IAA 25/26 Q1 Tardor - Laboratori

November 10, 2025

Abstract

Enunciat del laboratori, individual. Aquest s'ha d'entregar en format document de text (**no més de 30 pàgines**), abans del 28 de Desembre a les 11:59. Incloure figures i taules al document. També s'ha d'entregar un fitxer comprimit amb el codi necessari per replicar els resultats de cada secció (marcar clarament quin tros de codi correspon a quina secció). La descripció de cada secció conté un estimat (podeu fer més o menys) del contingut de suport visual (figures, taules), que ha d'anar acompanyat d'una explicació textual. També referencies a les seccions de codi (fitxers i/o línies) responsables de cada secció.

Cal argumentar de manera explícita totes les decisions de rellevància preses sobre les dades i el model. La pràctica s'avaluará sobre les explicacions i justificacions aportades, NO sobre el rendiment final del model. D'igual manera, totes les figures i taules han d'estar explícitament comentades al text, han de contindre informació dels eixos i una *caption* descriptiva.

El document ha d'incloure de la Secció §1 a la Secció §5. La secció de bonus és opcional.

Tots els dubtes metodològics seran resposts a classe, a hores de consulta i per correu (*e.g.*, *te sentit que faci això?*). Els dubtes tècnics (*e.g.*, *perquè aquest codi no fa el que vull que faci*) seran resposts només durant les classes, mai per correu.

Objectiu

El propòsit d'aquest laboratori és crear un model que, basant-se en paràmetres de laboratori, marcadors genètics, característiques de neuroimatge i demogràfics, predir si un pacient té esquizofrènia resistent al tractament (TRS) o esquizofrènia no resistent al tractament (NTRS). La base de dades d'entrenament té un total de 9.000 observacions que podeu trobar en el fitxer: *trs_train.csv*. Tracteu la variable *TRS* com a variable objectiu. Farem servir un nou conjunt de dades per evaluar la generalització del vostre model en una competició de Kaggle: El fitxer *trs_eval.csv*, que conté un conjunt d'evaluació sense variable objectiu *TRS*.

Teniu més informació sobre els paràmetres en el següent article. També teniu més informació en el link de la competició de Kaggle. **Els equips de Kaggle seran com a màxim de dos estudiants. Cada estudiant haurà de crear un compte de Kaggle i fer un merge d'equip amb el seu company. S'ha de fer servir el correu d'estudiantant de la UPC.**

1 Anàlisis i preprocessat de dades

- Anàlisi estadístic de les variables de manera independent. Comentaris i discussió sobre els resultats. Al notebook, podeu crear una taula (mean, var, min, max, ...) i el codi per estudiar les distribucions per variable. A l'informe només cal detallar un exemple d'una o dues variables que us sembli interessant comentar.

- Estudi de balanceig de classe objectiu. 1 Figura (histograma amb freq. per classe) en els conjunts de train i test donats. En aquest punt, decidim si fem servir algun mètode per balanceig de classes. Justificar aquesta decisió i detallar les conseqüències.

- Missings. Identificació i proposta de gestió. Si es fa imputació cal fer-ho després del particionat de dades.

- Outliers. Identificació i proposta de gestió, si cal.
- Recodificació de variables, si cal. Justificar la motivació.
- Cal particionar el dataset de train en noves particions? (Recorda que el dataset de eval no es pot fer servir per entrenar i que es farà servir només per crear un submission.csv, més informació en Kaggle). Incloure una taula a l'informe amb la mida de les diferents particions i justifiqueu la motivació. Teniu en compte que:
  - Si imputeu missings heu de particionar abans de la imputació.
  - Si feu servir un mètode de balanceig heu de particionar abans de balancejar les classes.

2 Preparació de variables

- Normalització de variables.
- Anàlisi de variables categóriques i variable objectiu. (1 Figura per variable categórica i variable objectiu)
- Eliminació de variables numèriques redundants o sorolloses fent servir la correlació i tenint en compte la tasca objectiu.
- Estudi de dimensionalitat amb PCA. ¿És necessari reduir variables? (1 Figura amb la variança explicada i número de dimensions). Justificar la motivació.

3 Definició de models

Entrenar 3 models: Un SVM, un XGBoost i una *custom* regressió logística. Pel cas de la *custom* regressió logística, heu de crear una classe logistic regression (estil scikit-learn) que faci servir descens de gradient fent servir minibatch.

- Definició de metriques d'entrenament i validació de generalització. Motivació.
- Entrenament:
  - Incloure figures/taules del resultats de l'aprenentatge.
  - Discussió dels hiperparàmetres disponibles, i dels valors usats. Incloure una taula amb la llista d'hyperparametres i valors provats. En el cas de la regressió logística *custom*, fer un estudi de la mida del batch: Incloure una figura amb el rendiment en funció de la mida del mini-batch.
  - Anàlisi de resultats i iteracions:
    - Nova preparació de variables i/o selecció (secció 2)
    - Selecció d'hiperparàmetres.
    - Inferencia amb validació. Taules amb resultats, per valor d'hiperparametre, i observació del nivell d'overfitting o underfitting.
- Resultat final obtingut. Incloure una taula comparativa del rendiment per els millors classificadors.

4 Selecció de model

- Motivació del millor model triat. Descripció del model triat. Si és més d'un, argumentar perquè.
- Característiques desitjables respecte al problema (complexitat, interpretabilitat, hiperparámetres, volum de dades, *etc.*).
- Anàlisi de les limitacions i capacitats del model.
- Generar prediccions sobre la partició d'evaluació (Kaggle submission.csv) i pujar el fitxer a la competició. Comparar els resultats amb el resultat obtingut en el conjunt d'entrenament. El notebook ha d'incloure una secció final on s'entreni el model triat i generi el fitxer submission.csv

5 Model Card

Documentació del model seguint l'estructura d'una model card.

Bonus 1

Per aquells estudiants que assoleixin aquest punt, i vulguin fer un pas extra, poden entrenar un model de EBM (sense fer servir PCA) i comparar amb els models anteriors. (1 Taula amb mètriques + 1 Figura amb les variables més importants en train/test).

Bonus 2

Pots guanyar la competició utilitzant tècniques d'aprenentatge alternatives? Si decideixes provar altres models, inclou en el notebook el millor model i la generació del seu submission.csv