

Examen Machinaal Leren OPLOSSINGEN

Eerste zittijd
19 januari 2022

Deel A

Voornaam: _____

Familienaam: _____

DRAAI DEZE PAGINA PAS OM NA HET SEIN VAN DE BEGELEIDERS

- Dit examen is gesloten boek. Er zijn geen hulpmiddelen toegelaten.
- Haal deze bundel niet uit elkaar.
- Noteer je antwoorden in deze bundel. Bij sommige vragen moet je een uitdrukking of waarde invullen in de daartoe voorziene ruimte én je antwoord motiveren. Vergeet deze motivatie niet. Zonder motivatie krijg je geen punten, zelfs al is je antwoord correct.
- Schrijf leesbaar.
- Achteraan deze bundel vind je een aantal lege bladen. Duid op elk blad goed aan of het over een kladblad of antwoordblad gaat.

1. Performantiemetrieken [4 punten]

De bijsluiter van een COVID19 zelftest toont onderstaande confusion matrix die de resultaten van de zelftest (Antigeen) vergelijkt met die van een PCR-test (die als ground-truth wordt beschouwd).

	PCR positief	PCR negatief	Totaal
Antigeen positief	85	4	89
Antigeen negatief	17	431	448
Totaal	102	435	537
Gevoeligheid	83.3% (95%CI: 74.7% - 90.0%)		
Specificiteit	99.1% (95%CI: 97.7% - 99.7%)		

- (a) Gevoeligheid en Specificiteit zijn synoniemen voor respectievelijk true positive rate en true negative rate. Leg beide termen uit aan de hand van dit voorbeeld. Je uitleg moet begrijpbaar zijn voor iemand zonder voorkennis over machine learning of statistiek.

Oplossing:

De gevoeligheid, ook wel recall of true positive rate genoemd, van de zelftest (antigeen) wordt berekend als het percentage van het echte aantal zieken (= aantal zieken positief op PCR) die door de zelftest als covid-positief worden aangeduid. Deze metriek geeft je m.a.w. een idee over de kans dat je zelftest positief is als je besmet bent. In dit voorbeeld is dit dus $85/102 = 83.3\%$.

De specificiteit is de true negative rate. Dit is het percentage gezonde mensen (volgens de antigeen) binnen het totaal aantal mensen dat als gezond wordt beschouwd door de zelftest. Deze metriek geeft m.a.w. informatie over hoe groot de kans is dat je effectief gezond bent, als de zelftest jou als gezond markeert. In dit voorbeeld is dit dus $431/435 = 99.1\%$.

- (b) Aan de hand van deze resultaten promoot de fabrikant de zelftest door een accuracy van 96% te beloven. Geef twee redenen waarom dit een vertekend beeld geeft.

Oplossing:

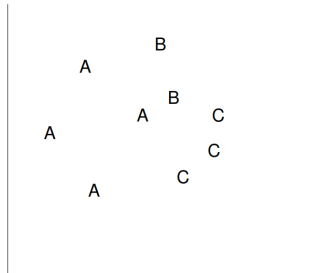
De dataset waarop de nauwkeurigheid is berekend was zeer ongebalanceerd: hij bevatte veel meer gezonde (448) dan zieke (89) personen. Nauwkeurigheid is nooit een goede metriek bij ongebalanceerde datasets.

Een PCR test zal enkel gebruikt worden door mensen die reeds symptomen vertonen, of die nog geen symptomen vertonen maar weten dat ze in aanraking zijn geweest met een ziek persoon. De kans dat die personen positief zijn is dus veel groter. De test zal dus gebruikt worden door een populatie die hoofdzakelijk zieke mensen bevat, terwijl de nauwkeurigheid is berekend op een populatie die grotendeels gezonde mensen bevatte.

