ISWZ3402 Inteligencia Artificial II

Ejercicio práctico

Elaborado por: Mario González

RESULTADO DE APRENDIZAJE DE LA CARRERA:

RC7. Adquiere y aplica nuevos conocimientos según sea necesario, utilizando estrategias de aprendizaje apropiadas.

Indicadores de desempeño:

- Indaga técnicas y herramientas necesarias acorde al problema planteado
- Argumenta las técnicas y herramientas disponibles
- Aplica las técnicas y herramientas pertinentes al problema planteado

OBJETIVO PROPUESTO DE LA CONSIGNA:

This workshop aims to explore the critical question of determining an optimal sample size for machine learning problems, focusing on a comparative analysis between random sampling and semi-supervised sampling using k-means clustering. Participants will delve into the considerations, advantages, and limitations of each sampling approach. Through practical exercises, attendees will gain hands-on experience in implementing these methodologies, enabling them to make informed decisions about sample sizes based on the specific requirements and nuances of their machine learning projects. By the conclusion of the workshop, participants will be empowered to navigate the trade-offs between different sampling techniques for enhanced model performance.

INDICACIONES:

- The number of samples needed in machine learning depends on various factors, including the complexity of the problem, the complexity of the model, the amount of variability in the data, and the desired level of model performance.

Some considerations to take into account for your project:

- **Data Exploration:** Conduct exploratory data analysis to understand the distribution of your data. This can help you identify potential challenges, such as imbalances, outliers, or skewed distributions, which may affect the amount of data needed.

- **Model Complexity:** More complex models often require more data to generalize well. If you're using a simple model, you might need fewer samples, but complex models with many parameters might require more data to avoid overfitting.
- **Data Variability:** If the data is highly variable or noisy, you might need more samples to capture the underlying patterns and reduce the impact of randomness.
- **Cross-Validation:** Using techniques like cross-validation can help you assess how well your model generalizes to new, unseen data. It also provides insights into the stability of your model's performance across different subsets of the data.
- **Learning Curve Analysis:** Plotting learning curves, which show the model's performance as a function of the number of training samples, can help you understand how performance improves with additional data. This can guide decisions about whether more data is beneficial.
- **Problem-Specific Considerations:** Some problems inherently require more data. For example, natural language processing tasks or image recognition tasks might benefit from larger datasets.
- **Resource Constraints:** Consider the computational resources and time available. Training on larger datasets may require more computational power and time.

Detailed instructions

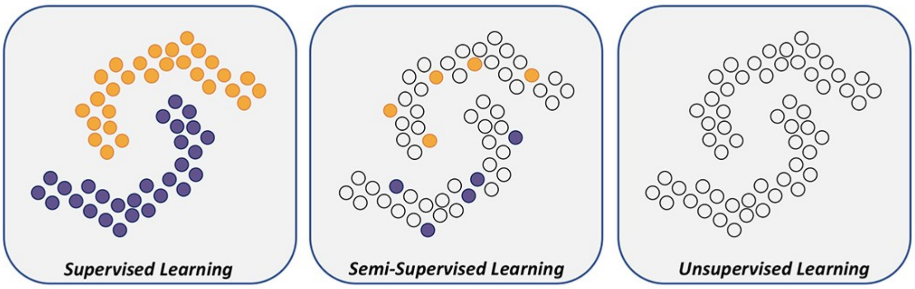
- Data: Work on the MNIST dataset.
    - ❒ Otherwise, your project will be evaluated with 1 point. Equivalent to not delivering the project.
    - ❒ Data:

```
import pandas as pd

mnist = pd.read_csv('https://raw.githubusercontent.com/sbussmann/kaggle-mnist/master/Data/train.csv')
```

Semi-supervised learning

- Random pattern selection vs. clustering selection



- Select a "small" subset to train the model (n instances)
- The subset is selected by random sampling vs using an unsupervised model (KMeans)
- Using the labeled data (the subset of n instances). In the example, we train an MLP classifier.
    - ❒ You must choose an additional ML model and compare it with the MLP.
- Compare the performance for both subsets in both models.
- KMeans selection should be better, until a certain value of n where both subset selections may converge.

