Facultat d'Informàtica de Barcelona

FIB UPC
Xarxes Neuronls i Deep Learning

Pràctica 1 (XNDL): Perceptró Multicapa

Alumnes :
Granja Bayot, Jordi
Gutierrez Kitajima, Luis Kazuto

Tutors :
Hinjos, Daniel
Arias, Anna
Belanche, Luis Antonio

May 7, 2024

Contents

1 Introducció

Aquest document presenta la solució a l'anàlisi i modelatge d'un conjunt de dades real de precipitacions a Austràlia. La nostra tasca inclou una anàlisi exploratòria de dades, estratègies de preprocessament i l'ús de models lineals i xarxes neuronals per a la predicció de pluja. L'objectiu és comprendre tant les característiques del conjunt de dades com construir un model neuronal efectiu per a abordar aquesta tasca. Adjuntem juntament amb aquest document notebooks amb gráfics que complementen les explicacions.

2 Càrrega, neteja i split

El conjunt de dades escollit ha sigut el *rain_ data*. Una primera neteja ha consistit en detectar la presència d'instàncies les quals tenien la columna objectiu *RainTomorrow* absent. Ja que aquestes instàncies no són d'utilitat, s'han eliminat de primera mà.

Tot seguit, hem realitzat el **split**. En el nostre cas i per les aplicacions pel qual el volem, hem creat tres particions: train, test, i val. Els percentatges són 70%, 15%, 15%. A més a més, tot i que la columna objectiu està bastant equilibrada, hem estratificat les mostres segons la columna *RainTomorrow*. Hem pres la decisió de dur a terme la partició de dades a aquestes altures, ja que és una bona pràctica per a evitar el *data snooping*, on, per tant, ens estalviarem la visualització de patrons en dades que en la pràctica no tindríem. Reservem el test per a mostrar els resultats de l'únic model escollit, i el validation per a tria d'hiperparàmetres i altres configuracions.

3 EDA

En aquest apartat discutirem l'exploració de les dades feta prèviament al modelatge. L'objectiu recau en l'estudi de la natura de les variables, característiques estadístiques, així com les interaccions entre variables. Una gran part d'aquesta exploració estarà condicionada respecte a la variable objectiu *RainTomorrow*. Les explicacions a continuació, es basa en la síntesi de les taules i plots creats i mostrats al Jupyter Notebook adjunt.

3.1 Anàlisi estadística

3.1.1 Variables numèriques

Pel que fa als màxims i els mínims, després de ser comprovats, estan dins dels màxims i mínims històrics a la regió d'Austràlia, fet que ens facilitarà la feina de detecció d'outliers. Trobem màxims i mínims molt semblants en les variables de pressió. Sembla que tenen un domini molt reduït.

Les variables *Rainfall* i *Evaporation* tenen uns màxims molt allunyats de la mitjana, i quantils amb valors molt baixos i propers. Tot sembla que tenen una natura *Power-Law*. Possiblement, haurem de tractar-les en un futur.

Finalment, sembla que les variables de temperatura i núvols tenen una dispersió relativament alta. On a més aquest últim grup de variables sembla que podrà ser multimodal per la forma dels quantils. Tot i això, es revisarà en els gràfics univariants.

