Pràctica 1 (XNDL): Perceptró Multicapa

Daniel Hinjos, Anna Arias

1 Introducció

El propòsit d'aquesta pràctica és que sigueu capaços d'abordar un cas més realista al camp de l'aprenentatge automàtic i profund. La vostra missió serà, en primer lloc, treballar amb un conjunt de dades específic, on haureu de realitzar una anàlisi exploratòria de dades (EDA) i idear una estratègia de tècniques de preprocessament adequada al context. En segon lloc, haureu de dur a terme un procés de modelatge utilitzant un model linear base i una xarxa neuronal de perceptró multicapa (MLP), sobre la qual haureu de treballar iterativament. L'objectiu és comprendre l'estructura i les característiques inherents al conjunt de dades, alhora que construïu i perfeccioneu un model neuronal capaç d'abordar la tasca elegida.

2 Dataset

En primer lloc, haureu de triar un conjunt de dades dels que us proporcionem a la carpeta data. Si voleu triar un altre, el que escolliu haurà de tenir les següents característiques:

- Haurà de tenir variables numèriques i categòriques.
- Haurà de contenir més de 10 variables.
- No haurà d'estar preprocessat, és a dir, serà necessari que es puga fer un preprocessament sobre aquest conjunt de dades.
- Haurà de contenir més de 500 mostres.
- El conjunt de dades no haurà sigut generat sintèticament
- No podrà ser un problema simple utilitzat habitualment o ja resolt multitud de vegades a internet.
- El problema haurà de ser prou complex perquè no obtinga un encert o un R2 gairebé perfecte usant una regressió lineal/logística.

3 Passos

Aquests seran els diferents passos que haureu de realitzar i reportar a l'informe que haureu de lliurar.

1. Anàlisi Exploratoria de Dades (EDA).
    - Fer un estudi estadístic de les dades per comprendre les variables.
    - Visualitzar les dades utilitzant gràfics adequats per identificar patrons, relacions i possibles anomalies.

- Explorar correlacions entre les variables i la seua importància per a la tasca en qüestió.

2. Estratègia de preprocessament.

   - Idear una estratègia de preprocessament adequada, tenint en compte, com a mínim, recodificació de variables numèriques, tractament de valors perduts, tractament de valors anòmals i normalització.

3. Remostreig.

   - Dividir el conjunt de dades en conjunt d'entrenament, test i validació (si escau) triant un percentatge.
   - Utilitzar tècniques de validació creuada per avaluar el rendiment dels models de manera robusta.

4. Model lineal base.

   - Entrenar i avaluar un model de regressió lineal o regressió logística segons la natura del problema (regressió o classificació).
   - Interpretar els resultats obtinguts (mètriques de classificació/regressió, coeficients, etc.).

5. Procés iteratiu - Perceptró Multicapa (MLP).

   En aquesta secció s'haurà de superar el model lineal base. Per a cada iteració:

   - Diagnosticar la situació actual del model en funció de les eines de diagnòstic (corbes de pèrdua, mètriques de rendiment, matrius de confusió, *model summary*). Entre els possibles diagnòstics estan l'*underfit*, *fit*, *overfit*, inestabilitat del procés d'aprenentatge, manca de convergència, convergència excessivament ràpida/lenta, comportament aleatori, etc.
   - Proposar una millora potencial en funció del diagnòstic anterior, com per exemple canvis a l'arquitectura del model o ajustament d'hiperparàmetres.
   - Experimentar amb la millora seleccionada i avaluar-ne l'impacte en el rendiment del model.

   Aquest procés es podrà repetir un total de 4 vegades. Per a la primera iteració, proveu només amb una sola capa, perquè tingueu marge de millora.

6. Model guanyador i conclusions.

   Fer una taula comparativa amb tots els models testejats, explicant el model que creieu que millor aborda al problema que esteu intentant resoldre. Finalment, detalleu quins són els problemes que heu trobat així com les conclusions que extraieu.

4 Lliurables

- Un informe que incloga els resultats de l'EDA, la justificació de les estratègies utilitzades de preprocessament, el procés i les decisions preses durant l'entrenament, l'avaluació dels models així com les conclusions extretes. El document tindrà una extensió màxima de **10 pàgines** (incloent-hi portada, índex i referències).

- Codi utilitzat per fer l'anàlisi i el modelatge, en un entorn com Jupyter Notebook o Google Colab.

5 Data de lliurament

Aquesta pràctica es realitzarà en parelles i s'hauran d'acomplir les tasques de forma col·laborativa. La data límit per lliurar el treball dut a terme serà el 5 de maig de 2024 fins a les 23:59h.

6 Rúbrica d'avaluació

Criteri	Satisfactori (10-9)	Millorable (8-5)	Malament (4-0)
Document **2.5 punts**	El document té una introducció, desenvolupament i conclusió. No supera l'extensió màxima. Presenta les idees de manera lògica i amb coherència. No hi ha errors ortogràfics ni sintàctics. El llenguatge utilitzat és acurat. Es fan servir elements visuals que ajuden a la comprensió.	El document té una introducció, desenvolupament i conclusió. No supera l'extensió màxima. Les idees es presenten en un ordre lògic, generalment amb coherència i fluïdesa. Hi ha alguns errors ortogràfics o sintàctics. El llenguatge utilitzat és just. S'utilitzen elements visuals, però no sempre de manera apropiada o no sempre ajuden a la comprensió del treball.	El document no conté introducció, desenvolupament o conclusió. Supera l'extensió màxima. Hi ha errors ortogràfics i sintàctics. El document utilitza un llenguatge poc precís. No s'usen elements visuals.
Estudi del conjunt de dades. **1 punt**	Es fa un estudi estadístic de les dades, es visualitzen les distribucions de les variables, s'analitzen raonadament les seues relacions i la seua natura.	Es calcula l'estadística descriptiva bàsica de les dades. Les visualitzacions o l'anàlisi són pobres.	L'estudi es limita a descriure la tipologia de dades, sense cap mena d'anàlisi.
Preprocessament de les dades **2 punts**	El preprocessament de dades és adequat al conjunt de dades, i s'adapta correctament a les particularitats de les dades. Cada decisió al respecte està degudament explicada al document. El preprocessament s'aplica adequadament a les diferents particions de dades.	S'ha fet algun tipus de preprocessament. No s'expliquen les decisions sobre el preprocessament.	No s'ha realitzat cap preprocessament de les dades o el preprocessament realitzat no té sentit. El processament previ s'aplica incorrectament a les diferents particions de les dades.
Ús de les dades **1 punt**	S'han usat les tres particions de les dades correctament.	Alguna de les particions no ha estat utilitzada per al seu propòsit.	Les particions han estat utilitzades de manera que els resultats i/o el model no són fiables.
Experiments amb el model lineal base **1 punt**	S'ha fet servir un model lineal de regressió o classificació. S'analitzen i s'interpreten els resultats obtinguts.	S'ha fet servir un model lineal però no de manera adequada o no s'analitzen els resultats.	No es fa servir un model lineal.
Cicles i tria del model **2.5 punts**	S'han realitzat els experiments de manera coherent. Tant els diagnòstics, les propostes de millora com les experimentacions són coherents, estan raonades i completes.	Els experiments són excessius o incomplets. Els diagnòstics, les millores o les experimentacions no estan completament raonades, són parcialment incoherents o contradictoris.	Tant el nombre d'experiments com el contingut és pobre o no és adequat.

Taula 1: Rúbrica.