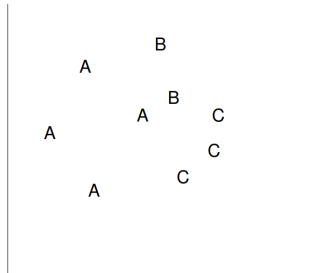
De meest relevante metriek is dus de gevoeligheid, maar die bedraagt 'slechts' 83.3% en is dus veel minder aantrekkelijk om te adverteren.

2. Classificatiemodellen [3 punten]

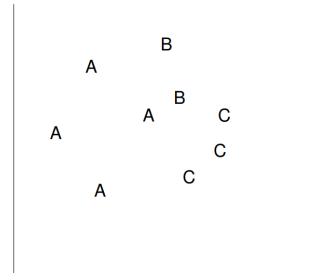
Onderstaande figuren stellen een classificatieprobleem voor met drie klassen en twee input features. We trainen drie verschillende classificatiemodellen op deze data: een decision tree (multiclass), een 1-nearest neighbor (multiclass) en een SVM (one-vs-all voor klasse A, geen kernel, C zeer groot). Teken een mogelijke decision boundary voor elk van deze modellen op onderstaande grafiek.



(a) Decision tree

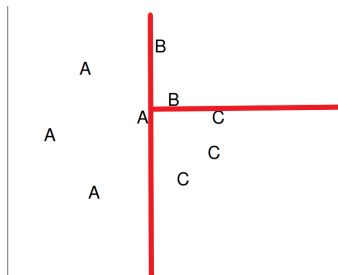


(b) 1-nearest neighbor

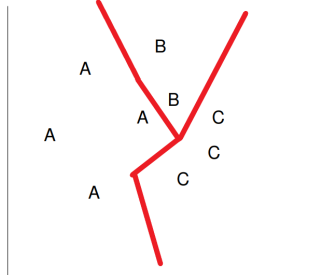


(c) SVM (one-vs-all)

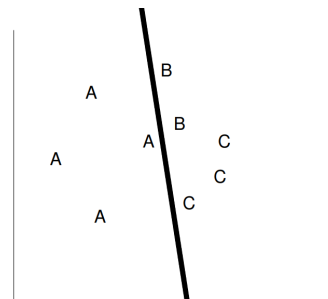
Oplossing:



(a) Decision tree



(b) 1-nearest neighbor



(c) SVM (one-vs-all)

Bij een decision tree wordt steeds gekeken naar het meest informatieve feature om het zoekgebied lineair te scheiden.

Bij een 1-nearest neighbour wordt enkel gekeken naar de afstand tot de meest nabije buur. Je krijgt dus een aaneenschakeling van lineaire grenslijnen.

Bij een SVM krijgen we een lineaire scheidingslijn. Aangezien de hyperparameter zeer groot is, neemt SVM genoeg met een zeer kleine marge, maar mogen geen datapunten aan de verkeerde kant liggen.

3. Juist of fout? Verklaar telkens heel kort je antwoord. [3 punten]

- (a) Door cross-validatie te gebruiken om hyperparameters te selecteren zijn we zeker dat het model niet kan overfitten.

Oplossing: Fout. Het model kan nog steeds overfitten als het eenvoudigweg te veel parameters bevat.

- (b) De “Random” in random forests slaat op het feit dat de beslissingen bij de decision trees telkens gebeuren op basis van een random gekozen feature.