Pseudocode*

- Select a subset of size n to train the model (e.g. n=10 instances), for this minimum all 10 classes in the MNIST must be represented in each sample, try to keep it balanced, or compare what happens if not.
- The subset is selected by random sampling and using an unsupervised model (KMeans)
- Using the labeled data (the subset of n instances) we train an MLP classifier and an additional ML model
- Compare the performance for both subsets (Random vs Means) and both ML models.
- For Random sampling use a train-test split or perform your own sampling without replacement
- For KMeans sampling find the more representative instances (observations), that is the instances that are the closest observation to each centroid.
- You will end up with n observations (n clusters).
- The centroid is a representative point that summarizes the characteristics of the group.
- Feed (train) the model with the instances selected in the process above.
- Build a curve of N (sample size) vs the accuracy of the model.
- That is compare Quantity (X) vs Quality (Y) for both sampling mechanisms.
- Compare random vs KMeans for selecting the sample N.
- Show your results in a graph like the one depicted below.

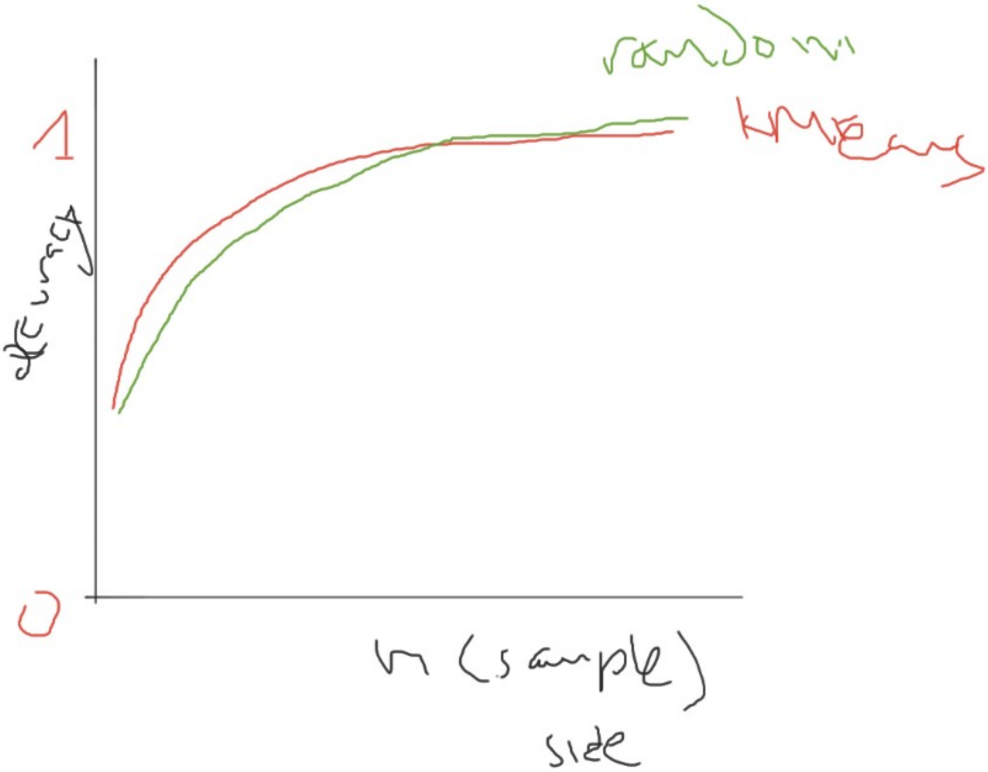
Why should work?

- KMeans finds more representative instances (observations).
- Given that the instances were the closest observation to each centroid.
    - ❒ The centroid is a representative point that summarizes the characteristics of the group.
- The best instances feed the supervised model.
- Compare with a random selection.
- You train on a relatively small data set, but test on all data.
- Conclusion: feeding the model with better data (not necessarily more) results in a better performance.
    - ❒ Quality over Quantity.
- Try more clusters
- Build a curve of n (sample size) vs the accuracy of the model.
    - ❒ That is compare Quantity (x) vs Quality (y).

❒ Compare random vs KMeans for selecting the sample.

  ❍ Compare for both ML models selected.

❒ The figure below is a schematic representation of what you should get for a single model.



General instructions

- (-1 pts) Specify the group members in the first cell of the notebook.
  - ❒ Give an itemized list: each member in a line.
- (-1 pts) Deliver two files into a zip folder with the following name format:
  - ❒ P2_Project_IA202410_II1II2II3.zip: with two files: html **(or pdf)** and ipynb.
  - ❒ For exampel: Nikita Martínez, Nestor Lozada, Damián Briones, results in a file name: P2_Project_IA202410_NM1NL2DB3.zip.
- (-1 pts) **All group members should upload the same file**
- The report (notebook) will be evaluated as:
  - ❒ (-5 pts) Poor: only the code was delivered.
  - ❒ (-5 pts) Fair: commented code, with sections.
  - ❒ (-2 pts) Good: commented code, with sections and explanations of each part of the process.

