Stony Brook University

Machine Learning Overview

Vibha Mane
The COSINE Lab
Department of Electrical and Computer Engineering

What is Machine Learning?

- Machine Learning is the task of **learning from data**, that is, generating models from data, and then **making predictions**.
- In particular, we want to **unveil possible hidden patterns** and structures from data and use this information for analysis and understanding of the nature of data.
- We want to develop **mathematical algorithms** to train a model from data, and utilize it to make predictions or decisions, in an automated manner.

Machine Learning Definition
Tom Mitchell,1997

- A widely quoted definition, as given by Tom M. Mitchell:

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

- **Example Handwritten Digit Recognition**
  - Task T: recognize and classify digits
  - Performance Measure P: accuracy of classification
  - Experience E: training with images

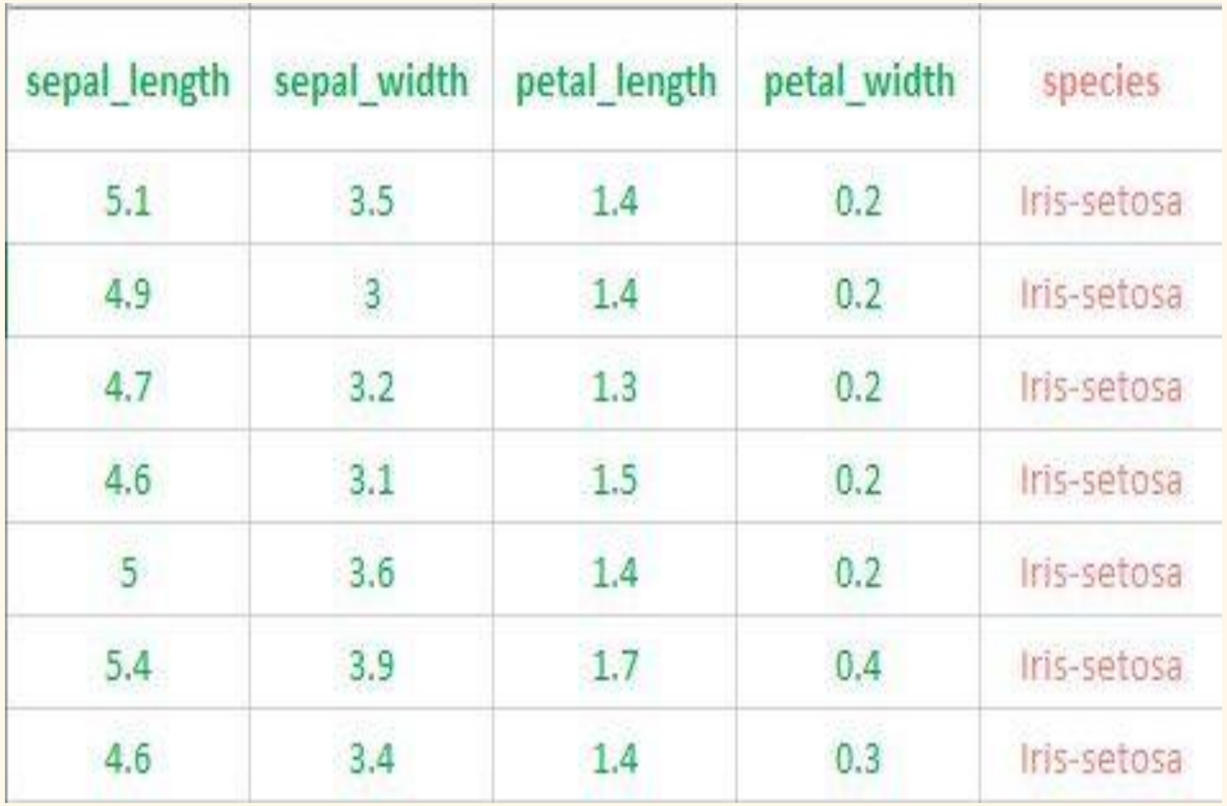
Machine Learning Example - Classification

- As an example, we have the **Iris Flower Data Set** or Fisher’s Iris Data Set (Fisher, 1936) which consists of **50 samples or observations** from each of the three species of Iris: Iris Setosa, Iris Virginica and Iris Versicolor. The **three species** are called **class labels** or **target features**.

