IAA 22/23 Q1 Tardor - Laboratori 2

November 21, 2022

Abstract

Enunciat del segón laboratori, individual. Aquest s'ha d'entregar en format document de text, abans del 23 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).

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.

Contexte

Disposem de dades d'una empresa proveïdora de telefonia i internet. En particular tenim dades dels clients seus clients en un periode de temps, incloent els que han marxat. L'empresa vol saber:

- Quin tipus de persona marxa?
- Que es pot fer per evitar-ho?
- Quina gent no marxa?
- Amb quina precisió podem predir la gent que marxa?

1 Anàlisis i preprocessat de dades: Profiling

- Anàlisis de correlació de les variables. Incloure matriu. Proposta de merge/delete si s'escau.
- Anàlisi estadístic de les variables de manera independent. Incloure: Distribució, rang, outliers, missings. Proposta de recodificació si s'escau.
- Anàlisi de riscos i biaixos de les variables. Proposta de eliminació o vigilància (si no es vol eliminar) per aquelles variables on calgui.

2 User profiling

- Estudi de diferencia entre la població que marxa, i la que no, per variable. Incloure comparacións de distribucions.
- Descripció del client típic que marxa i del que no, i recomanacions per la companyia.

3 Preprocessat de dades: Classifier

- Estudi de balanceig de respecte a la variable objectiu.
- Definició del particionat (train-val/test). El test ha d'estar perfectament balancejat respecte a la variable objectiu.
- Definició de l'estrategia per mitigar el desbalanceig en train i val. Undersampling, oversampling, cross-validation.
- Normalització de dades, basada en train-val.

4 Classificador

Per cadascun dels models entrenat a continuació, descriure quines característiques son desitjables per al problema i quines no. Discussió dels hiperparàmetres disponibles, i dels valors usats.

- Definició de mètriques.
- Entrenament d'un KNN.
- Entrenament d'un arbre de decisió.
- Entrenament d'un random forest o XGBoost (opcional)
- Entrenament de dues SVMs (una lineal, una radial)
- Comparació final (amb val i test) i proposta de model.

5 Classifier: Model Card

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

6 Clustering

Usar mínim dos mètodes de clustering d'entre k-means, hierarchical clustering i DBSCAN. Per els resultats obtinguts amb cadascun, visualitzar els clústers usant alguna tècnica de reducció de dimensionalitat. Discussió dels resultats del clústering.