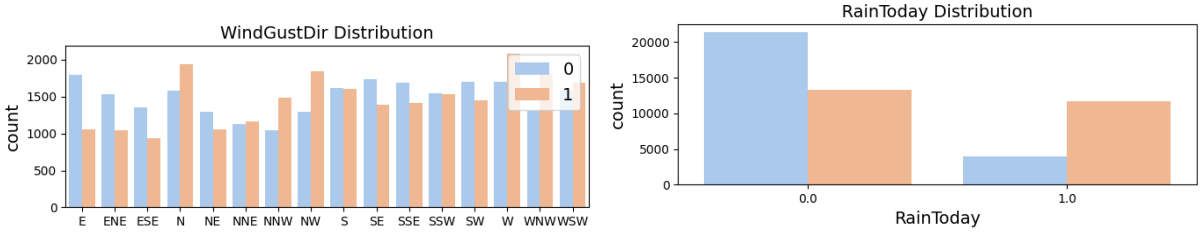
3.1.2 Variables categòriques

En primer lloc, ens hem fixat en els valors únics, ja que poden dificultar l'aprenentatge. Les variables *Date*, *Location*, *WindGustDir*, *WindDir9am*, i *WindDir3pm* presenten moltes categories. Fet que haurem de tractar posteriorment a l'extracció de característiques. D'altra banda, hem vist com totes les categories estan prou equilibrades exceptuant *RainToday*. Veurem aquest equilibri amb la descriptiva univariant.

3.2 Estudi univariant

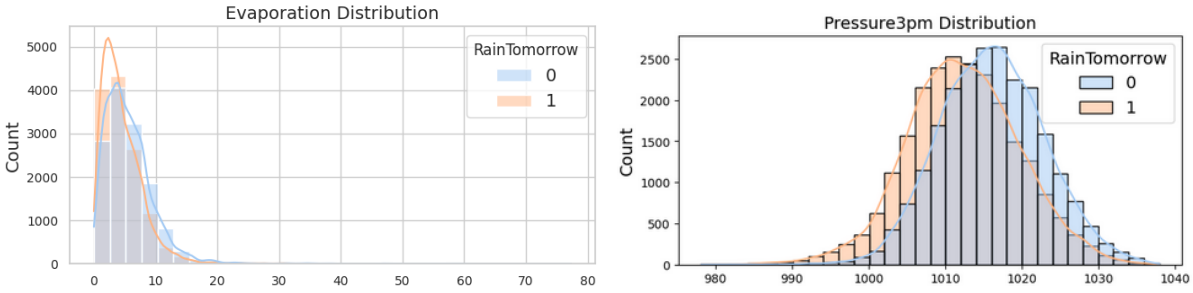
En aquest apartat, discutirem l'anàlisi univariant de les variables, així com la seva distribució condicionada a la variable objectiu. Tenint en compte la restricció de longitud del document, mostrarem els gràfics condicionals, que permeten una visió tant del comportament univariant com de l'estructura general. És rellevant assenyalar que prèviament hem estratificat les dades respecte a la variable resposta *RainTomorrow*, obtenint una proporció equilibrada de 50/50.

En el cas de les variables categòriques, la variable *Date* mostra valors extrems al principi, probablement a causa de la baixa quantitat de mostres en els valors inicials. Pel que fa a l'estacionarietat, aquesta variable generalment oscil·la al voltant dels mateixos valors, tot i que es pot observar una certa estacionalitat que podrà ser útil en anàlisis futures, particularment si es considera una variable que representi l'estació. Respecte a *Location*, la qual ha estat representada en una taula (per la quantitat de categories), sembla mostrar una certa importància per a la predicció. Per acabar, cal esmentar les tres variables relacionades amb el vent, així com *RainToday*. Com que les tres variables de vent tenen un comportament similar, en representarem només una:



Podem observar que en el cas del vent tenim diferències entre categories, de la mateixa manera que amb *RainToday*. En aquesta segona podem veure clarament com si avui no ha plogut, demà és menys probable que plogui. D'altra banda, si avui ha plogut, demà és més probable que plogui.

Quant a les variables numèriques, tenim dues principals natures de distribucions: normal i Power-Law, on a la segona categoria cauen les variables *Rainfall* i *Evaporation* com hem dit anteriorment, i a la primera totes les altres:



Un cop vistes totes les distribucions presents al Jupyter Notebook, podem dictaminar que les variables numèriques més influents de cara a predir la pluja són: **Sunshine**, **Humidity**, i **Clouds**. Altres variables moderadament influents són: **Pressure**, **Temp3pm** o **WindGustSpeed**. A més a més, hi ha variables com *Rainfall* o *Clouds* que tenen molts valors a 0 (pels dies que no plou), o variables amb certa multimodalitat com *Cloud9am* i *Cloud3pm*.

