IAA 23/24 Q1 Tardor - Laboratori 1

November 27, 2023

Abstract

Enunciat del laboratori, individual. Aquest s'ha d'entregar en format document de text, 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. Les seccions de bonus son opcionals.

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 fer un model que, donat 17 característiques clíniques predigui la supervivència del pacient amb cirrosi hepàtica. Tracteu la variable *Status* com a variable objectiu. Podeu descarregar el dataset de https://archive.ics.uci.edu/dataset/878/cirrhosis+patient+survival+prediction+dataset-1

1 Anàlisis i preprocessat de dades

- Anàlisi estadístic de les variables de manera independent. Comentaris i discussió sobre els resultats. 1 Taula (mean, var, min, max, ...) + 1 Figura (distrib.) per variable.
- Estudi de balanceig de classes. 1 Figura (histograma amb freq. per classe). 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. 1 Taula (num. misings per var). Si es fa imputació cal fer-ho després del particionat de dades.
- Outliers. Identificació i proposta de gestió, si cal. 1 Figura per variable amb outliers amb la distribució amb i sense outlier.
- Recodificació de variables, si cal. Justificar la motivació.
- Particionat del dataset en train/test o train/validació/test. A partir d'aquest punt, no useu la partició de validació (si heu decidit fer servir la partició de validació) fins l'entrenament de models. La partició de test no la podeu fer servir fins l'avaluació del model final. 1 Taula (mida de les particions).
  - 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. 1 Figura per variable (distrib. després de variable).
- Anàlisi de correlacions entre variables númeriques. 1 matriu (correlació per parella).
- Anàlisi de variables categóriques i variable objectiu. (1 Figura per variable categórica i variable objectiu)
- Eliminació de variables redundants o sorolloses, 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 KNN, un arbre de decisió i un SVM.

- Definició de metriques. Motivació.
- Motivació del primer model triat. Característiques desitjables respecte al problema (complexitat, interpretabilitat, hiperparámetres, volum de dades, *etc.*).
- Discussió dels hiperparàmetres disponibles, i dels valors usats. 1 Taula (llista d'hyperparametres i valors provats).
- Primer entrenament amb train. Figures/Taules. (resultats del aprenentatge)
- Anàlisi de resultats, i iteració:
  - Preparació de variables (secció 2)
  - Selecció d'hiperparàmetres.
  - Inferencia amb validació. Taules amb resultats per train i val, per valor d'hiperparametre, i observació del nivell d'overfitting.
- Resultat final obtingut per el primer model (train i val). 1 Taula (resultats).
- Motivació del segon model triat.

4 Selecció de model

- Descripció del model triat. Si és més d'un, argumentar perquè.
- Anàlisi de les limitacions i capacitats del model.
- Resultats en partició de test, en comparació amb train i val. 1 Taula.
- Model card.

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 i comparar amb els models anteriors. (1 Taula amb mètriquees + 1 Figura amb les variables més importants en train/test).

Bonus 2

Fer anàlisi no supervisat de les dades. Traient la variable *"Status"*, podem identificar clústers? A què corresponen? Ens poden servir per quelcom en la nostra tasca? I fora de la nostra tasca?