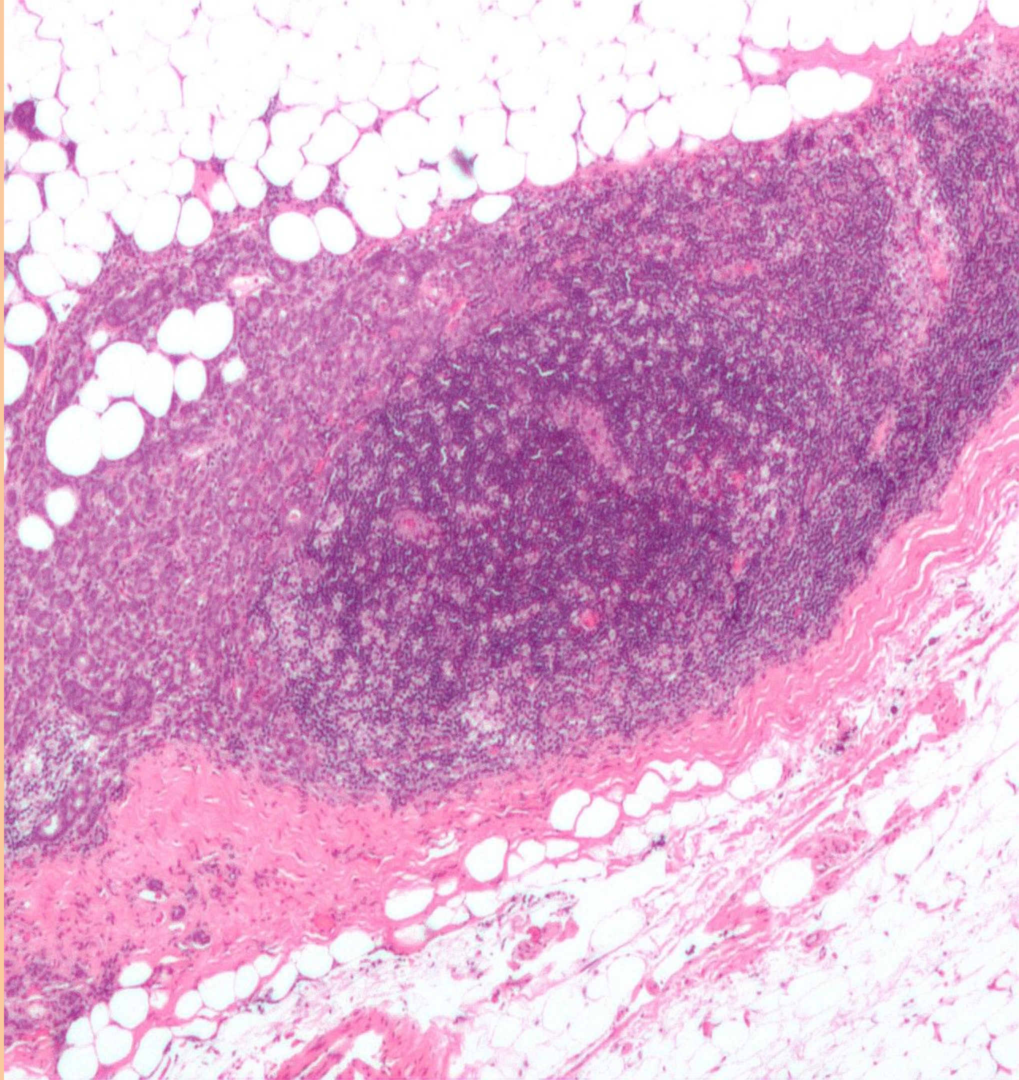
LAB GOAL: Create a narrative from your analysis of the Ripples dataset, where you mention the following:

- Explore this dataset and note your observations
- First, model this with an easily interpretable regressor
- Apply at least half a dozen other different pregressors, that need not be interpretable
- Compare the results, and create justifications for the differential performances

Dataset

Data is in the file: `regressors/ripples.csv`

Breast Cancer Dataset



Explore the Breast-Cancer dataset

This is a much-studied dataset, where Fine-Needle Aspiration gives tissue samples, from which the original authors did a classic exercise of feature engineering.

CONCEPTUAL GOAL

- Read the original paper to appreciate the process of feature engineering
- What would be the crucial elements of your predictive model-diagnostics?
- Can you explain why some models work so well?