3.3 Estudi d'interaccions i correlació

Pel que fa a l'estudi d'interaccions entre múltiples variables, hem explorat diverses combinacions, destacant especialment la relació entre els núvols, el vent i la seva direcció. Tot i això, l'anàlisi no ha revelat resultats significatius.

Quant a les correlacions, hem realitzat tant el coeficient de correlació de Pearson com el de Spearman, amb l'objectiu de detectar tant correlacions lineals com no lineals.

L'anàlisi basada en les matrius de correlació mostrades al Jupyter Notebook indica que les temperatures no només mostren una alta correlació entre les seves mesures a diferents hores, sinó també amb la temperatura mínima i màxima. A més, la pressió en diferents moments presenta una correlació directa d'1.

D'altra banda, les característiques del vent mostren una alta correlació entre elles. Finalment, els núvols i la humitat presenten una correlació elevada amb l'índex sol diari, això sí, amb una relació negativa.

Les fortes correlacions presents al nostre conjunt de dades ens portaran a provar mètodes com **PCA**, ja que pot ajudar a complir les assumpcions que fan els models sobre les dades, així com millorar resultats en alguns dels casos.

4 Preprocessing

4.1 Encoding

La primera variable que s'ha hagut de recodificar és la variable *Date*, que s'ha passat a format *Datetime*. Tot seguit, les variables binàries *Raintoday* i *RainTomorrow* s'han convertit a 0 i 1 per al funcionament dels algorismes.

4.2 Missing Values

L'objectiu sería imputar les nostres dades de manera que no variï massa les distribucions originals, evitant eliminar variables. Per poder realitzar aquesta tasca cal tenir en compte

que imputar valors en una instància amb massa dades faltants podria introduir una quantitat significativa d'incertesa i potencialment distorsionar les relacions dins de les dades. Per tant abans de fer la imputació caldria desfer-nos d'aquestes instancies.

Hem considerat que tenir 7 valors mancants sería considerat massa, per tant després de aquesta eliminació, ens queden **44.284** instàncies que només contenen 6 o menys valors mancats cada una, permetent una imputació que preserva la integritat i qualitat de les dades. Per els valors mancants que encara romanen a la base de dades, els imputem de la següent manera:

- **Variables Numeriques:** Entrenem un model de Imputador MICE amb el que imputem les dades faltants. Aquest imputador fara ús dels veïns més propers (KNN) per predir els valors a imputar, iterant fins convergencia.
- **Variables Categóriques:** S'ha considerat que els valors mancants en les variables categòriques podrien tenir un significat ocult, com ara la falta de registre per raons meteorològiques. Per tant, s'ha creat una nova categoria *missing* per evitar-ne la pérdua d'aquesta informació

4.3 Outliers

La detecció d'outliers s'ha fonamentat en dos passos: detecció d'anomalies, i anàlisi multivariant.

La detecció d'anomalies s'ha realitzat mitjançant l'observació de màxims i mínims per variable en el cas de les variables numèriques, i en la cerca de categories no existents per les variables categòriques. En el cas de les variables numèriques tots els valors eren vàlids (dins els barems històrics d'Austràlia). En el cas de les categòriques no hi havia cap valor fora de l'habitual.

D'altra banda, s'ha fet l'anàlisi multivariant. S'ha escollit fer-ho així, ja que de forma univariada no hi ha cap valor fora del normal, on a més a més retallar distribucions diferents de manera igual (mitjançant mètodes com per exemple la distància interquartil) no seria correcte, pel fet que no compliriem les assumpcions.
El primer pas ha sigut representar les dades en baixes dimensions mitjançant la tècnica *UMAP*, on després hem entrenat un *IsolationForest* perquè predigués quines instàncies eren outliers.
El resultat mostrat al notebook, ens mostra com no podem diferenciar outliers com a tal dins d'aquest subespai vectorial, encara més, podem observar com el IsolationForest treu unes 6000 instàncies que semblen estar dins del núvol de punts. Per tant, hem decidit no eliminar outliers.