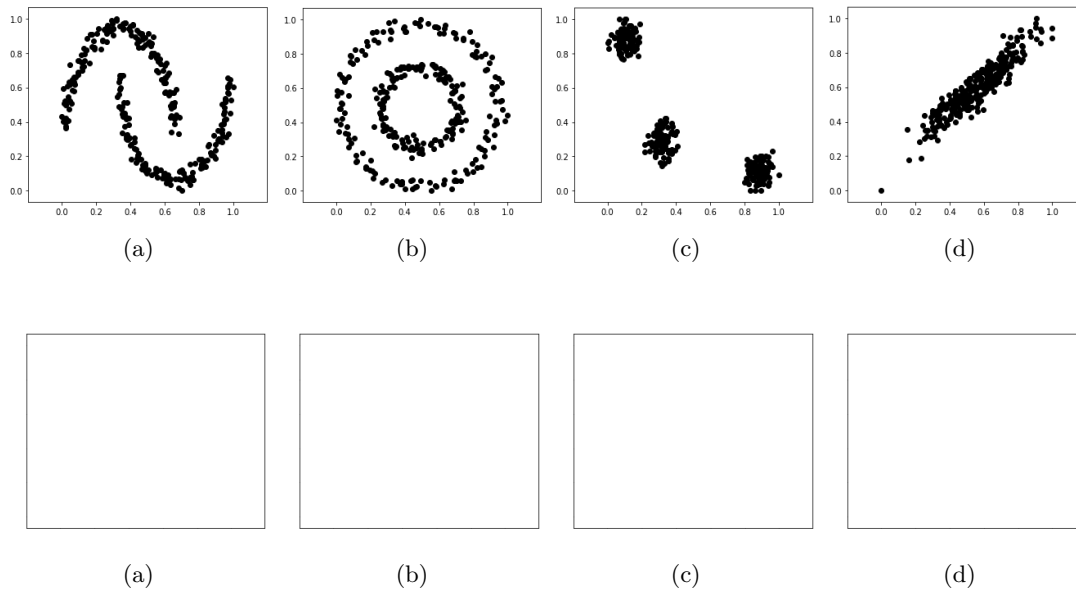
Oplossing: Fout. De beslissingen gebeuren op basis van de “beste” feature, gekozen uit een random subset van alle features.

- (c) Door extra features toe te voegen neemt het risico op overfitting af.

Oplossing: Juist. Meer features toevoegen maakt het probleem moeilijker, waardoor het model al zijn parameters moet gebruiken om een betere performantie te halen, maar het geen kans krijgt om de dataset ‘van buiten’ te leren. Hier gaan we er natuurlijk van uit dat de extra features extra informatie bijdragen en niet gewoonweg duplicate of redundante features zijn.

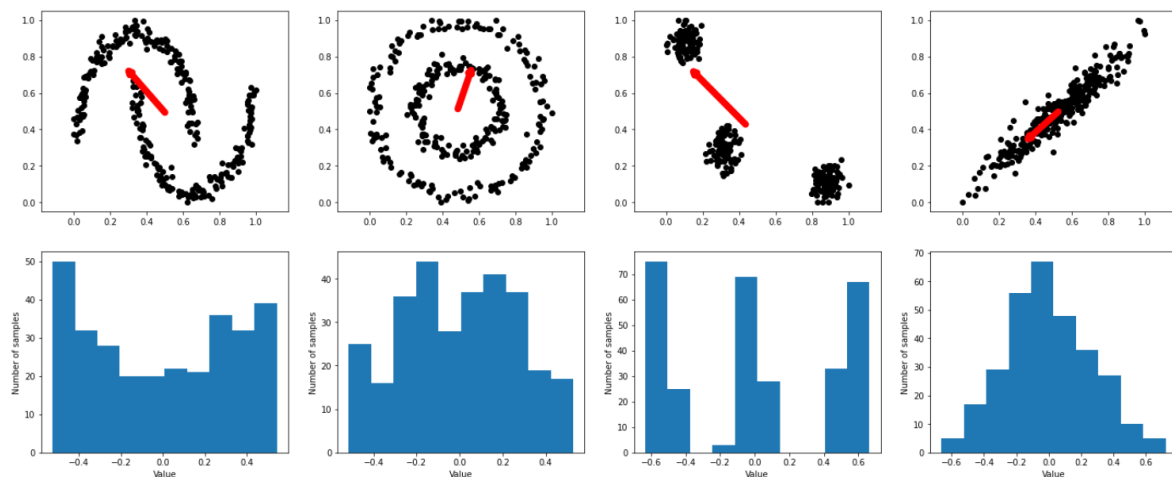
4. PCA [5 punten]

Onderstaande figuur toont vier datasets ($x_i \in \mathbb{R}^2$). Beide assen hebben dezelfde schaal en alle data ligt tussen 0 en 1. We willen de dimensionaliteit verminderen naar één. Hiervoor gebruiken we PCA.