- For each of the 150 samples, **four features** were measured: sepal length, sepal width, petal length and petal width. These four features are called **descriptive features or** just **features**

- The objective here is to **learn the model** from the given data set, so that we can identify which species a new observation belongs to. This task is known as **classification**.

Iris Flower Data

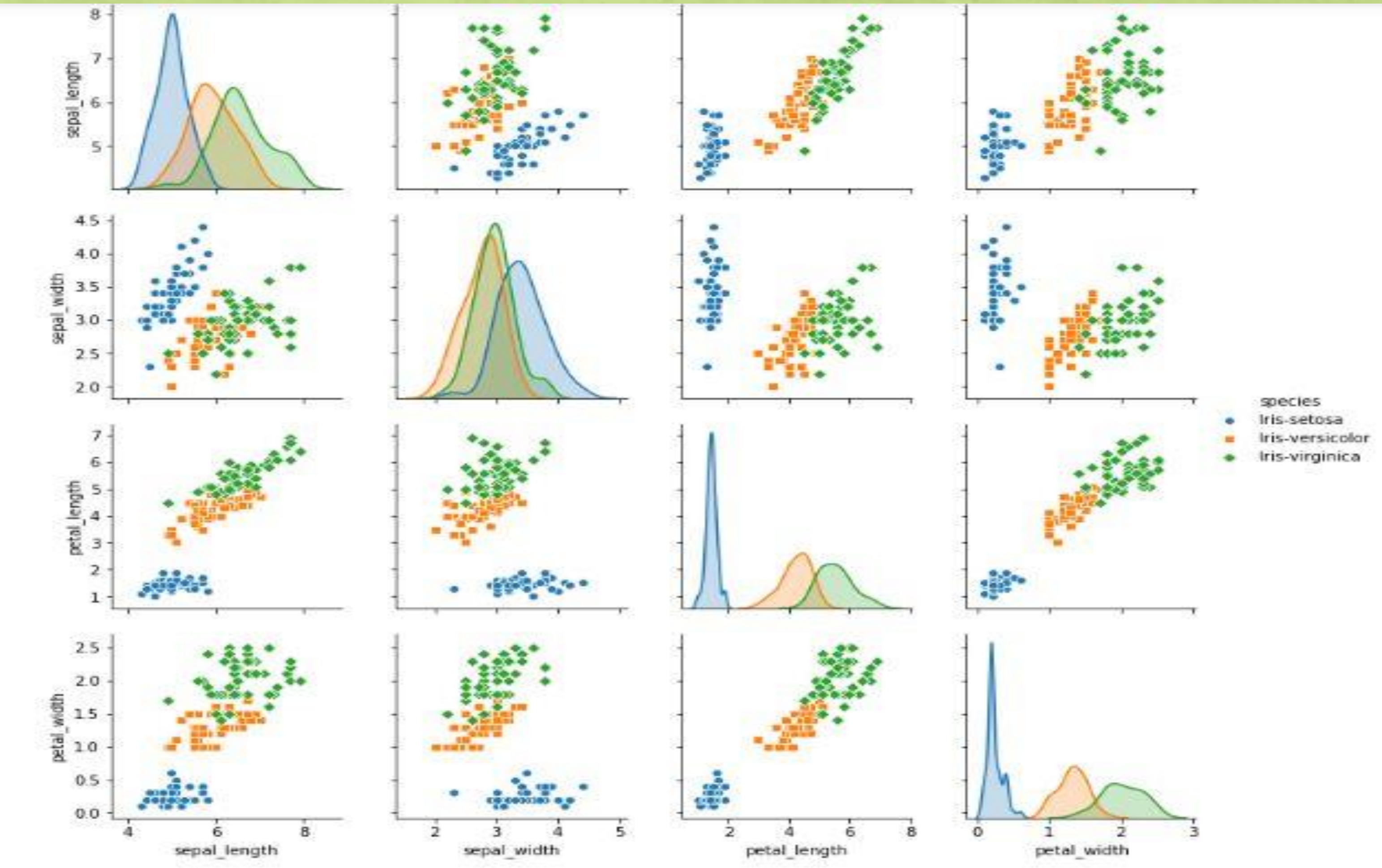


sepal_length	sepal_width	petal_length	petal_width	species
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
5	3.6	1.4	0.2	Iris-setosa
5.4	3.9	1.7	0.4	Iris-setosa
4.6	3.4	1.4	0.3	Iris-setosa

A subset of the Iris Data Set:

- Each of the first four columns contains a feature.
- The last column is the class label, for the three different species.
- Each row is an observation of a different iris flower; these observations are known as **samples**.

Scatter Plot Matrix of the Iris Data Set