4.4 Transformacions

Per a entrenar els diferents models, s'ha hagut de transformar les variables perquè fossin numèriques i estiguessin en la mateixa escala. En la primera de les condicions, s'ha aplicat *OneHotEncoding* a les variables categòriques amb més de dues categories. Les binàries han estat reduïdes a 0 i 1 segons un *OrdinalEncoder*. Això és degut al fet que en ser binària, un dummie d'una de les categories és prou per expressar la distribució.

Finalment, s'ha aplicat el *StandardScaler* a totes les variables numèriques perquè tinguessin un domini comú. Fet que ens permet aprendre i treure conclusions amb més facilitat.

5 Feature Engineering

En ordre de millorar el rendiment del model, manipularem les variables de ***Date, Location, WindGustDir, WindDir9am, WindDir3pm*** per extreure noves que capten la mateixa informació de manera més eficient.

5.1 Feature extraction

- **Date:** Durant l'EDA, vam notar una estacionalitat en la variable resposta. Per capturar aquesta tendència sense dependre de la variable temporal, vam crear les categories Months i Season. Això ens va permetre observar que hi ha més pluja durant els mesos de juny, juliol i agost (estació d'estiu).
- **Wind Directions:** Aquestes variables utilitzen direccions de la brúixola, però la seva estructura no captura les relacions cícliques. Per abordar això, s'han creat les variables X i Y que representen les coordenades polars equivalents a les direccions de la brúixola. Això permet una representació més precisa de la direcció del vent.
- **Location:** La variable Location, amb 49 ciutats diferents, s'ha simplificat en una nova variable anomenada Region, amb 9 categories corresponents a les regions d'Austràlia. Aquesta nova variable ajuda a agrupar les observacions i reduir la dimensionalitat del model.

5.2 Feature selection

Hem decidit utilitzar totes les variables disponibles a la base de dades excepte Date, Location, WindGustDir, WindDir9am, WindDir3pm i Month. Aquestes s'exclouen perquè s'han creat altres variables més eficaces per explicar la mateixa informació de manera més precisa. La variable Month es descarta ja que la variable Season ofereix una representació més adequada de l'estacionalitat de la sèrie temporal, amb menys categories, rendint l'entrenament del model més eficient.

6 Modelling

6.1 Model lineal base

En aquesta secció, utilitzarem les dades preprocessades per a entrenar un model de regressió logística. Mitjançant aquest model, buscarem comprendre i predir els factors que influeixen en la predicció de pluja a Australia (**RainTomorrow**). Aprofundirem en el procés de construcció i avaluació del model.

6.1.1 Hiperparàmetres

Els hiperparàmetres que hem fet servir per entrenar el nostre model de regressió logística són ***ratio de L1*** (proporció de regularització L1 aplicada) i ***C*** (força de la regularització global del model). Per tal d'entrenar un model amb els hiperparàmetres òptims, emprem el nostre conjunt de dades de validació per dur a terme una **validació creuada**. Mitjançant aquest procés, identifiquem els hiperparàmetres que minimitzen el valor de pèrdua de la funció **Cross-Entropy** per a les nostres dades de validació.

Els millors hiperparámetres que trobem amb la validació creuada és de **0.8431633** per **Ratio de L1** i **0.6440911** per **C**. Amb aquests hiperparàmetres, fem servir el nostre conjunt de dades de entrenament per entrenar el model de regressió logística.