LAB GOAL: Create a narrative from your analysis of the Breast-cancer dataset, where you mention the following:

- Start with an exploratory data-analysis with visualizations
- Apply a battery of classifiers to the data, where you apply the necessary hyperparameter tuning to get the most out of each classifier
- Provide a comparative analysis of the models

Dataset

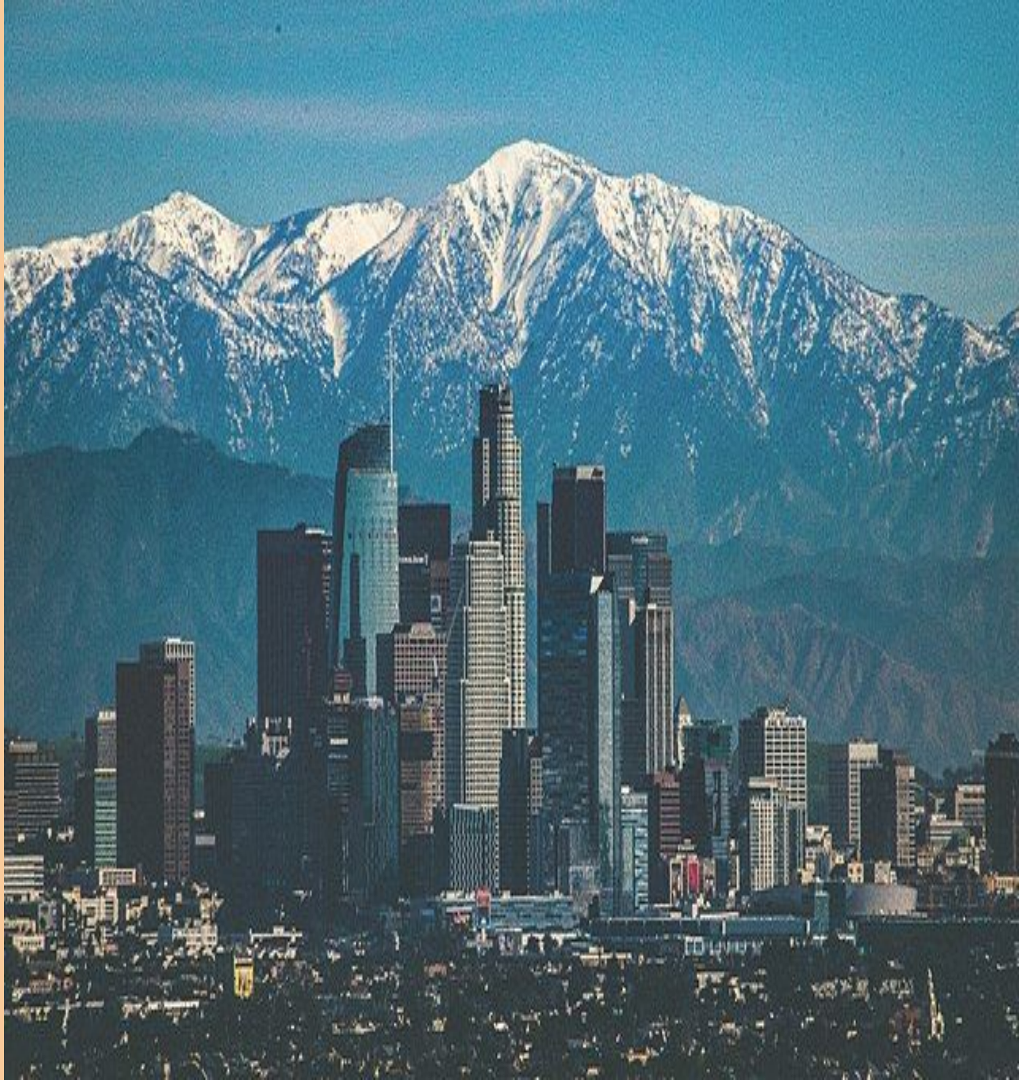
Use the UCI Wisconsin Breast-Cancer Dataset

https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html#sklearn.datasets.load_breast_cancer

California Housing

Predicting the house prices in California



Explore the California Housing dataset

The house-prices are affected by many factors. It is therefore illustrative to build a predictive model for it.

CONCEPTUAL GOAL

- Explain the presence of clusters in the data
- What is the data missing, and where can you get it?
- What features would you extract from the raw-data?