- **Scatter plot matrix** of the Iris data set with 4 features.
- It can visually represent multiple features of a data set to explore their relationship and discover hidden patterns.
- The plots on the diagonal are density plots, where each color represents a different species.

The Machine Learning Problem

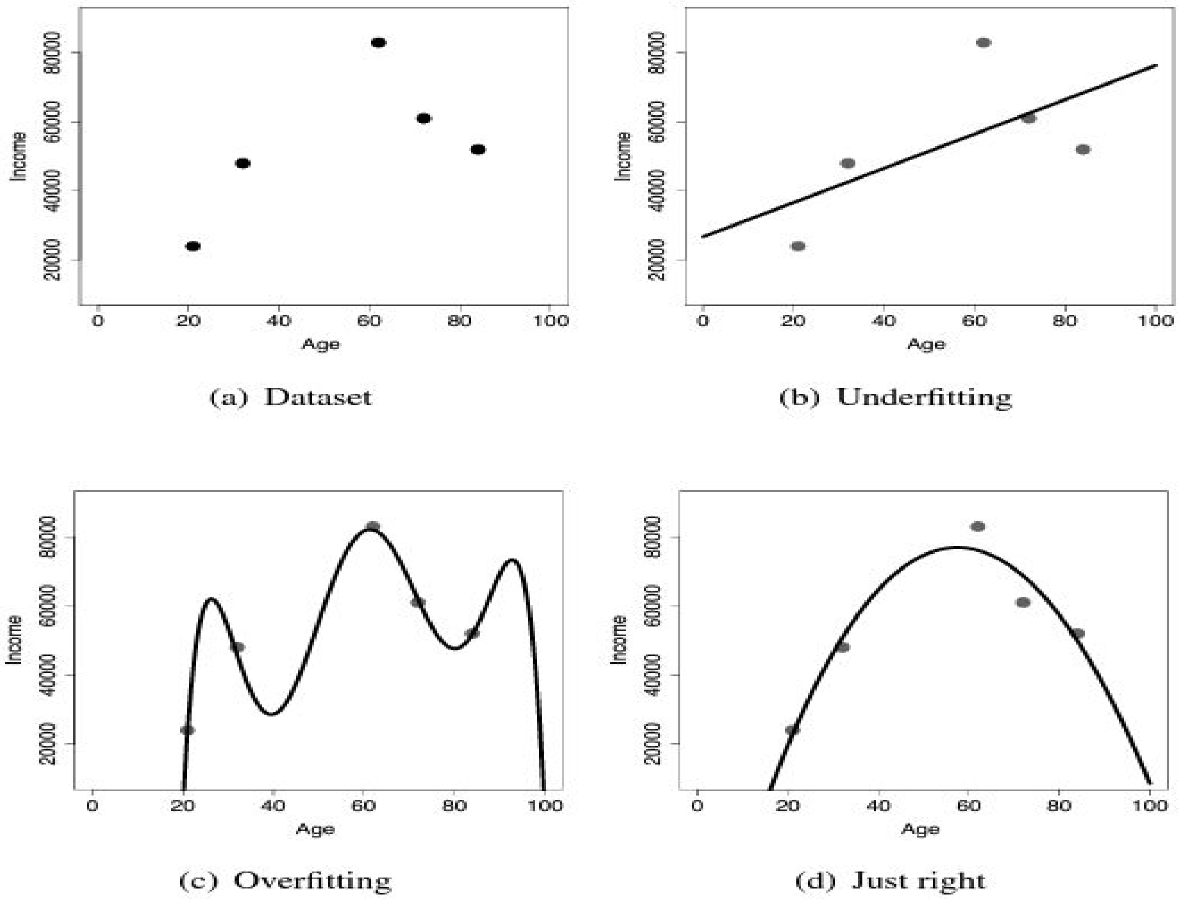
- Machine Learning algorithms work by searching through a set of possible prediction models for the **model that best captures the relationship between the features and the class labels**.
- We want a model that is consistent with the given data. However, the given data set typically represents only a small sample subset of all the possible instances. For example, with m features and n instances, there are m^n possible models, **many of which would fit the given data set**.
- Further, there may be noise in the data. Therefore, we say that machine learning is a complex problem.