6.1.2 Anàlisi dels resultats

Amb el nostre conjunt de Test fem prediccions. Per evaluar el rendiment del nostre model fem servir les métriques de classificació:

Classe	**Precisió**	**Recall**	**F1-score**	**Support**
No Plou	0.7814	0.7959	0.7886	4782
Sí Plou	0.7920	0.7773	0.7846	4782
Accuracy global	0.7866			
Precisió mitjana (macro)	0.7867			
Precisió mitjana (ponderada)	0.7867			

L'anàlisi de les mètriques revela que el nostre model exhibeix una capacitat equilibrada per classificar tant les instàncies de pluja com les que no plou, la qual cosa suggereix una absència de biaix cap a alguna classe específica. Aquest equilibri és crucial per garantir la fiabilitat de les nostres prediccions. A més, l'alt valor de l'AUC (veure Notebook) de **0.87** indica una excel·lent capacitat discriminatoria del model entre les classes.

Pel que fa al context específic de predicció de pluja, prioritzem la capacitat del model per identificar correctament els casos de pluja. Observem un recall de **0.7773** per a la predicció de pluja, la qual cosa, tot i no ser perfecta, demostra una utilitat potencial en aplicacions professionals.

Finalment, també és d'interès explorar els pesos atribuïts a les diferents variables pel model, amb l'objectiu de comprendre com impacta cada variable impacta en la predicció de la pluja. Trobem que les variables que més impacten són **Sunshine, WindGustSpeed, Humidity3pm, Pressure9am, Pressure3pm, Region, RainToday**.

6.2 Perceptró Multicapa

En aquest apartat discutirem el procés iteratiu realitzat per a arribar al perceptró multicapa escollit per modelar en nostre problema. El procés iteratiu ve motivat per el diagnòstic i correcció de models amb l'objectiu de millorar els resultats.

6.2.1 Condicions de l'experimentació.

En primer lloc i com hem introduit anteriorment, hem realitzat un anàlisi de components principals (PCA) per a descorrelacionar les nostres variables. Com es pot observar en el

primer model que presentarem, aquesta transformació ha portat a una millora significativa en el rendiment del model. Amb l'objectiu de garantir comparacions vàlides, hem establert un learning rate i un batch size comuns com a punt de partida, tot i que cal assenyalar que la modificació d'aquests paràmetres podria influir en la millora del model, així com canviar d'una iteració a l'altra. A més, cal destacar que tot el procés s'ha dut a terme amb una llavor fixada per a controlar la variabilitat aleatòria.

Pel que fa a l'ús dels conjunts de dades, hem emprat el conjunt de validació per diagnosticar els models a les corbes d'aprenentatge. A més, per a garantir una avaluació més robusta, hem complementat això amb la validació creuada. Els gràfics es troben al notebook adjunt.

6.2.2 Iteració 1

La primera iteració implica un model simple amb una sola capa oculta que conté 8 neurones amb activació ReLu, i la capa de sortida utilitza activació sigmoide per a la classificació. A partir de les corbes de pèrdua i les mètriques de rendiment en els conjunts de test i entrenament, es pot observar un ajustament satisfactori, amb una diferència entre les corbes mínima. També s'aprecia una lleugera inestabilitat en el procés d'aprenentatge, possiblement a causa de l'ús d'un batch size de 32. Tot i això, s'ha considerat que un batch size petit podria ser més adequat, ja que permet capturar patrons més detallats i evitar mínims locals. Dins del comportament de l'entrenament, també es nota un procés d'aprenentatge relativament lent a partir de les primeres iteracions.

Amb l'esperança de millorar el rendiment, i considerant que les corbes de entrenament i validació estan molt properes, s'ha diagnosticat el model com un cas lleu de underfit. El següent pas per millorar consistirà en augmentar la capacitat del model.

6.2.3 Iteracó 2

A la segona iteració, el model s'ha ampliat. Ara tenim una capa oculta que consta de setze neurones. Analitzant les corbes i les mètriques,es fa evident un clar **overfitting**. Malgrat que es manté la inestabilitat esmentada anteriorment, observem una convergència més ràpida.