(a) Schets de eerste principal component op elke figuur.

(b) Schets een histogram van elke dataset na PCA dimensionality reduction in de vakken op de onderste rij.



De principal component van de eerste dataset ligt diagonaal. De richting is moeilijker te bepalen (beide werden goedgekeurd).

In dataset (b) kan de principal component twee datasets kan in willekeurige ligging liggen, omdat de variantie in beide features dezelfde is. De getoonde oplossing is dus maar 1 van de vele.

(c) Sorteer de vier datasets volgens dalende variantie verklaard door de eerste principal component.

Oplossing: De verklaarde variantie is hoger als er minder informatie overblijft na projectie op de principal component.

De principal component van dataset (d) verklaart de meeste variantie, gevolgd door dataset (c). Dataset (c) heeft iets meer spreiding in de richting orthogonaal op de principal component dan dataset (d). Dataset (a) heeft nog iets meer spreiding, en in dataset (b) is de spreiding het hoogst.

De oplossing is dus, in dalende volgorde: (d), (c), (a), (b).

- (d) PCA is een lineaire dimensionaliteitsreductie techniek. Verklaar de term “Linear”. Wat zijn hiervan de nadelen? Geef twee mogelijke oplossingen om dit te vermijden.

Antwoord:

Oplossing: Lineaire technieken herleiden de dimensionaliteit door datapunten te projecteren naar een andere basis. In essentie komt dit neer op een lineaire transformatie die uit te drukken is met een matrix. De basisvectoren die het nieuwe coördinatensysteem bepalen zijn een lineaire combinatie van de oorspronkelijke basisvectoren.

De projectie zelf herleidt de dimensionaliteit niet. De reductie gebeurt omdat we, na projectie, enkel bepaalde dimensies overhouden.

Het grootste nadeel van lineaire dimensionaliteitsreductie is dat je veronderstelt dat de dataset zonder al te veel verlies van informatie kan geprojecteerd worden op een hypervlak. Dit is in veel gevallen een onterechte veronderstelling.

Alternatieven zijn niet-lineaire technieken zoals Linear Local Embedding (waarbij slechts lineariteit op korte afstand wordt verondersteld) en kernel PCA.

5. Neural Networks [5 punten]

- (a) We bouwen een eenvoudig neuraal netwerk met Keras aan de hand van onderstaande code.

```
model = Sequential()  
model.add(Dense(10, activation='relu', input_shape=(16,)))  
model.add(Dense(20, activation='relu'))  
model.add(Dense(30, activation='relu'))  
model.add(Dense(2, activation='softmax'))
```

Hoeveel parameters heeft dit model ?

Antwoord:

Oplossing: $10 * 16 + 10 + 20 * 10 + 20 + 30 * 20 + 30 + 2 * 30 + 2$

- (b) Overfitting is een vaak voorkomend probleem bij neurale netwerken. Hieronder kan je een aantal mogelijke strategieën vinden. Geef telkens aan of dit de mate van overfitting zal verminderen, vermeerderen of dat ze hierop geen effect zal hebben. Motiveer steeds kort je antwoord.

- Een extra laag toevoegen.

Antwoord:

- Vergroten van de L^2 regularisatie.

Antwoord:

Oplossing: a. Door een extra laag toe te voegen zal je de capaciteit van het model vergroten en loop je dus meer risico op overfitting.
b. De L^2 regularisatie forceert het model om kleinere gewichten te gebruiken. Dit helpt meestal om overfitting tegen te gaan.

Klad