PROGETTO MACHINE E DEEP LEARNING
a.a. 2019/20

Obiettivo del progetto è realizzare un sistema di apprendimento automatico mediante l'utilizzo del linguaggio di programmazione Python.

Ogni studente dovrà analizzare due dataset che appartengono alle seguenti famiglie di dati:
- immagini,
- dati sequenziali di una sola fra le seguenti tipologie: testi, audio.

La tipologia di coppia di dataset da analizzare dovrà essere selezionata dall'elenco disponibile sulla piattaforma didattica.

Per ognuno dei due dataset i task da risolvere sono:
- classificazione,
- semi-supervised anomaly detection.

Classificazione. I modelli da valutare per questo task sono:
- AdaBoost
- SVM
- Reti Neurali
- Stima di densità (uno fra Bayes, Naive Bayes, LDA, GMM, KDE, KNN, altro)

Per ognuna di queste tecniche occorre consegnare almeno un classificatore. È possibile consegnare anche più classificatori appartenenti alla stessa famiglia (ad esempio una rete densa e una rete convoluzionale) se lo si ritiene opportuno.
Si ricorda che è possibile applicare un classificatore binario al problema della classificazione multiclasse utilizzando le strategie One-versus-All e All-Pairs (si rimanda alla sezione 17.1 del libro di testo).

Anomaly Detection. Per questo task si deve usare almeno una fra le tecniche elencate per la classificazione. In questo caso il training set dovrà essere composto dalla classe più numerosa del dataset ed il test set da quelle restanti. Per la valutazione della qualità del detector utilizzare le curve di ROC[1].

La selezione della funzione di loss (esempio MSE, Cross-Entropy, altro), così come di ogni altro elemento che concorre alla definizione del modello (esempio funzione kernel in SVM, funzioni di attivazione e architetture di reti neurali, ottimizzatori, iperparametri, eventuale distanza e altro), fa parte delle scelte progettuali dello studente.

Fra i risultati da presentare deve essere inserita l'accuratezza ottenuta mediante 10-fold cross-validation. Inoltre è richiesta la consegna dei modelli appresi (uno per tipologia) e di uno script che ne calcola l'accuratezza su un validation set codificato nello stesso formato dei dati di partenza.

Si dovrà produrre una relazione scritta, indicativamente di una decina pagine, a cui va allegato il codice. La relazione e il software dovranno essere consegnati entro 2 giorni dalla data della discussione e potranno essere integrati in sede di discussione. Per la presentazione dell'attività progettuale, che avrà la durata di 20 minuti, è possibile avvalersi di slides.

1 Si veda ad esempio https://en.wikipedia.org/wiki/Receiver_operating_characteristic.