Data Analytics, a.a. 2020-2021
Proposte di Progetti

Docenti: Prof. Giuseppe Lisanti, Prof. Marco Di Felice
{giuseppe.lisanti, marco.difelice3}@unibo.it

December 11, 2020

Regole generali

- **COME** Il progetto può essere svolto singolarmente o in gruppi di massimo DUE unità.
- **COME** Il progetto può essere svolto su traccia proposte dallo studente (previa approvazione da parte dei docenti) oppure sviluppando le tracce descritte qui sotto. Quest'ultime devono essere considerate delle specifiche di massima; è possibile estenderle/modificarle/customizzarle a piacimento, a patto di non ridurre la complessità.
- **COME** La consegna avviene attraverso la piattaforma VIRTUALE di UNIBO, nella pagina del corso e nella sezione dedicata al Progetto.
- **QUANDO** Sono previste le seguenti deadline di consegna del progetto: 1 Febbraio 2021, 1 Marzo 2021, 1 Aprile 2021, 1 Maggio 2021, 1 Giugno 2021, 1 Luglio 2021, 1 Settembre 2021.
- **COSA** Occorre consegnare **i sorgenti** ed una **relazione** (strutturata nelle seguenti sezioni: Introduzione, Metodologia, Implementazione, Risultati), lingua a scelta (inglese preferibile).
- **COSA** A valle della consegna, occorre preparare una **presentazione con slides** per la discussione del progetto: durante la presentazione, verrà chiesto di effettuare una demo del progetto.
- **NOTA 1** Tutti i membri del gruppo devono essere presenti durante la discussione, e devono conoscere il 100% del progetto svolto. La ripartizione del lavoro deve essere equa tra i membri del gruppo.
- **NOTA 2** A valle della discussione del progetto, è previsto l'esame orale con domande di teoria sugli argomenti trattati durante il corso. Il voto finale viene calcolato come media del progetto e della prova orale.

Specifica del Progetto

Il progetto consiste nella realizzazione di uno studio di analisi di dati, a partire dalla loro acquisizione fino all' estrazione di conoscenza dagli stessi. A tal fine, lo studio deve prevedere la progettazione ed implementazione di una **data pipeline** che includa TUTTE le fasi illustrate durante il corso. Nello specifico, devono essere considerate le fasi di:

- **Data Acquisition**: lettura del dataset e caricamento in memoria attraverso opportune librerie Python.
- **Data Pre-processing**: utilizzo delle tecniche illustrate durante il corso, quali data cleaning, dimensionality reduction, standardization, etc. La scelta delle tecniche da introdurre dipende dal task del processo di analisi (vedi Sezione successiva).
- **Modeling**: costruzione del modello di classificazione/regressione, sulla base del task del processo di analisi (vedi Sezione successiva). Approfondire il processo di tuning degli hyperparametri.
- **Performance analysis**: splitting dei dati (train, validation, test) e confronto delle prestazioni di tecniche differenti e/o di architetture di reti neurali differenti.
- **Visualization**: tracciamento e visualizzazione delle metriche del modello (es. usando TensorBoard[1]) e/o delle caratteristiche salienti del dataset (es. scatter plot matrix).

A seguire, sono illustrate tre tracce di Task di data analytics, con descrizione dei relativi dataset da utilizzare.

TASK 1: Classificazione di misurazioni di sensori

Dataset:	Transportation Mode Detection Dataset (TSD)
URL	http://cs.unibo.it/projects/us-tm2017/index.html
Obiettivo	Classificazione di attività umane
Tecniche	MLP + Tecniche classiche (Random Forest, NB, SVM)

Il dataset contiene dati di crowdensing relativi a misurazioni di sensori embedded dello smartphone (es. accelerometro, giroscopio, magnetometro). Il dataset è annotato con la modalità di trasporto usata in quel momento dall'utente (es. autobus, macchina, bici, etc). L'obiettivo dello studio è riconoscere la modalità di trasporto corrente, a partire da un campionamento istantaneo del valore dei sensori presenti sullo smartphone.
Per l'analisi del dataset, **è possibile prendere spunto dalla pubblicazione** disponibile al link seguente:
http://dl.acm.org/citation.cfm?id=3034534

[1] https://www.tensorflow.org/tensorboard

TASK 2: Classificazione di testo

Dataset:	Quicksign OCRized Text Dataset (QS-OCR)
URL	https://github.com/Quicksign/ocrized-text-dataset
Obiettivo	Classificazione di testo
Tecnica	Recurrent Neural Network (RNN)

Il dataset contiene 400000 documenti testuali, appartenenti a 16 diverse categorie (es. letter, email, scientific report). Il dataset è stato ottenuto da quello del Task 1 applicando tecniche automatiche di rinoscimento del testo, quindi potrebbe contenere molto rumore. L'obiettivo dello studio è classificare un documento, riconoscendo la categoria cui appartiene. Come nel caso precedente, **date le dimensioni del dataset, si consiglia l'utilizzo della GPU per la costruzione del modello**. **In alternativa (GPU non disponibile), si consiglia di utilizzare il dataset "ridotto"**, disponibile allo stesso link. In quest'ultimo caso, si consiglia di usare tecniche di fine-tuning per il training.
Per l'analisi del dataset, **è possibile prendere spunto dalla pubblicazione** disponibile al link seguente:
https://arxiv.org/abs/1408.5882

TASK 3: Classificazione di Immagini

Dataset:	Ryerson Vision Lab Complex Document Information Processing (RVL-CDIP)
URL	https://www.cs.cmu.edu/ aharley/rvl-cdip/
Obiettivo	Classificazione di immagini
Tecnica	Convolutional Neural Network (CNN)

Il dataset contiene 400000 immagini grayscale relative alla scansione di documenti, appartenenti a 16 diverse categorie (es. letter, email, scientific report). L'obiettivo dello studio è classificare una scansione, riconoscendo il tipo di categoria cui appartiene il documento. **Date le dimensioni del dataset, si consiglia l'utilizzo della GPU per la costruzione del modello. In alternativa (GPU non disponibile), si consiglia di utilizzare il dataset "ridotto"**, disponibile a questo URL:
https://www.kaggle.com/patrickaudriaz/tobacco3482jpg
Nel caso del dataset "ridotto", si consiglia di usare tecniche di fine-tuning per il training. Per l'analisi, **è possibile prendere spunto dalla pubblicazione** disponibile al link seguente:
https://arxiv.org/abs/1704.03557

Sono disponibili ulteriori spunti per progetti (e/o tirocini/tesi), contattando direttamente i docenti del corso.