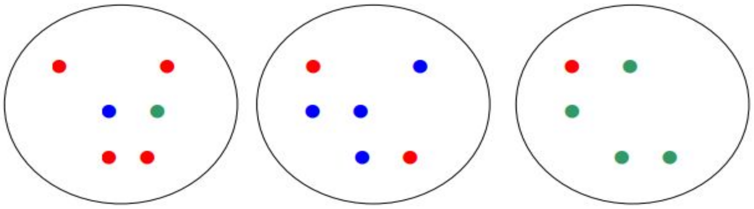
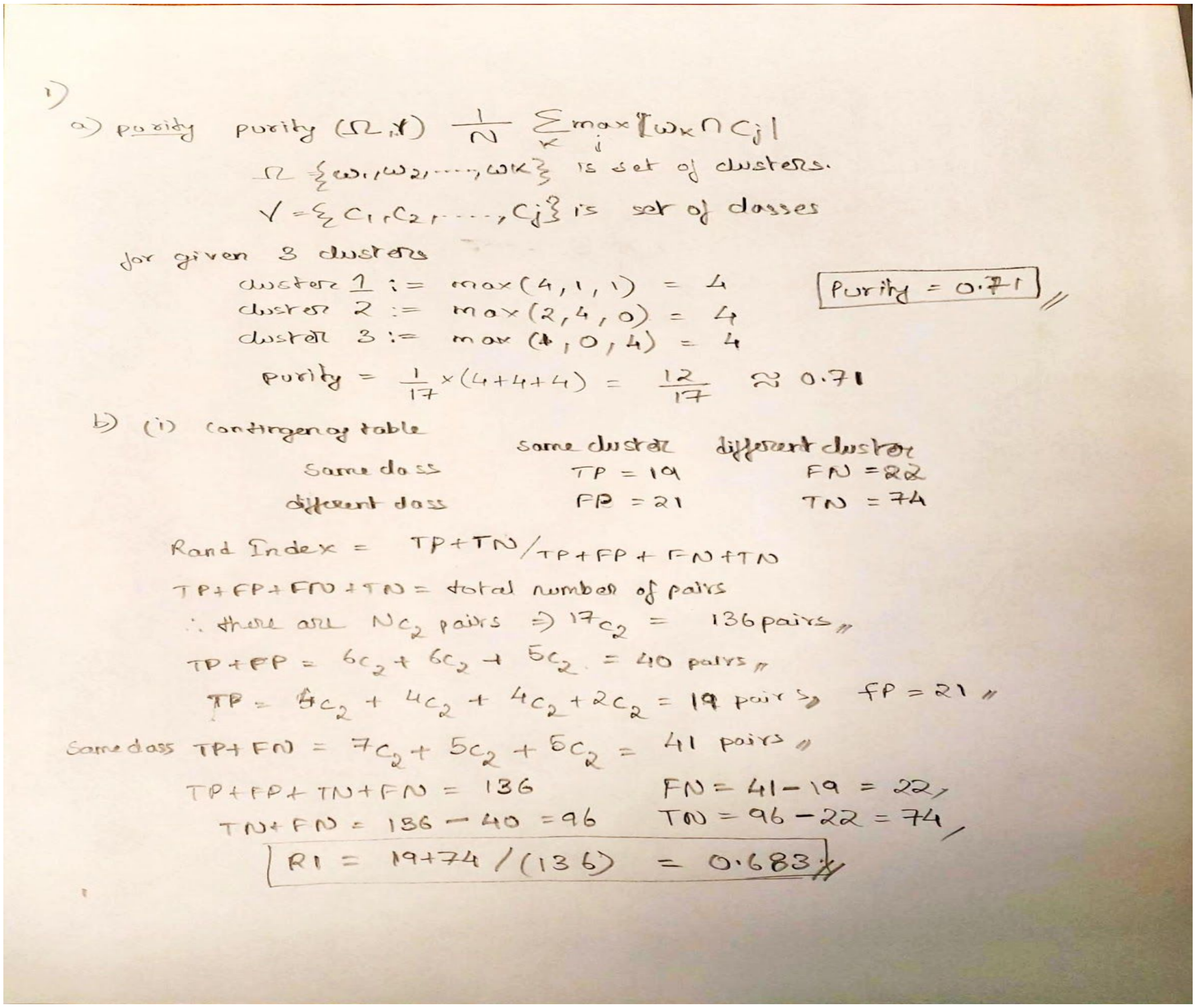
CS 620/DASC 600 : Introduction to Data Science | Homework 5

Herambeshwar Pendyala | 01130541

hpend001@odu.edu

1) (30 points) Consider the following 3 clusters.
   a) Calculate the Purity
   b) Create the contingency table (confusion matrix) and using the contingency table,
      i) Calculate the Rand index
      ii) Calculate the Balanced F measure





1)

a) purity purity (Ω, γ) $\frac{1}{N} \sum_{k} \max_{j} |\omega_k \cap c_j|$

Ω $\{\omega_1, \omega_2, \dots, \omega_K\}$ is set of clusters.

