Reconhecimento de Padrões: Atividade 04 − Comparação de Classificadores

Carlos Emmanuel Pereira Alves
Curso de Bacharelado em Ciência da Computação
Universidade Federal do Agreste de Pernambuco (UFAPE)
Garanhuns, Brasil
carlos.emmanuel.236@gmail.com

1) Realize 100-fold cross validation estratificado na base Skin Segmentation utilizando o classificador 1-NN com distância Euclidiana então realize os procedimentos abaixo.

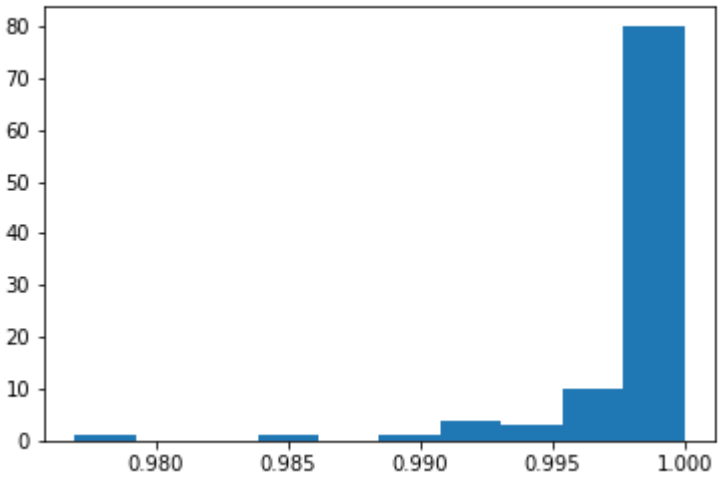
a) Mostre a média, o máximo e o mínimo da medida-F.

Medida-F Classe 1
Max: 1.0
Med: 0.99831
Min: 0.97692

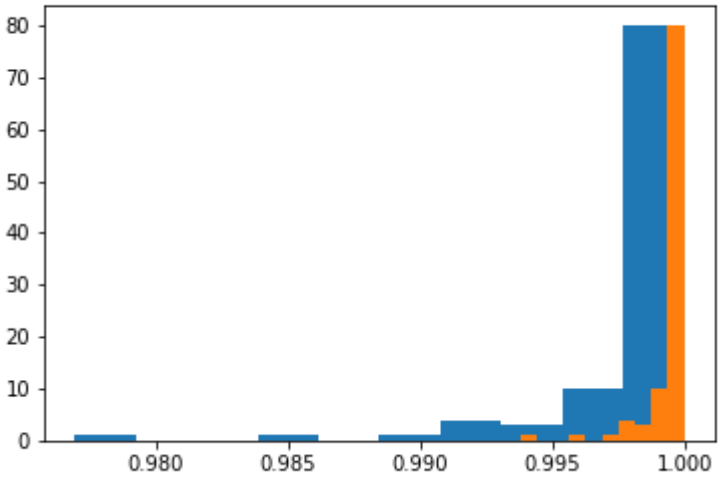
Medida-F Classe 2
Max: 1.0
Med: 0.99955
Min: 0.99378

b) Mostre o histograma da medida-F.

Classe 1



Classe 2



c) Calcule o intervalo de confiança da medida-F.
Intervalo de confiança da Classe 1: [0.99166 1.00496]
Intervalo de confiança da Classe 2: [0.99778 1.00132]

d) Qual a medida-F mínima que você espera ao aplicar este classificador, sob as mesmas condições de treinamento, para dados nunca vistos?
A medida-F mínima esperada vai ser de 99,166% para a Classe 1 e de 99,778% para a Classe 2, que é o menor limite calculado no intervalo de

confiança. Temos o nível de confiança em 95%, por isso podemos ter um número menor que não foi previsto em nenhum dos intervalos.

e) Qual a medida-F esperada para o classificador quando aplicada a dados nunca antes vistos.
Para dados não vistos antes teremos a média de 99,831% para a Classe 1 e de 99,955% para a Classe 2.

2) Realize um experimento pareado com 100 repetições de Holdout 50/50 utilizando o classificador 1-NN com distância Euclidiana. Utilize duas versões da base Wine archive.ics.uci.edu/ml/datasets/Wine para este experimento, a primeira versão é a base original, a segunda versão é a base sem a última coluna. Após calcular 100 taxas de acerto para cada uma das versões da base, realize os procedimentos abaixo.

a) Calcule a diferença das 100 taxas de acerto.

```
Diferença das 100 taxas de acerto:
[-0.1011236  -0.15730337 -0.08988764 -0.11235955 -0.07865169 -0.19101124
 -0.13483146 -0.07865169 -0.11235955 -0.11235955 -0.14606742  0.03370787
 -0.16853933 -0.21348315 -0.08988764 -0.15730337  0.03370787 -0.03370787
 -0.1011236  -0.25842697 -0.05617978 -0.19101124 -0.15730337 -0.11235955
 -0.13483146 -0.11235955 -0.1011236  -0.1011236  -0.06741573 -0.15730337
 -0.08988764 -0.21348315 -0.1011236  -0.11235955 -0.07865169 -0.06741573
 -0.04494382 -0.12359551  0.02247191 -0.07865169 -0.15730337 -0.14606742
 -0.02247191 -0.08988764 -0.04494382 -0.08988764 -0.12359551 -0.01123596
 -0.06741573 -0.14606742 -0.1011236  -0.08988764 -0.08988764 -0.11235955
 -0.1011236  -0.08988764 -0.03370787 -0.25842697 -0.11235955 -0.16853933
 -0.16853933 -0.08988764 -0.16853933 -0.15730337 -0.13483146 -0.1011236
 -0.12359551 -0.12359551 -0.07865169 -0.21348315 -0.13483146 -0.1011236
 -0.14606742 -0.12359551 -0.20224719 -0.16853933 -0.16853933 -0.23595506
 -0.03370787 -0.12359551 -0.1011236  -0.1011236  -0.12359551 -0.13483146
 -0.11235955 -0.03370787 -0.12359551 -0.16853933 -0.11235955 -0.06741573
 -0.15730337 -0.03370787 -0.01123596 -0.17977528 -0.14606742 -0.05617978
 -0.07865169 -0.16853933 -0.1011236  -0.08988764]
```

