## Universidade Federal do Rio Grande do Sul Instituto de Informática Aplicada Biologia Computacional - Lista 7 Prof. Dr. Márcio Dorn

Nome: Vicente Merlo Matrícula: 244950

O arquivo de treinamento foi carregado usando a biblioteca pandas, apropriada para manipulação de datasets.

Usei o sklearn para realizar o KMeans. Para fazer isso, transformei o objeto DataFrame obtido pelo pandas no carregamento do CSV para listas da linguagem.

Executei o KMeans com k = 2, e obtive dois clusters de centróides:

(os dois centróides são vetores de dimensão 7128, então salvei no arquivo centroides k2.txt)

Esses centróides são de 2 clusters obtidos "cegamente", não levando em consideração qual o diagnóstico real. Para comparar se os resultados obtidos estão correspondendo com a realidade, iterei sobre todas as amostras de treinamento e verifiquei a qual cluster elas pertencem:

Amostra 0

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 2

Amostra 1

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 2

Amostra 2

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 1

Amostra 3

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 1

Amostra 4

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 1

Amostra 5

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 2

Amostra 6

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 1

Amostra 7

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 1

Amostra 8

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 1

Amostra 9

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 1

Amostra 10

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 1

Amostra 11

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 2

Amostra 12

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 1

Amostra 13

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 1

Amostra 14

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 1

Amostra 15

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 1

Amostra 16

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 1

Amostra 17

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 2

Amostra 18

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 2

Amostra 19

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 1

Amostra 20

Esperado: AML (GRUPO 2)

Obtido: GRUPO 2

Amostra 21

Esperado: AML (GRUPO 2)

Obtido: GRUPO 2

Amostra 22

Esperado: AML (GRUPO 2)

Obtido: GRUPO 2

Amostra 23

Esperado: AML (GRUPO 2)

Obtido: GRUPO 2

Amostra 24

Esperado: AML (GRUPO 2)

Obtido: GRUPO 2

Amostra 25

Esperado: AML (GRUPO 2)

Obtido: GRUPO 2

Amostra 26

Esperado: AML (GRUPO 2)

Obtido: GRUPO 2

Amostra 27

Esperado: AML (GRUPO 2)

Obtido: GRUPO 2

Amostra 28

Esperado: AML (GRUPO 2)

Obtido: GRUPO 2

Amostra 29

Esperado: AML (GRUPO 2)

Obtido: GRUPO 2

Amostra 30

Esperado: AML (GRUPO 2)

Obtido: GRUPO 1

Amostra 31

Esperado: AML (GRUPO 2)

Obtido: GRUPO 2

Amostra 32

Esperado: AML (GRUPO 2)

Obtido: GRUPO 2

Amostra 33

Esperado: AML (GRUPO 2)

Obtido: GRUPO 2

Amostra 34

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 2

Amostra 35

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 2

Amostra 36

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 2

Amostra 37

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 2

Amostra 38

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 1

Amostra 39

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 2

Amostra 40

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 2

Amostra 41

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 2

Amostra 42

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 2

Amostra 43

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 2

Amostra 44

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 2

Amostra 45

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 2

Amostra 46

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 1

Amostra 47

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 2

Amostra 48

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 1

Amostra 49

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 1

Amostra 50

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 2

Amostra 51

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 2

Amostra 52

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 1

Amostra 53

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 1

Amostra 54

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 1

Amostra 55

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 2

Amostra 56

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 2

Amostra 57

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 1

Amostra 58

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 2

Amostra 59

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 1

Amostra 60

Esperado: ALL (GRUPO 1)

Obtido: GRUPO 2

Amostra 61

Esperado: AML (GRUPO 2)

Obtido: GRUPO 2

Amostra 62

Esperado: AML (GRUPO 2)

Obtido: GRUPO 2

Amostra 63

Esperado: AML (GRUPO 2)

Obtido: GRUPO 2

Amostra 64

Esperado: AML (GRUPO 2)

Obtido: GRUPO 2

Amostra 65

Esperado: AML (GRUPO 2)

Obtido: GRUPO 2

Amostra 66

Esperado: AML (GRUPO 2)

Obtido: GRUPO 2

Amostra 67

Esperado: AML (GRUPO 2)

Obtido: GRUPO 1

Amostra 68

Esperado: AML (GRUPO 2)

Obtido: GRUPO 2

Amostra 69

Esperado: AML (GRUPO 2)

Obtido: GRUPO 2

Amostra 70

Esperado: AML (GRUPO 2)

Obtido: GRUPO 2

Amostra 71

Esperado: AML (GRUPO 2)

Obtido: GRUPO 2

Naturalmente, ocorre um overfit, pois a amostra de treinamento é a mesma amostra que a de testes.

Como as primeiras centróides são escolhidas aleatoriamente, os resultados divergem: Em algumas execuções, a taxa de acerto é de 98%, e em outras, de 50%. Por isso, fiz a média entre 10 execuções, e obtive a taxa de acerto média de:

68% de acerto

E com k = 3, obtive três centróides:

(salvos no arquivo centroides\_k3.txt)