Werkgroep 7: Classificatie en discriminatie analyse

Opdracht 1 – Check your comprehension:

- What are the possible measurement levels of independent and dependent variables in a discriminant analysis?
  Independent: interval (binair)
  Dependent: Nominaal
- What is the role of generalised Euclidean distance in the assignment of individuals to groups?
  De afstand berekenen van subject-punten naar alle groep-punten. Daarna ieder subject toewijzen aan de groep met de kortste afstand.
- What is wrong (or missing) with the PAC as a measure of accuracy of classification?
  PAC is een 'global measure', dus het maakt geen onderscheid tussen soorten error (false positive vs. false negative).
- What is the difference between the sensitivity and the positive predictive value of some test?
  Sensitivity = $p(X^+|Y^+)$. Dit is de conditionele kans dat iemand met een ziekte (Y^+) een positieve diagnose (X^+) krijgt. De 'positive predicitive value' = $p(Y^+|X^+)$. Dit is de conditionele kans dat iemand met een positieve diagnose (X^+) de ziekte ook echt heeft (Y^+).
- What is the difference between the specificity and the negative predictive value of some test?
  Specificity = $p(X^-|Y^-)$. Dit is de conditionele kans dat iemand zonder de ziekte (Y^-) een negatieve diagnose (X^-) krijgt. 'Negative predictive value' = $p(Y^-|X^-)$. Dit is de conditionele kans dat iemand met een negatieve diagnose (X^-) de ziekte ook echt niet heeft (Y^-).
- Why is the base rate of some disease usually very important for the positive and negative predictive value of a test for the disease, but not at all important for its sensitivity and specificity?
  De *'base rate'* zal de percentages correcte diagnoses niet zal veranderen als de groepen (met ziekte en zonder ziekte) de populatie vertegenwoordigen, dus '*sensitivity'* en '*specificity'* zullen niet veranderen. Echter, een veranderende *'base rate'* heeft wel effect op het aantal *'true positives'* en *'false positives'*. Voorbeeld: als de *'base rate'* voor 'geen ziekte' het hoogst is, zullen er meer *'false positives'* zijn in vergelijking met het aantal *'false negatives',* dus de *'positive predicitive value'* wordt lager.

Opdracht 2 – Find an optimal cut-off point

- For each of these seven cut-off points, calculate the numbers of false positives and false negatives.
  Zeven cut-off points:
  Cut-off point X_c:
    - Diagnose positief (X+) als X is groter of gelijk aan X_c
    - Diagnose negatief (X-) als X is kleiner dan X_c. Voorbeeld $X_c = 3,5$

obstipatie (Y+) controle (Y-)

Cut-off point	0,5	1,5	2,5	3,5	4,5	5,5	6,5
Valse neg. (n)	0	0	5	15	45	125	200
Valse pos. (p)	200	180	130	70	30	5	0

- Which cut-off point do you choose if we regard both types of errors as equally bad?
  Als men allebei de fouten (positief en negatief) even erg vindt:

som = n + p	200	180	135	85	75	130	200

- Which cut-off point do you choose if we regard false negatives as twice as bad as false positives?
  Als men negatieve fouten zwaarder wil laten wegen (2x) dan positieve fouten. Het laagste punt verschuift naar links (eerst 75 en nu 100).

Som = 2n + p	200	180	140	100	120	255	400

Opdracht 3 – Assignment of individuals to groups

Drummer Joe ("Animal") von Karajan has scores 11 on speed, 14 on basic rhythms, and 13 on complex rhythms. Using the assignment method on p. 4-6 of this week's text, what kind of drummer is Joe?

$$d(\text{Joe, rock}) = \sqrt{(11-14)^2 + (14-11)^2 + (13-10)^2} = 5{,}2$$

Dit doe je ook voor Jazz = 3,0 en voor Klassiek = 3,7.
De afstand bij Jazz is het kleinst dus Joe hoort bij de jazz groep.

