**BJ: to Lab-3, PUM txt-1** (г)

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

BM Dataset Characteristics:

1. There is a case study, added to Orange system: Bone Marrow (BM) mononuclear cells with AML (M=1000, N= 1004). What is this dataset?

Main features, categorical indices, details, etc.

**1. Understanding the BM Dataset**

**a. Description:**

The BM dataset in Orange likely represents a collection of features extracted from bone marrow mononuclear cells of patients diagnosed with Acute Myeloid Leukemia (AML). It contains information about M = 1000 samples (observations) and N = 1004 features (variables).

**b. Main Features (Hypothetical):**

Due to the lack of specific details about the dataset, here are some common features you might encounter in a bone marrow mononuclear cell dataset for AML analysis:

* **Morphological features:** Cell size, shape, nucleus-to-cytoplasm ratio, presence of abnormal granules, etc.
* **Immunophenotyping:** Expression levels of specific cell surface markers that can help identify different types of white blood cells.
* **Cytogenetic features:** Chromosomal abnormalities associated with AML.
* **Gene expression data:** Levels of RNA transcripts for various genes.
* **Clinical data:** Patient demographics, medical history, treatment information, etc.

**c. Categorical Indices:**

Categorical features in the dataset could include:

* **Diagnosis:** AML subtype (e.g., M0, M2, M4)
* **Patient demographics:** Age, gender, ethnicity
* **Treatment:** Type of chemotherapy, radiation therapy, etc.
* **Outcome:** Response to treatment, survival time

**d. Additional Details (Exploration Needed):**

* **Data types:** The specific data types of each feature (numerical, categorical, text) can be crucial for choosing appropriate analysis techniques.
* **Missing values:** The presence and handling of missing values can affect analysis results.
* **Normalization/scaling:** Depending on the data types and analysis goals, features might need normalization or scaling.

**2. BM and k-Nearest Neighbors (kNN)**

**a. Potential Applications:**

kNN is a non-parametric supervised learning algorithm. In the context of the BM dataset, kNN could be used for:

* **Classification:** Classifying new bone marrow samples as belonging to a specific AML subtype based on their similarity to existing labeled samples in the dataset.
* **Clustering (unsupervised):** Grouping similar bone marrow samples together based on their features, potentially revealing new subtypes or patterns.

**b. Considerations:**

* **Feature selection:** Choosing the most informative features can improve kNN performance.
* **Number of neighbors (k):** The value of k can significantly impact classification accuracy. Experiment with different k values to find the optimal setting.
* **Distance metric:** The metric used to measure similarity between samples (e.g., Euclidean distance) can influence results. Choose a metric suitable for your data characteristics.

**3. BM and t-SNE**

**a. Application:**

t-SNE (t-distributed Stochastic Neighbor Embedding) is a dimensionality reduction technique used for visualizing high-dimensional data in a lower-dimensional space (typically 2D or 3D) while preserving the relationships between similar samples.

**b. Use Case:**

* Visualize the BM dataset to explore potential clusters or groupings of bone marrow samples based on their features.
* Identify outliers or anomalies in the data that might require further investigation.

**c. Considerations:**

* **Choosing the perplexity parameter:** This parameter controls how tightly packed the points are in the lower-dimensional space. Experiment with different values for optimal visualization.

**4. BM and PCA (Principal Component Analysis)**

**a. Application:**

PCA is another dimensionality reduction technique that identifies the directions of greatest variance in the data. It projects data points onto these principal components, reducing dimensionality while capturing most of the information.

**b. Use Case:**

* Reduce the dimensionality of the BM dataset for further analysis with other machine learning algorithms that might be less efficient with high-dimensional data.
* Identify the most informative features that contribute most to the variance in the data.

**c. Considerations:**

* **Number of components to retain:** The number of principal components chosen for analysis depends on the desired level of information preservation and computational efficiency.

**5. BM and Neural Networks in Orange**

**a. Potential Applications:**

Neural networks are powerful machine learning models capable of learning complex relationships between features and targets. In the context of the BM dataset, neural networks could be used for:

* **Classification:** Similar to kNN, neural networks can classify new bone marrow samples into AML subtypes or predict patient outcomes.
* **Unsupervised learning:** Autoencoders, a type of neural network, can be used for dimensionality reduction or anomaly detection in the BM dataset.

**b. Considerations:**

* **Architecture selection:** Choosing the right neural network architecture (number of layers, neurons, etc.) is crucial for optimal performance.
* **Training parameters:** Hyperparameters like learning rate, batch size, and optimizer need to be tuned for the specific problem and dataset.
* **Overfitting:** Neural networks are prone to overfitting, especially with high-dimensional data. Techniques like regularization and early stopping can help prevent this.
* **Evaluation metrics:** Choose appropriate evaluation metrics (e.g., accuracy, precision, recall, F1-score) aligned with your analysis goals.

**c. Orange Capabilities:**

Orange provides a user-friendly interface for building and training neural networks. You can explore various network architectures (multilayer perceptrons, convolutional neural networks, etc.) and experiment with different hyperparameters. Orange also offers visualization tools to help you understand your network's behavior.

**Additional Tips:**

* **Data preprocessing:** Ensure proper data cleaning, handling missing values, normalization/scaling, and feature selection before applying any machine learning algorithms.
* **Exploration is key:** Use visualization techniques like histograms, scatter plots, and boxplots to understand the distribution of features and identify potential relationships.

By understanding the BM dataset, its features, and the strengths and limitations of different machine learning algorithms, you can effectively leverage Orange for various analyses related to bone marrow mononuclear cells and AML research.