The Machine Learning Models

- **Underfitting** occurs when the prediction model selected by the algorithm is too simplistic to represent the underlying relationship in the data set between the features and the class labels.
- **Overfitting** occurs when the prediction model selected by the algorithm is so complex that the model fits the data set too closely and becomes sensitive to noise.
- Striking just the right balance between underfitting and overfitting is called the **Goldilocks model**.

The Machine Learning Models (contd)

John D. Kelleher at al., 2015



(a) Dataset

(b) Underfitting

(c) Overfitting

(d) Just right

Types of Machine Learning Tasks

The machine learning tasks are broadly classified into three categories:

- **Supervised Learning:** The data set consists of several example inputs and outputs, and the goal is to find a general rule that maps the input into the outputs, such as image classification.
- **Unsupervised Learning:** The data set contains no labels (or target features), and the goal is to find interesting or hidden patterns in the data, for example, discovering graph structures.
- **Reinforcement Learning:** The data is given as a feedback, and the program takes action so as to maximize a reward, for example, playing the mancala game.

Model Selection for Supervised Learning

- We want to develop a general-purpose algorithm, focused on solving a particular problem.
- To this effect, we want to collect, preprocess and utilize the given data in an effective manner.
- For supervised learning, the machine learning model is built in **two stages: training stage and test stage**.
- In the **training stage**, the parameters of the model are learned from the training data set utilizing a **supervised learning algorithm**.
- In the **test stage**, the fitted model is then used to make predictions on the test data set, so as to measure how well the model is trained.

