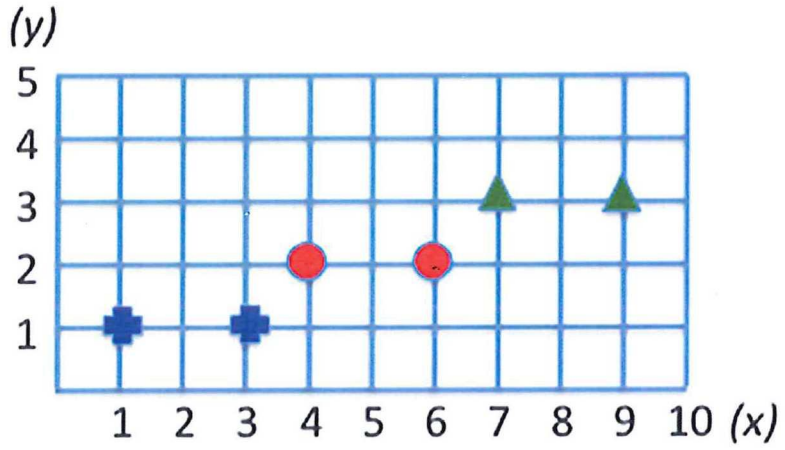
Delft University of Technology
Faculty Electrical Engineering, Mathematics and Computer Science
Pattern Recognition & Bioinformatics Group
Multimedia Signal Processing Group

**Data Mining (TI2730-C)**
Examination, Tuesday April 9, 2013, 14:00 – 17:00

- Geef, indien mogelijk, niet alleen de uitkomsten, maar ook de tussenliggende berekeningen.
- Beantwoording van de vragen mag in het Nederlands of Engels, maar hou het wel kort en bondig.
- Geef de antwoorden voor iedere vraag op een afzonderlijk blad, want de opgaven worden voor het nakijken gesplitst!
- Vergeet niet je naam en studienummer te zetten op ieder blad!
- Het totaal aantal punten dat je voor de opgaven kunt behalen is 33.

**Opgave 1 (7pt)**

In figuur 1 is een drie-klassen probleem gegeven. Om dit drie-klassen probleem te classificeren wordt gebruikt gemaakt van de nearest mean classifier.




*Figuur 1 : Een drie klassen probleem met voor elk object twee kenmerken x en y.*

a, 1pt) Een nieuw object heeft kenmerken (3,2). Welke klasse voorspeld de nearest mean classifier die getrained is op de data in figuur 1 voor dit object.

b, 1pt) Wat is de apparant error van de nearest mean classifier getraind op de data in figuur 1.

c, 1pt) Bereken de 6-fold cross-validatie error van de nearest mean classifier gegeven de data in figuur 1.

d, 1pt) Wat is het verschil tussen de 'true error' en de 'bayes error'.

e, 1pt) De nearest mean classifier is een Bayes classifier. Geef zoveel mogelijk aannames die worden gemaakt in een nearest mean classifier.

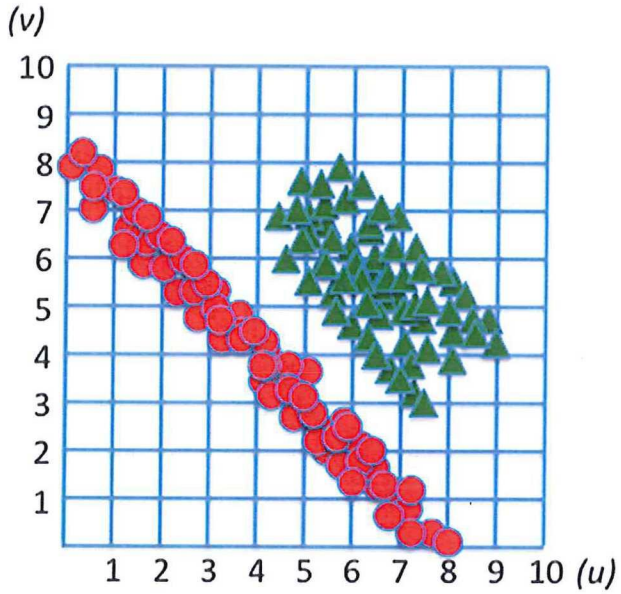
f, 2pt) Aan welke aanname wordt het minst voldaan.