LAB GOAL: Create a narrative from your analysis of the California dataset, where you mention the following:

- Create multiple visualizations, illustrating the various relationships
- Build a predictive model for the prices
- Feature-engineer something useful
- What external data can you join with?
- Tell a story by relating to something you know about California and its history

Dataset

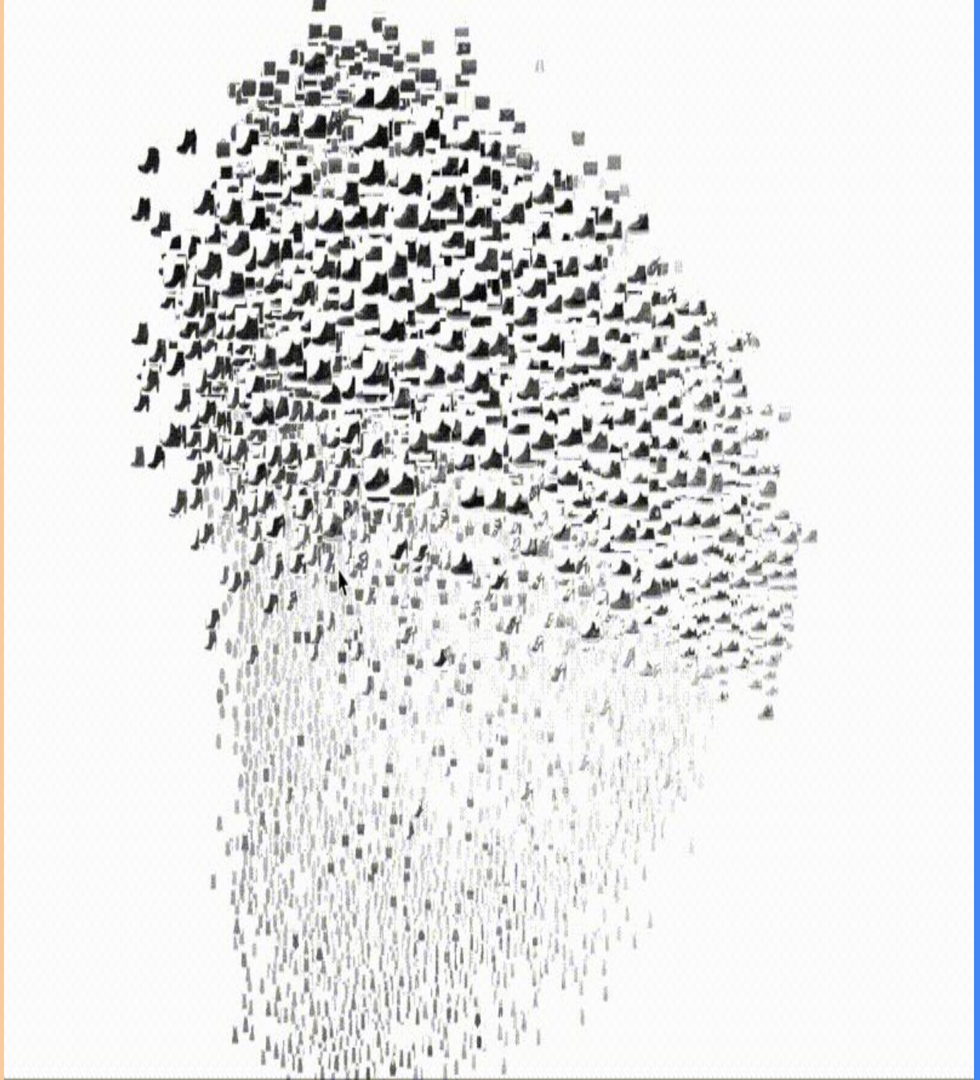
Use the California Housing Dataset

https://scikit-learn.org/stable/datasets/index.html#california-housing-dataset

https://www.kaggle.com/camnugent/california-housing-prices

MNIST & FASHION MNIST

The Hello-World of Machine-Learning



Explore the MNIST and FASHION-MNIST datasets

Over the years, the MNIST dataset has gone from being a benchmark for machine-learning algorithms to the poster-child of success of modern algorithms. Fashion-MNIST is a better alternative for comparative study of algorithms and it retains backward compatibility with MNIST

CONCEPTUAL GOAL

-
- What would be the crucial elements of your predictive model-diagnostics?
- Can you explain why some models work so well

LAB GOAL: Create a narrative from your analysis of the MNIST dataset, where you mention the following:

- Start with an exploratory data-analysis with visualizations.
- Perform dimensionality reduction based visualizations of the MNIST and Fashion-MNIST
- Apply a battery of at least 10 different classifiers besides CNN to the data
- (Do not simply use CNN architectures and be done!)
- Provide a comparative analysis of the models