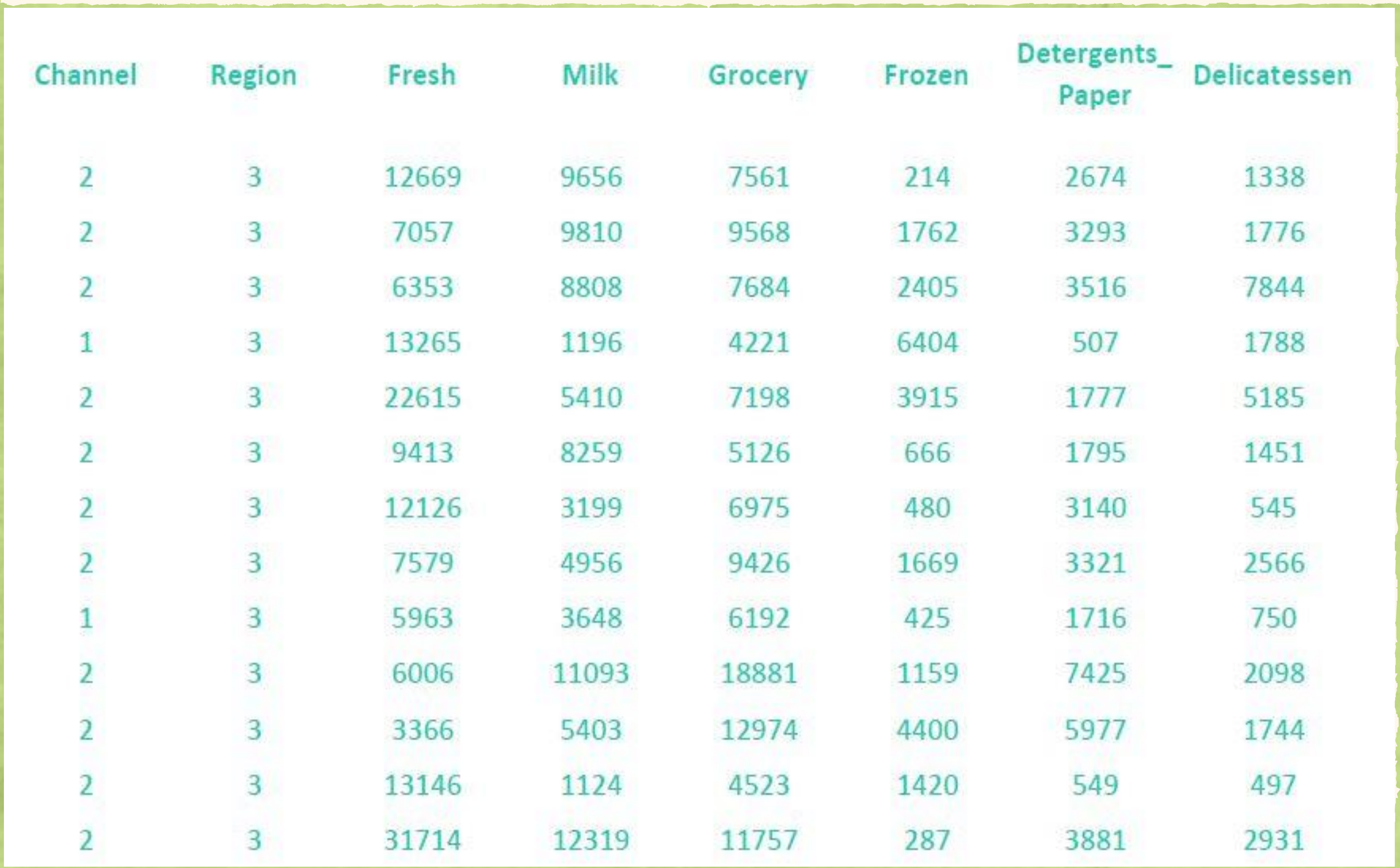
Model Selection for Supervised Learning Example

- The **classification** of the Iris flower species from the given features is an example of supervised learning.
- The data set contains 50 samples (the number of rows). We split the data into training set , with 40 instances, and test set with 10 instances.
- We utilize the training data set and a classification algorithm, say Naïve Bayes classifier, **to learn the parameters of the model**.
- Next, we utilize the test data set and the learned parameters, to classify the species. A measure of error in classification, such as least square error is used to **validate the model**.

Model Selection for Unsupervised Learning

- Quite often, the data set contains **no target features or labels**.
- The wholesale customer data, shown in the following slide, is an example of data with no labels.
- In this case, the goal is to find interesting patterns or similar groups in the data.
- **Clustering** is an example of unsupervised learning. It is the task of finding **homogeneous groups** in data.
- The model, that is number and types of clusters, may be selected based on business needs or some other specified criteria.