- ❒ (-0 pts) Excellent: commented code, with sections and explanations of each part of the process including formulas, research, and explanation of your ML model, the dataset exploration, the problem you want to solve, etc. Figures with labels, titles legends, etc.

Sources

- Check:
  - ❒ Kmeans clustering
  - ❒ Semi-supervised learning

FORMA DE TRABAJO:

La propuesta se la desarrollará en grupos de máximo de 5 integrantes.

ESPECIFICACIONES DE ENTREGA:

El estudiante debe entregar un informe completo y detallado en formato ipynb (jupyter notebook) donde se detalle cada una de las fases. Debe exportar el informe a html o pdf, y adjuntar junto con el notebook en ipynb.

Identifica necesidades de aprendizaje

- Define el problema de muestreo: random vs. Kmeans.
- Represente de forma gráfica los resultados comparando ambas formas de muestreo

Selecciona fuentes de información

- Investiga referencias que describen el uso de los modelos de ML seleccionados

Aplica las técnicas y herramientas pertinentes al problema planteado

- Realiza el preprocesamiento de los datos de entrada (si fuese necesario)
- Implementa el escalado adecuado de los datos (recuerde que algoritmo de ML tiene su particularidad de procesar el input)
- Implementa los modelos de ML para el problema y realiza los experimentos indicados
- Describe la solución (combinación de hiperparámetros) con su respectiva justificación
- Conclusiones

RÚBRICA:

CRITERIOS	EXCELENTE	MUY BUENO	BUENO	REGULAR	INSUFICIENTE
IDENTIFICA NECESIDADES DE APRENDIZAJE	Identifica necesidades de aprendizaje de **manera autónoma** y aplica estrategias **apropiadas y relevantes** que le permiten ampliar su conocimiento.	Identifica necesidades de aprendizaje de **manera autónoma** y aplica estrategias **apropiadas** que le permiten ampliar su conocimiento.	Identifica necesidades de aprendizaje de **manera autónoma** y aplica estrategias **generales** que le permiten ampliar su conocimiento.	**Requiere apoyo** para identificar necesidades de aprendizaje y aplica estrategias **poco apropiadas** que le permiten ampliar su conocimiento.	**Requiere apoyo** para identificar necesidades de aprendizaje y **no aplica** estrategias que le permiten ampliar su conocimiento.
SELECCIONA FUENTES DE INFORMACIÓN	Selecciona **de manera proactiva** fuentes de información adicional y **persigue permanentemente** experiencias educacionales más allá de los requerimientos de su entorno de aprendizaje.	Selecciona **de manera activa** fuentes de información adicional y **persigue permanentemente** experiencias educacionales más allá de los requerimientos de su entorno de aprendizaje.	Selecciona **de manera básica, aunque apropiada,** fuentes de información adicional y **persigue regularmente** experiencias educacionales más allá de los requerimientos de su entorno de aprendizaje.	Selecciona de **manera básica y poco apropiada** fuentes de información adicional y **persigue ocasionalmente** experiencias educacionales más allá de los requerimientos de su entorno de aprendizaje	**No selecciona** fuentes de información adicional y **no persigue** experiencias educacionales más allá de los requerimientos de su entorno de aprendizaje.
APLICA DESTREZAS Y CONOCIMIENTOS	Aplica de una **manera innovadora (nueva y creativa)** las destrezas y conocimientos, demostrando comprensión y	Aplica de una **manera adecuada** las destrezas y conocimientos, demostrando comprensión y un **óptimo desempeño**	Aplica de una **manera básica** las destrezas y conocimientos, demostrando comprensión y **buen desempeño** frente a	Aplica de una **manera parcial** las destrezas y conocimientos, demostrando **poca** comprensión y **un desempeño regular**	Aplica de forma **errónea** las destrezas y conocimientos, demostrando un **desempeño deficiente** frente a

	excelente desempeño frente a nuevas situaciones.	frente a nuevas situaciones.	nuevas situaciones.	frente a nuevas situaciones.	nuevas situaciones.