$\gamma = \{c_1, c_2, \dots, c_j\}$ is set of classes

for given 3 clusters

cluster 1 := max(4,1,1) = 4

cluster 2 := max(2,4,0) = 4

cluster 3 := max(1,0,4) = 4

$\boxed{\text{Purity} = 0.71}$ //

purity $= \frac{1}{17} \times (4+4+4) = \frac{12}{17} \approx 0.71$

b) (i) contingency table

	same cluster	different cluster
same class	TP = 19	FN = 22
different class	FP = 21	TN = 74

Rand Index = TP+TN / TP+FP+FN+TN

TP+FP+FN+TN = total number of pairs

$\therefore$ there are $^{N}C_2$ pairs $\Rightarrow$ $^{17}C_2$ = 136 pairs //

TP+FP = $^{6}C_2 + {}^{6}C_2 + {}^{5}C_2$ = 40 pairs //

TP = $^{4}C_2 + {}^{4}C_2 + {}^{4}C_2 + {}^{2}C_2$ = 19 pairs, FP = 21 //

Same class TP+FN = $^{7}C_2 + {}^{5}C_2 + {}^{5}C_2$ = 41 pairs //

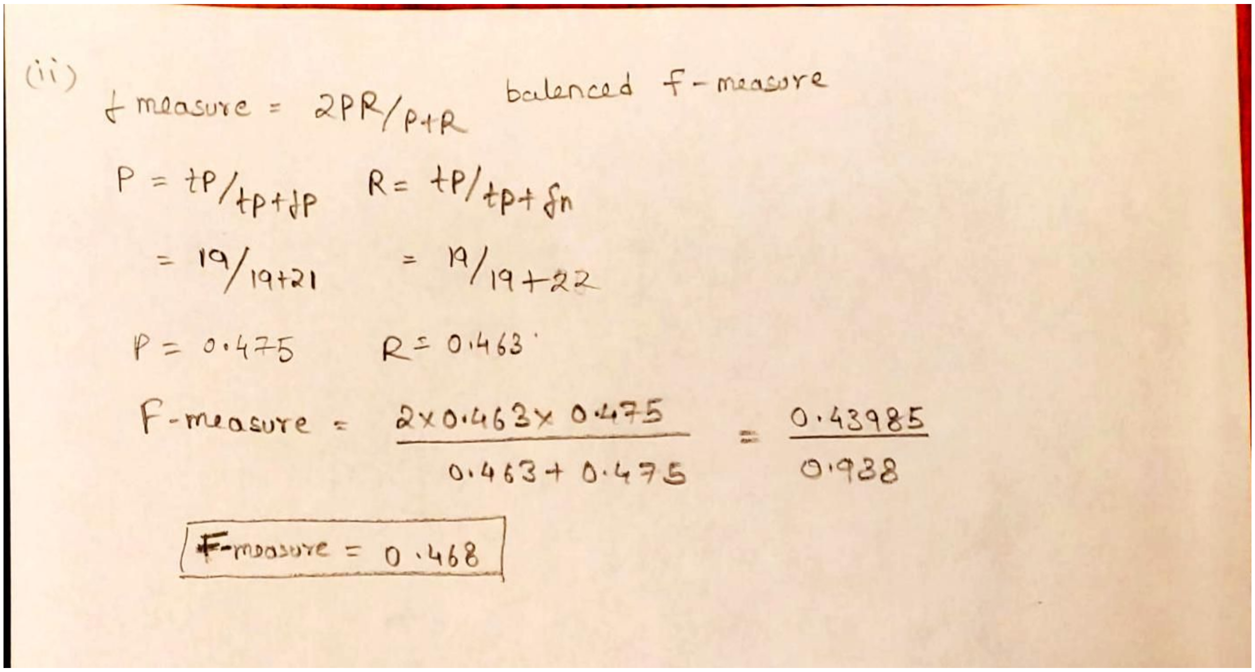
TP+FP+TN+FN = 136

FN = 41 − 19 = 22,

TN+FN = 136 − 40 = 96

TN = 96 − 22 = 74

$\boxed{RI = 19+74/(136) = 0.683}$ //



(ii)