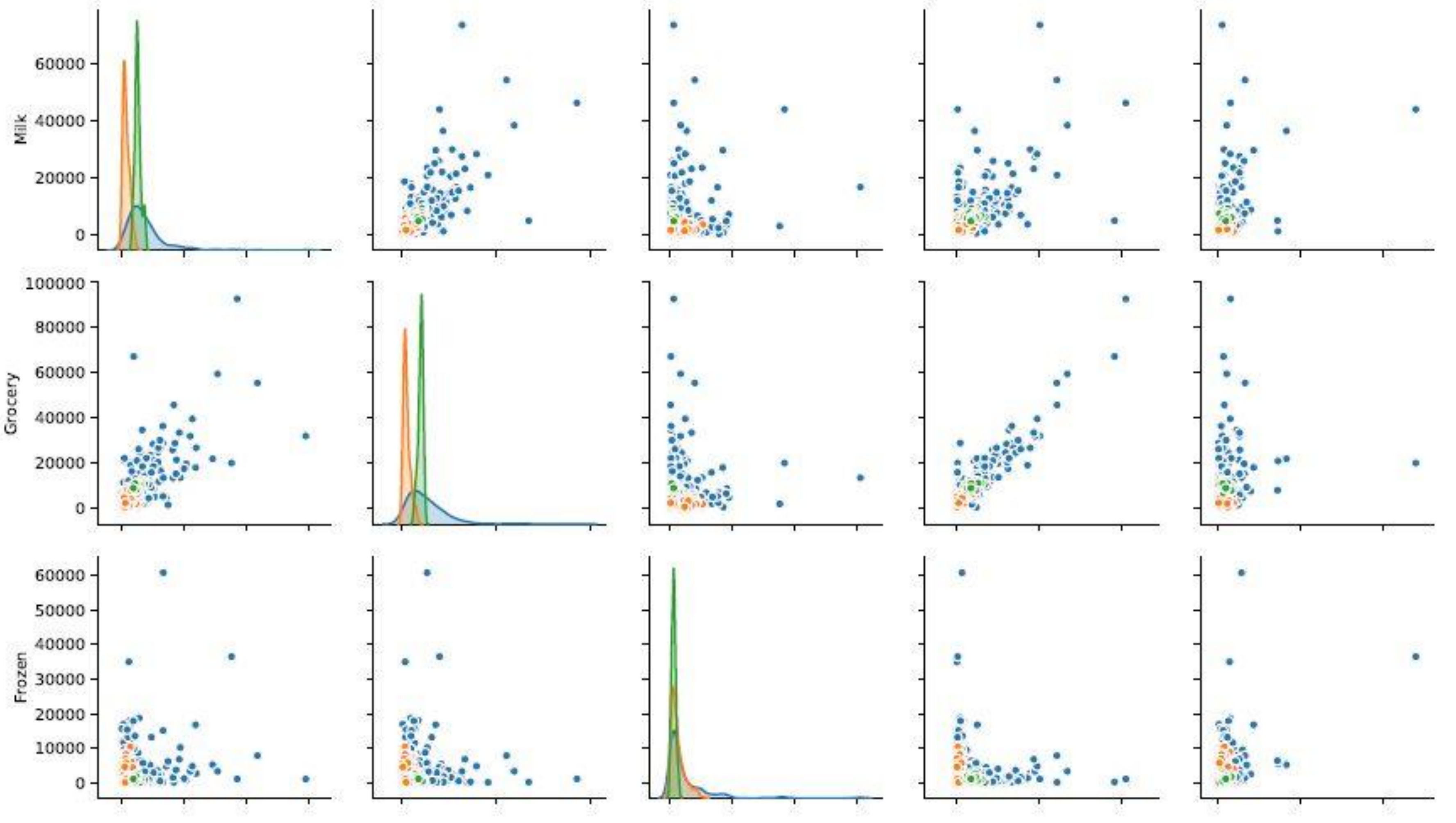
Model Selection for Unsupervised Learning Example (1 of 3)



Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_ Paper	Delicatessen
2	3	12669	9656	7561	214	2674	1338
2	3	7057	9810	9568	1762	3293	1776
2	3	6353	8808	7684	2405	3516	7844
1	3	13265	1196	4221	6404	507	1788
2	3	22615	5410	7198	3915	1777	5185
2	3	9413	8259	5126	666	1795	1451
2	3	12126	3199	6975	480	3140	545
2	3	7579	4956	9426	1669	3321	2566
1	3	5963	3648	6192	425	1716	750
2	3	6006	11093	18881	1159	7425	2098
2	3	3366	5403	12974	4400	5977	1744
2	3	13146	1124	4523	1420	549	497
2	3	31714	12319	11757	287	3881	2931