b) Calcule o intervalo de confiança destas diferenças.
Intervalo de confiança: [-0.2238 -0.00227]

c) Realize o teste de hipótese sobre estas diferenças para verificar se a diferença da taxa de acerto é significativa entre as duas versões. Mostre sua conclusão para o teste.

Como o 0 está fora do intervalo, rejeitamos o H_0, ou seja, os classificadores não tem o mesmo erro. Analisando vemos que, a base sem a última coluna tem uma taxa de acerto significativamente maior.

d) Calcule o intervalo de confiança da taxa de acerto para cada versão da base.
Intervalo de confiança da taxa de acerto da base completa:
[0.64918 0.79217]
Intervalo de confiança da taxa de acerto da base sem a última coluna:
[0.76324 0.90418]

e) Realize o teste de hipótese de sobreposição dos intervalos de confiança. Mostre sua conclusão para o teste.
Existe sobreposição entre os intervalos de confiança, portanto, não podemos afirmar se os classificadores possuem ou não taxas de acerto diferentes.

3) Qual o número máximo de características que podem ser removidas da base Iris archive.ics.uci.edu/ml/datasets/iris sem reduzir significativamente a taxa de acerto? Defina a metodologia utilizada para justificar sua resposta.
Utilizando o 1-nn com distância euclidiana, 100 repetições e holdout 50/50, e a base completa:
Intervalo de confiança da taxa de acerto: [0.92012 0.99135]

Utilizando o 1-nn com distância euclidiana, 100 repetições e holdout 50/50, e a base sem a última coluna:
Intervalo de confiança da taxa de acerto: [0.87729 0.97338]

Utilizando o 1-nn com distância euclidiana, 100 repetições e holdout 50/50, e a base sem as duas últimas colunas:
Intervalo de confiança da taxa de acerto: [0.64748 0.81812]

Utilizando o 1-nn com distância euclidiana, 100 repetições e holdout 50/50, e a base sem as três últimas colunas:
Intervalo de confiança da taxa de acerto: [0.49659 0.72874]

Podemos ver que retirando 1 coluna o intervalo de confiança da taxa de acerto não diminui tanto, ainda temos sobreposição. Mas quando retiramos 2 colunas o intervalo de confiança da taxa de acerto cai significativamente. E, ainda, quando retiramos 3 colunas o intervalo de confiança da taxa de acerto cai drasticamente. Então o número máximo de colunas que podem ser retiradas sem afetar significativamente o intervalo de confiança da taxa de acerto é 1.

4) Utilizando o classicador k-NN na base Wine archive.ics.uci.edu/ml/datasets/Wine, teste os valores k = 1, . . . , 15. Para qual valor de k o classificador apresenta uma taxa de acerto significativamente maior? Defina a metodologia utilizada para justificar sua resposta.

1-nn
Média de acertos: 0.71506
Intervalo de confiança: [0.62015 0.80996]

2-nn
Média de acertos: 0.72472
Intervalo de confiança: [0.63793 0.81151]

3-nn
Média de acertos: 0.70449
Intervalo de confiança: [0.62537 0.78362]

4-nn
Média de acertos: 0.70124
Intervalo de confiança: [0.6193 0.78317]

5-nn
Média de acertos: 0.71191
Intervalo de confiança: [0.63211 0.79171]

6-nn
Média de acertos: 0.71169
Intervalo de confiança: [0.6387 0.78467]

7-nn
Média de acertos: 0.71416
Intervalo de confiança: [0.63804 0.79027]

8-nn
Média de acertos: 0.70652
Intervalo de confiança: [0.63182 0.78122]

9-nn
Média de acertos: 0.71146
Intervalo de confiança: [0.63152 0.7914]

10-nn
Média de acertos: 0.71056
Intervalo de confiança: [0.63054 0.79058]

11-nn
Média de acertos: 0.7091
Intervalo de confiança: [0.62536 0.79284]

12-nn
Média de acertos: 0.70281
Intervalo de confiança: [0.62015 0.78547]

13-nn
Média de acertos: 0.71371
Intervalo de confiança: [0.64231 0.7851]

14-nn
Média de acertos: 0.70528
Intervalo de confiança: [0.62342 0.78714]

15-nn
Média de acertos: 0.71798
Intervalo de confiança: [0.64672 0.78924]

Vemos que todos os intervalos se sobrepõem, então não podemos rejeitar o H_0. Devemos olhar outras métricas para tentar classificá-los.