**Opgave 2 (7pt)**

Principal Component Analysis (PCA) is een techniek om data te transformeren.

a, 1pt) Geef aan waar en waarvoor PCA ingezet kan worden in een datamining process.

In figuur 2 is een twee-klassen probleem gegeven.




Figuur 2. *Een twee klassen probleem met voor elk object twee kenmerken u en v.*

b, 1pt) Geef een schatting van de PCA componenten (eigen vectoren) van deze dataset.

c, 1 pt) Geef een schatting van de eigenwaarden die bij deze eigenvectoren horen.

d, 1pt) Teken de distributie van beide klassen als de data geprojecteerd wordt op de *eerste* eigenvector (als je het antwoord op *b* niet weet, neem dan een eigenvector aan).

e, 1pt) Teken de distributie van beide klassen als de data geprojecteerd wordt op de *tweede* eigenvector (als je het antwoord op *b* niet weet, neem dan een eigenvector anders dan in antwoord d aan).

f, 1pt) Voor welke van de twee projecties (in antwoorden *d* en *e*) is het twee-klassen probleem het best scheidbaar.

g, 1pt) Geef een (wiskundige) maat die deze scheidbaarheid kan bepalen.

**Opgave 3 (9pt)**

Assume we have 5 objects with 2 features (x,y):

(1,2)
(1,4)
(3,2)
(3,3)
(6,1)

In order to find clusters the k-mean clustering algorithm is used.

a, 2pt) Draw the objects in a 2-dimensional feature space. And assume now that k=2. The starting cluster centers are (1,0) and (3,0). Determine the final clusters and give the coordinates of these cluster centers.

b, 3pt) Explain what the Davies-Bouldin index is measuring and what the idea is behind the formula. Compute the Davies-Bouldin index for the final result of a).

For the remaining of this exercise a new object (2,12) is added to the group of the 5 objects.

c, 1pt) Again one wants to use the k-means algorithm. One can select k=2 or k=3. Which one do you prefer. Elucidate your answer.

There is also another object of which unfortunately the y-value is missing. This object is: (4,y).

d, 3pt) Determine for this object the y-value in case of mean imputation and hot deck imputation, respectively.
Which one do you prefer in this case? Elucidate your answer.

**Opgave 4 (10pt)**

In a database, a tuple of gender, age and occupation is defined as a quasi-identifier, meaning that this tuple can be used to identify individuals. Based on this information and the table below, answer the following questions.

| Gender | Age | Occupation | Sickness |
|---|---|---|---|
| Male | 21-30 | Writer | Diabetes |
| Female | 21-30 | Singer | Flu |
| Female | 21-30 | Writer | HIV |
| Female | 31-40 | Writer | Flu |
| Male | 11-20 | Singer | Flu |
| Male | 21-30 | Actor | Diabetes |
| Female | 31-40 | Singer | Diabetes |
| Male | 21-30 | Writer | Flu |

a) (2pt) Give a brief definition for K-anonymity. What is the value of K for the quasi-identifier in the table above?

b) (1pt) Alter the table in such a way that the table has 3-anonymity for the quasi-identifier (gender-age-occupation).

c) (2pt) Bob wants to learn Alice's sickness. He knows that she is a writer and 32 years old. Give the probabilities for possible sicknesses Alice might have, that is P(HIV), P(Diabetes), P(Flu).

d) (2pt) Give a brief definition for L-diversity. What is the value of L for the sickness attribute? (Remember that L-diversity is defined for anonymized sub-groups. Therefore, the entries should be grouped first based on the similar tuples.)

e) (1pt) Alter the table in such a way that the sickness attribute has 3-diversity.

f) (2pt) What is the effect of anonymization on the utility? Explain briefly with an **example**.