$f\ measure = 2PR/P+R$ balenced f-measure

$P = tP/tp+fP$ $R = tP/tp+fn$

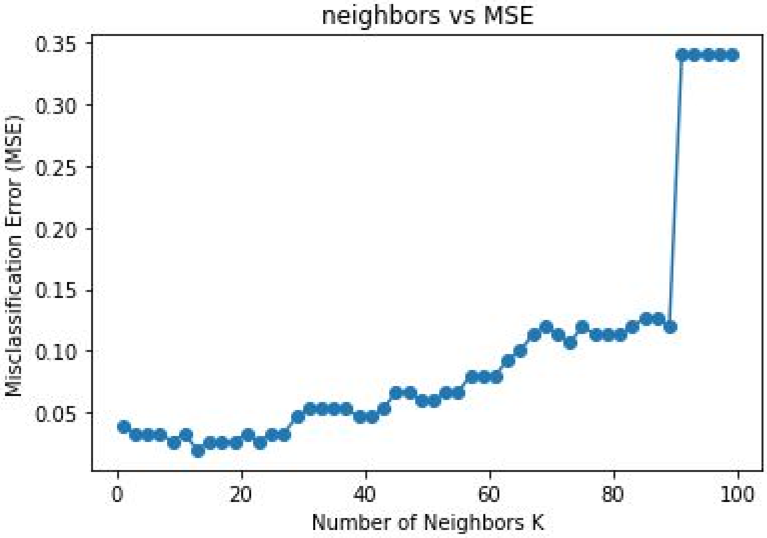
$= 19/19+21$ $= 19/19+22$

$P = 0.475$ $R = 0.463$

$$F\text{-measure} = \frac{2\times0.463\times0.475}{0.463+0.475} = \frac{0.43985}{0.938}$$

$$\boxed{F\text{-measure} = 0.468}$$

2) Generate a Plot “neighbors” vs “MSE” and also find and print the optimal K using the “MSE” list. Include the plot as a figure in your pdf.



```
The optimal 'k' is 13 with misclassification error(MSE) of
0.01999999999999907
```

3) (20 points) Describe your thoughts about what you think it means to work as a data scientist. You may therefore – if you like – be very personal and describe your own plans and fears for your future career, criticism (or appreciation) for your education, skills you

need to develop further, and soon. This question is intended to encourage you to reflect about yourself and your future career, and will therefore be graded generously!

Ans)

As a Data scientist, I believe what we do actually helps organizations and its people to get a more clear understanding about their data. Every problem we solve makes the difference and every product we design or create helps people connect with each other and help them . According to me we are not just programmers, problem solvers we are also an artist and a good story teller.

Plans for future career

- In my early childhood days, I was interested in the field of robotics which made me to pursue engineering. It was during the undergraduate time I developed an interest in Data Science, Machine learning and its applications. So I am aligned to pursue my career in the field of data science.
- I aim to work as a data scientist working in the field of Natural Language Processing and Computer Vision, solving real world problems.

Fears for future career.

- As computer science is an evolving field, we need to keep ourselves updated and as of today it is developing at a very fast pace and I fear one day, I may not be able to catch up with it.
- Maintaining work life balance is necessary and this is one of the fears.
- Fear of getting into a field where I cannot utilize my technical expertise and getting used to monotonous work, this not only destroys the skill but also narrows the person's thinking capability as he/she is no longer doing any new activity.

Criticism/praise for education

- The education until undergraduate level earned me the ability indirectly to adapt to upcoming new challenges, handle them with ease and gain appraisal by implementation.
- To illustrate this with my very own past experience : During my tenure as a software engineer in a multinational company, I worked on building up new non existing model in

an unhealthy environment where not many resources are provided to cope up with the challenges I faced. My learning abilities and skills has certainly paved the way in achieving the goal of the project and earn client appreciations.

Skills to develop in future

- As an aspiring Data scientist I plan to learn new skills while updating the old skills.
- As part of technical skills, I want to increase my skills in the area of High Performance Computing, as expertise in this area is required to be efficient in implementing deep learning models.
- I plan to improve my computer programming, problem solving skills by participating in programming contests.
- Along with good technical skills, strong interpersonal and communication skills are also required to achieve the career goals and I aim to improve my skills to grow my career.