Opdracht 4 – Discriminant analysis with SPSS

- What are the independent and dependent variables in this analysis?
  Afhankelijk: Groep (= leerstoornis) □ de voorspelling
  Onafhankelijk: PERF, INFO, VERBEXP, AGE □ voorspellend
- How do the three groups appear to be distinct from one another?
  De 'memory groep' lijkt vooral lager te schoren op INFO dan de andere groepen (M = 7.00 vs. 11.67 en 9.67).
  De 'perceptie-groep' lijkt vooral lager te scoren op PERF dan de andere groepen (M= 87.67 vs. 98.67 en 100.33)
  De 'communicatie-groep' lijkt vooral lager te scoren op VERBEXP dan de andere groepen (M= 28.33 vs. 36.33 en 38.33)
- Is the prediction better than we would expect on the basis of chance?
  Wilks'lambda = 0.010,
  X^2 (8) = 20.51,
  $p < 0.01$
  Dus het antwoord is ja, de voorspelling is beter dan verwacht op basis van kans.
  (let op: Neem altijd de bovenste Wilk's lambda)
- Calculate the PAC, sensitivity and specificity.
  Dit dient per groep bekeken te worden.
  PAC = (3+3+3)/9 = 1
  Sensitiviteit (memory) = 3/3 = 1. Hetzelfde geldt voor perceptie en communicatie.
  Specificiteit (memory) = (3+3+0+0)/6 = 1. Hetzelfde geldt voor perceptie en communicatie.
- Suppose that one wrongly classified child is added to the sample (the child has a visual perception disorder, but receives a memory disorder as the prediction). Create a new classification table yourself, and re-calculate the PAC, sensitivity and specificity. Which measures change, and which ones stay the same?
  PAC = (3+3+3)/10 = 0.90 (was 1)
  Sensitiviteit:
  Memory: 3/3 = 1.
  Perceptie: 3/4 = 0.**75** (was 1).
  Communication = 3/3 = 1.
  Specificiteit:
  Memory: (3+3)/ 7 = 0.**86** (was 1).
  Perceptie (3+3)/ 6 = 1.
  Communication = (3 + 1+ 3)/7 = 1

Opdracht 5 – Classification and Bayer's rule

- For this classification table, calculate the percentage of "hits" (or "percentage accurately classified": PAC) and two probabilities, namely that a Discriminantosis patient or a healthy subject respectively will receive the diagnosis Discriminantosis. We will call these probabilities p(X+|Y+) and p(X+|Y-).
  PAC: (195 + 190) / 400 = 0.9625
  Sensitiviteit: p(X+ | Y+) 195/200 = 0.975
  False positives: p(X+|Y-) 10/200 = 0.05
- Calculate p(Y+|X+), the probability that someone with the diagnosis Discriminantosis actually has the disease, for the following two populations:
  • the population of the Netherlands, where the disease occurs in 1 in 10000 people;
  • psychological methodologists, a recognised high-risk group, where no less than 25 percent suffer from this dreaded occupational disability.
  Regel van Bayes: rekening houden met hoe vaak de ziekte voorkomt in de hele bevolking (corrigeren).
  Bevolking: P(Y+) = 0.0001, dus P(Y-) = 0.9999

$$P(Y+|X+) = \frac{P(X+|Y+)P(Y+)}{P(X+|Y+)P(Y+) + P(X+|Y-)P(Y-)}$$

  P(Y+|X+) = (0,975*0.0001)/(0,975*0,0001 + 0,05*0,9999) = 0,0019
  Vanwege de lage *'base rate'* is er een erg lage *'positive predicitive value'*.

  Methodologen: $P(Y^+) = 0,25$, dus $P(Y^-) = 0,75$
  $P(Y^+|X^+) = (0,975*0,25)/(0,975*0,25 + 0,05*0,75) = 0,867$
  Een hogere base rate (25%) □ bij een positieve diagnose nu wel een grote kans op discriminatore (86,7%)
- For both populations, calculate the PAC (although it is not really necessary, for convenience you may assume 10000 people in each population) under two conditions:
  i) without use of diagnostic information (i.e. if we allocate all subjects to the most common group);
  ii) with use of diagnostic information (i.e. if we allocate subjects to groups on the basis of DA). Does the use of diagnostic information result in a better prediction?
  Zonder diagnose X: voorspel bij iedereen de meest voorkomende categorie: niet ziek.
  Met diagnose : voorspel Y^+ na positieve diagnose X^+. Voorspel Y^- na negatieve diagnose X^-.

  Nederlandse bevolking: zonder diagnose: voorspel altijd Y-, want P(Y-) = 0,9999 □ $PAC_{zonderdiagnose} = 0.9999$

	Diagnose: D	Diagnose: N	Totaal
Werkelijk D	(0,975 * 1) = ongeveer 1	(0,025 * 1) = ongeveer 0	1
Werkelijk N	(0,05 * 9999) = ongeveer 500	(0,95 * 9999) = ongeveer 9499	9999
Totaal	501	9499	10000

Tabel: with diagnosis 1: classification table with numbers (N=10000).

$PAC_{with} = (1 + 9499) / 10000 = 0.95$. Door gebruik van diagnostische informatie is er sprake van een lagere PAC (0,95) dan zonder gebruik hiervan (0,9999).
Deze methode is niet precies aangezien het aantal mensen wordt afgerond, maar als een overzicht werkt het goed.

	Diagnose: D	Diagnose: N	Totaal
Werkelijke D	(0,975 * 0.0001)	(0,025 * 0.0001)	0.0001

	= 0.0000975	= 0.0000025	
Werkelijke ND	(0,05 * 0.9999) = 0.049995	(0,95 * 9999) = 0.949905	0.9999
Totaal	0.0500925	0.9499075	1

Tabel: with diagnosis 2: classification table with proportions.