- The wholesale customer data is taken from the **UCI Machine Learning Repository [3].**
- It shows the customer annual spending on various products, in monetary units.
- The data contains 8 features, and 440 samples.
- The feature **Frozen** gives the annual spending on milk products.
- The feature **Channel** refers to retail vs. restaurant/hotel.
- The data set has **no labels**.

Model Selection for Unsupervised Learning Example (2 of 3)



- The wholesale customer data is clustered using the **DBSCAN algorithm [2].**
- Six numerical features, namely, Fresh, Milk, Grocery, Frozen, Detergents_Paper and Delicatessen are used to find similar groups.
- The scatter plot shows **two clusters** in orange and green. The blue data points are the noise points.

Model Selection for Unsupervised Learning Example (3 of 3)

- DBSCAN is a **density-based clustering algorithm**. It groups together objects that are closely packed, marking as **outliers'** points that are in the low-density regions.
- The clusters in green and orange are separated in Milk vs. Grocery feature space (row 1, column 2 and row 2, column 1).
- The two clusters are similar in the feature Frozen. Therefore, this feature does not help discriminate between the clusters.
- This example will be studied in more details in the Clustering Module.

Machine Learning Data Structure

Machine Learning data encompasses structured, semi-structured and unstructured data:

- The Iris Data Set shown earlier, with a well-defined format and features, is an example of **structured data**.
- Text data with some discernible pattern, which enables parsing, is an example of **semi-structured data**.
- Data which has no inherent structure, such as email messages, PDF files, videos, photos and webpages are examples of **unstructured data**.

Machine Learning Data Structure, cont.

- About **90% of the data** being generated on the web and the real world is unstructured data.
- Most machine learning algorithms are employed on structured data. Therefore, we perform **feature selection task** on unstructured data.

Machine Learning Tasks

Some common Machine Learning tasks:

- Regression
- Classification
- Clustering
- Recommender Systems
- Market Basket Analysis
- Feature Selection
- Feature Reduction
- Optimization

Machine Learning Applications (1 of 3)

- **Regression** is the task of making predictions from features, such as predicting house prices from location, house size, plot size, number of rooms and so forth.
- Another example of **regression** is predicting the age of a viewer watching a YouTube video.
- In e-commerce, users are **clustered** into groups, based on their purchasing or web-surfing habits.
- Another example is **clustering** of astronomical data into types of galaxies or stars.

Machine Learning Applications (2 of 3)

- Handwriting recognition with the standard MNIST data set is an example of image **classification**.
- Another example of **classification** is landmark recognition challenge posed on Kaggle. Here the task is to recognize landmarks, such as the Coliseum or the Eifel tower, from images.
- In text analysis, such as document classification, we perform **feature selection**, such as frequency of uncommon words, to perform the classification task.
- Time-series data such as a Fetal Heart Rate is an example of unstructured data. To classify the data into reassuring or non-reassuring, we perform **feature selection** such as variability and accelerations in heart rate.

Machine Learning Applications (3 of 3)

- Companies such as Netflix and Amazon utilize **recommender systems**, based on user ratings, or similar groups, to recommend contents or products to users.
- In **market basket analysis**, the data set consists of items (or products) and transactions. The goal is to predict which other items, the consumer is likely to purchase, based on what is in their basket.

References

1. John Kelleher, Fundamentals of Machine Learning for Predictive Data Analytics, The MIT Press, 2015.

2. Martin Ester et al., *A density-based algorithm for discovering clusters in large spatial databases with noise*, KDD, 1996.

3. Dua, D. and Graff, C., UCI Machine Learning Repository, Irvine, CA, University of California, School of Information and Computer Science, 2019.