**BJ: to Lab-4, PR txt-1** (c,г)

**Bone Marrow (BM) mononuclear cells with AML (M=1000, N= 1004)**

BM Dataset Characteristics:

1. **Bone Marrow (BM) Mononuclear Cells with AML Dataset:**
   * **Description:** The Bone Marrow (BM) Mononuclear Cells with Acute Myeloid Leukemia (AML) dataset is a dataset included in the Orange system. It consists of data on mononuclear cells extracted from bone marrow samples of patients diagnosed with AML.
   * **Main Features:** The dataset likely includes various features characterizing the properties of mononuclear cells, such as cell size, shape, density, and other cellular characteristics measured through laboratory analyses.
   * **Categorical Indices:** Depending on the specific features included in the dataset, there may be categorical indices representing categorical variables, such as cell type, treatment regimen, or disease subtype.
   * **Details:** The dataset is structured with rows representing individual samples or patients and columns representing different features or attributes measured for each sample. It may also include additional metadata such as patient identifiers, sample collection dates, and clinical outcomes.
2. **BM and k-Nearest Neighbors (kNN):**
   * **Usage:** The BM dataset can be used with the k-Nearest Neighbors (kNN) algorithm in Orange for classification tasks.
   * **Approach:** kNN is a non-parametric classification algorithm that assigns a class label to a sample based on the majority class among its k nearest neighbors in the feature space.
   * **Implementation:** In Orange, the BM dataset can be loaded and preprocessed as needed before applying the kNN classifier to predict the AML status of new bone marrow samples based on the characteristics of existing samples in the dataset.
3. **BM and t-distributed Stochastic Neighbor Embedding (t-SNE):**
   * **Usage:** t-SNE is a dimensionality reduction technique commonly used for visualizing high-dimensional data in lower-dimensional space.
   * **Approach:** t-SNE aims to capture the underlying structure of the data by modeling similarities between data points in the high-dimensional space and preserving them in the lower-dimensional space.
   * **Implementation:** With Orange, the BM dataset can be used with the t-SNE visualization tool to explore and visualize the relationships between mononuclear cell samples in a lower-dimensional space, potentially revealing clusters or patterns indicative of different cell types or disease subtypes.
4. **BM and Principal Component Analysis (PCA):**
   * **Usage:** PCA is a dimensionality reduction technique used to transform high-dimensional data into a lower-dimensional space while preserving the most important variance in the data.
   * **Approach:** PCA identifies orthogonal axes (principal components) along which the variance of the data is maximized and projects the data onto these components, reducing the dimensionality while retaining the most important information.
   * **Implementation:** In Orange, the BM dataset can be utilized with PCA to reduce the dimensionality of the mononuclear cell data, enabling visualization, exploration, or further analysis while retaining the essential characteristics of the original dataset.
5. **BM and Neural Networks in Orange:**
   * **Usage:** Neural networks, including various architectures such as feedforward, convolutional, or recurrent neural networks, can be applied for classification, regression, or other tasks on the BM dataset.
   * **Approach:** Neural networks learn complex patterns and relationships from data by adjusting the weights of interconnected neurons through training iterations using optimization algorithms.
   * **Implementation:** Orange provides neural network modules that allow users to build, train, and evaluate neural network models using the BM dataset. This enables tasks such as classifying bone marrow samples into AML or non-AML categories based on the learned patterns in the data.

**Housing Dataset in Orange**

**1. Housing Dataset Characteristics:**

The Housing dataset in Orange contains 13 features (columns) that describe various aspects of housing in the Boston area. Here's a brief overview of each feature:

**1. CRIM:** Per capita crime rate by town. **2. ZN:** Proportion of land zoned for lots over 25,000 sq. ft. **3. INDUS:** Proportion of non-retail business acres per town. **4. CHAS:** Charles River dummy variable (1 if tract bounds river; 0 otherwise). **5. NOX:** Nitric oxides concentration (parts per 10 million). **6. RM:** Average number of rooms per dwelling. **7. AGE:** Proportion of owner-occupied units built before 1940. **8. DIS:** Weighted distances to five Boston employment centers. **9. RAD:** Index of accessibility to radial highways. **10. TAX:** Full-value property-tax rate per $10,000. **11. PTRATIO:** Pupil-teacher ratio by town. **12. B:** 1000(Bk - 0.63)^2 where Bk is the proportion of black people by town. **13. LSTAT:** Lower status of the population (percent).

**1.2 MEDV - Median Value of Owner-Occupied Homes:**

MEDV is the target variable in the Housing dataset. It represents the median value of owner-occupied homes in each town, expressed in thousands of dollars. The acronym MEDV stands for "Median Value."

**2. Classes of Applied Problems:**

**2.1 MEDV as a Continuous Variable:**

* **Regression:** Predicting the median home value based on the other features.
* **Time Series Analysis:** Forecasting future median home values.
* **Clustering:** Identifying groups of towns with similar housing characteristics.