Encara que s'ha produït una millora lleugera en el model, es diagnostica com a overfitting. Per tant, el següent pas serà dividir aquestes neurones en dues capes, amb l'objectiu de millorar la capacitat de generalització i captar patrons seqüencials en les dades.

6.2.4 Iteració 3

El tercer model consta de dues capes ocultes, una amb 8 neurones i l'altra amb 4. Després de modificar el model, observem una altra millora significativa. A més, el diagnòstic del model anterior ha facilitat el desenvolupament d'un model que generalitza de manera més efectiva, indicant la possibilitat d'un underfit. No es detecta cap component aleatòria ni inestabilitat.

Amb l'esperança d'incrementar encara més la capacitat després d'identificar el subajust, preveiem afegir més neurones i utilitzar tècniques de regularització per aconseguir un bon ajust.

6.2.5 Iteració 4

La quarta i última iteració presenta un model compós per una capa oculta de 90 neurones, seguida d'una capa de Dropout amb un ratio de 0.25. Després, s'incorpora una altra capa oculta de 90 neurones, seguida d'una capa Dropout amb un ratio de 0.5. Finalment, es troba la capa de sortida amb funció sigmoide. L'addició de la tècnica de Dropout, que desactiva un nombre determinat de neurones durant cada època d'entrenament, ha permès regularitzar el model en la mesura del possible, aconseguint una reducció de la pèrdua (cross-entropy) de 0.1. Malgrat aquesta millora, s'observa un sobreajust notable i una lleugera inestabilitat. No s'ha assolit el nivell d'ajust òptim esperat en la iteració anterior. En futures millores, s'exploraria la possibilitat de explorar i optimitzar els mètodes de regularització per a poder conservar aquesta lleugera millora sense perdre capacitat de generalització.

6.2.6 Anàlisi de resultats

Els resultats han estat obtinguts computant la mitjana aritmètica en els resultats de les diferents mètriques en 5 folds diferents per raons de robustesa.

Model	Loss	Accuracy	F1-Score	AUC	Precision	Recall
model1	0.404403	0.811345	0.810014	0.895810	0.812902	0.807203
model2	0.402449	0.813138	0.811949	0.897896	0.814234	0.809736
model3	0.405436	0.812432	0.811757	0.895503	0.811848	0.811755
model4	0.375933	0.825139	0.821868	0.911732	0.834519	0.809639

Tot i que no trobem diferencies molt significatives, el model 4 destaca per identificar lleugerament millor els patrons en les dades. És així com tot i ser un model amb més complexitat, l'escollirem com el definitiu.

Per a validar aquest model escollit, s'ha realitzat una corva roc, matriu de confusió i el comput de les mètriques anteriors sobre el cojunt de test.

Loss	**Accuracy**	**Precision**	**Recall**	**F1**	**AUC**
0.410937	0.80688	0.813904	0.795692	0.804695	0.89320

Tot i donar pitjor resultat, entra dins de l'esperat i presenta millora respecte el model lineal. També és important comentar que el model prediu millor quan no plou.

7 Conclusió

En conclusió, la solució plantejada per abordar el problema ofereix resultats robusts i prometedors per a la predicció de pluja a la regió d'Austràlia. En primer lloc, s'ha dut a terme una exhaustiva anàlisi exploratòria de les dades. Seguidament, s'ha realitzat un preprocessament i una enginyeria de característiques rigorosos amb l'objectiu d'extreure la màxima informació possible i adaptar-la per a alimentar dos models. Finalment, s'ha desenvolupat un model lineal de base com a referència i un perceptró multicapa mitjançant un procés iteratiu.

Tot i que els resultats del model lineal i del perceptró multicapa no mostren grans diferències, hem aconseguit millorar significativament les prediccions, així com realitzat un anàlisi del potencial rendiment i casos d'ús del model proposat.