Segundo Laboratorio de Introduction to Machine Learning

Universidad de Ingenieria y Tecnologia (UTEC) – Lima, Peru

September 2023 – Seccion Profesor Arturo Deza

Tema de Laboratorio : Classification

Alumnos/as : NOMBRES

3 points Find a new publicly available dataset between 3 to 10 classes, with no more than 1'000'000 data points, and no less than 1'000 data points. You can not use MNIST, but can use derivatives (you can use datasets other than images) *(Buscar un nuevo dataset publico entre 3 a 10 clases, que tenga no mas de 1'000'000 de puntos de data y no menos de 1'000. No puedes usar MNIST, pero si otras bases de datos asocidadas a MNIST : Ejemplo: Fashion MNIST. Tambien puedes usar bases de datos que no necesariamente tienen que ser imagenes).*

2 points Define your Training Set, Your Validation Set, and Testing Set. How will you partition your data? What is the percentile split? (example: 80% Training, 10% Validation, 10% Testing). Why did you pick this split? *(Definir tu data de entrenamiento, validacion y testeo. Como particionaras tu data? Cuales son los percentiles? (ejemplo: 80% Entrenamiento, 10% Validacion, 10% Testeo). Por que elejiste esta particion?)*

6 points Use a Multi-Class logistic regression classifier to perform a classification for the classes selected. Report the Confusion Matrices (as raw numbers) for the Training Set, the Validation Set, and the Testing Set. Properly label the Actual + Prediction dataset. *(Utiliza un classificador de multi-clase de regresion logistica para el numero de clases elegidas. Reportar la Matriz de Confusion (como numeros en bruto) para el dataset de entrenamiento, validacion y testeo. Tener cuidado en etiquetar correctamente las filas y columnas para etiquetas actuales y predecidas.)*

3 points Perform an L0, L1 or L2 regularization on the Multi-Class Logistic Regression for the previous classes. Re-plot the confusion matrices. What has changed? Is it better or worse than before adding regularization? Why?*(Hacer una regularizacion del tipo L0, L1 o L2 sobre la regresion logistica multi-clase sobre las clases previas. Re-plotear las matrices de confusion. Que ha cambiado? Las matrices de confusion se ven mejor o peor? Por que?)*

6 points 5-Fold Cross-Validation: From the total initial dataset using L0, L1 or L2 regularization: repeat the classification model 5 times through non-overlapping cross-validation. Report the Confusion Matrix on the Training, Validation and Testing Data for each partition (as raw numbers). You should be plotting a

total of 15 sub-figures (3 training/validation/testing x 5 validations). Do this only for 1 set of hyper-parameters. *(Cross-Validation de 5 particiones : Del dataset total inicial, y usando regularizacion L0, L1 o L2, repetir la clasificacion 5 veces usando cross-validacion con particiones no-superpuestas. Reportar las matrices de confusion en el dataset de entrenamiento, validacion y testing para cada particion (poner los numeros en bruto). Deberias terminar ploteando 15 sub-figuras (3 entrenamiento/validacion/testeo x 5 validaciones). Hacer esto solo para un set de hiper-parametros.)*

Obligatory Please list in your 2 page report : *(Favor agregar en tu informe de dos paginas)*:

(a) The contributions of each author. *(La contribucion de cada autor)*

(b) The list of all the python packages used. *(La lista de todos los paquetes de python y/o otro lenguaje utilizado)*

(c) The list of all toolboxes used (and links to datasets and dataset license) *(La lista de todos los toolboxes utilizados (y links de datasets y licensias de datasets))*

(d) The list of any AI tools (*e.g.* ChatGPT, Perplexity, You) used in your homework and how. *(La lista de todos los asistentes de AI utilizados (ejemplo: ChatGPT, Perplexity, You))*

(e) The list of all academic references used in your homework. *(La lista de todas las referencias academicas en tu tarea)*

(f) Attach a copy of all your code (this may extend 2 pages). *(Agregar una copia de todo to codigo – esto puede extender 2 paginas)*