**2.2 MEDV as a Categorical Variable:**

* **Classification:** Categorizing towns into different classes based on their median home value (e.g., affordable, luxury).
* **Decision Tree Learning:** Creating a decision tree to predict whether a home's value is above or below a certain threshold.

There are many other potential applications and analyses that can be performed with this dataset.

1. [harvard-iacs.github.io/2019-CS109A/labs/lab04/notebook/](https://harvard-iacs.github.io/2019-CS109A/labs/lab04/notebook/)
2. [towardsdatascience.com/apache-spark-mllib-tutorial-ec6f1cb336a9](https://towardsdatascience.com/apache-spark-mllib-tutorial-ec6f1cb336a9)
3. [github.com/AkashIngole/boston-housing-prediction-ml-app](https://github.com/AkashIngole/boston-housing-prediction-ml-app)

**Housing Dataset and Neural Networks in Orange**

Some potential tasks we can tackle using NNs in Orange with the Housing dataset:

**1. Regression with NNs:**

* **Predicting Median Home Value:** You can train a NN to predict the continuous target variable, MEDV (median home value), based on the 13 input features. This is a classic regression task where the NN learns the non-linear relationships between the features and the target variable.

**2. Classification with NNs (**Discretized MEDV **!):**

* **Categorizing Homes by Value:** You can discretize the continuous MEDV variable into categories (e.g., affordable, mid-range, luxury) based on specific value thresholds. Then, train a NN to classify new housing data points into these categories based on their features.

**3. Hybrid Approach (Regression + Classification):**

* **Two-Stage Prediction:** This approach involves two steps:
  + Train a first NN to predict the continuous MEDV value.
  + Train a second NN (or another classifier) to categorize the predicted MEDV values into the desired classes based on pre-defined thresholds.

**Benefits of Using NNs for these Tasks:**

* **Non-linear Modeling:** NN can capture complex, non-linear relationships between features that might be missed by simpler models like linear regression.
* **Feature Learning:** NNs can automatically learn important features from the data during training, potentially leading to improved performance compared to models that rely on hand-crafted features.

**Additional Considerations:**

* **Data Preprocessing:** Before training NNs, it's crucial to preprocess the data appropriately, such as scaling the features and handling missing values.
* **Hyperparameter Tuning:** The performance of NNs depends on various hyperparameters (e.g., number of neurons, learning rate). Experiment with different settings to optimize the NN for your specific task.
* **Evaluation Metrics:** Choose appropriate evaluation metrics (e.g., mean squared error for regression, accuracy, precision, recall for classification) to assess the effectiveness of your NN model.

By leveraging NNs in Orange, you can explore various approaches to analyze and predict housing prices using the Housing dataset.

**The IRIS dataset**

**The Iris Flower Dataset in Orange**

The Iris flower dataset you encountered in Orange is a classic and widely used benchmark dataset in ML for classification problems. Here's a breakdown of its details:

**History and Description:**

* The dataset was first published by Ronald Fisher in his 1936 paper "The use of multiple measurements in taxonomic problems."
* It describes three different species of iris flowers: Iris Setosa, Iris Versicolor, and Iris Virginica.
* Each data point in the dataset represents a single iris flower and contains measurements of four features:
  + Sepal Length (cm)
  + Sepal Width (cm)
  + Petal Length (cm)
  + Petal Width (cm)

**Column Information:**

* **Number of Columns:** 5 (including the target class label)
* **Quantitative Features:** All four features (Sepal Length, Sepal Width, Petal Length, Petal Width) are continuous numerical values (quantitative).
* **Categorical Feature:** There is one categorical feature, which is the target class label representing the three iris species (Setosa, Versicolor, Virginica).

**Goal of Use in Orange and Machine Learning:**

* **Classification:** The primary goal of using the Iris dataset in Orange and machine learning is to build classification models that can accurately predict the iris species based on the four flower measurements.
* **Benchmarking:** Due to its well-defined structure, small size, and clear separation between classes, the Iris dataset is often used as a benchmark for testing and comparing the performance of different machine learning classification algorithms.
* **Learning Tools:** The simplicity of the Iris dataset makes it a great starting point for beginners to learn about classification algorithms and experiment with different techniques in Orange.

**Benefits of Using the Iris Dataset:**

* **Simplicity:** Easy to understand and interpret, making it approachable for beginners.
* **Publicly Available:** Freely accessible, allowing everyone to experiment and compare results.
* **Well-Documented:** Extensive literature and resources available to understand the data and task.
* **Suitable for Various Algorithms:** Can be used to test and compare a wide range of classification algorithms.

**Additional Notes:**

* In some versions of Orange, the Iris dataset might be pre-loaded and displayed as a widget.
* You can use this data to train various classification models in Orange, such as:
  + k-Nearest Neighbors (kNN)
  + Support Vector Machines (SVM)
  + Decision Trees
  + Neural Networks
* By experimenting with different algorithms on the Iris dataset, you can gain valuable insights into their strengths and limitations for classification tasks.