Relatório de AM - Equipe 06

Diógenes Wallis de França Silva
Johnny Marcos Silva Soares
Rodrigo Vitor Alves de Mello

Dados

Características do *Dataset*

Data Set Characteristics:	Multivariate	**Number of Instances:**	1484	**Area:**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	8	**Date Donated**	1996-09-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	333591

(Sem *Missing Data)*

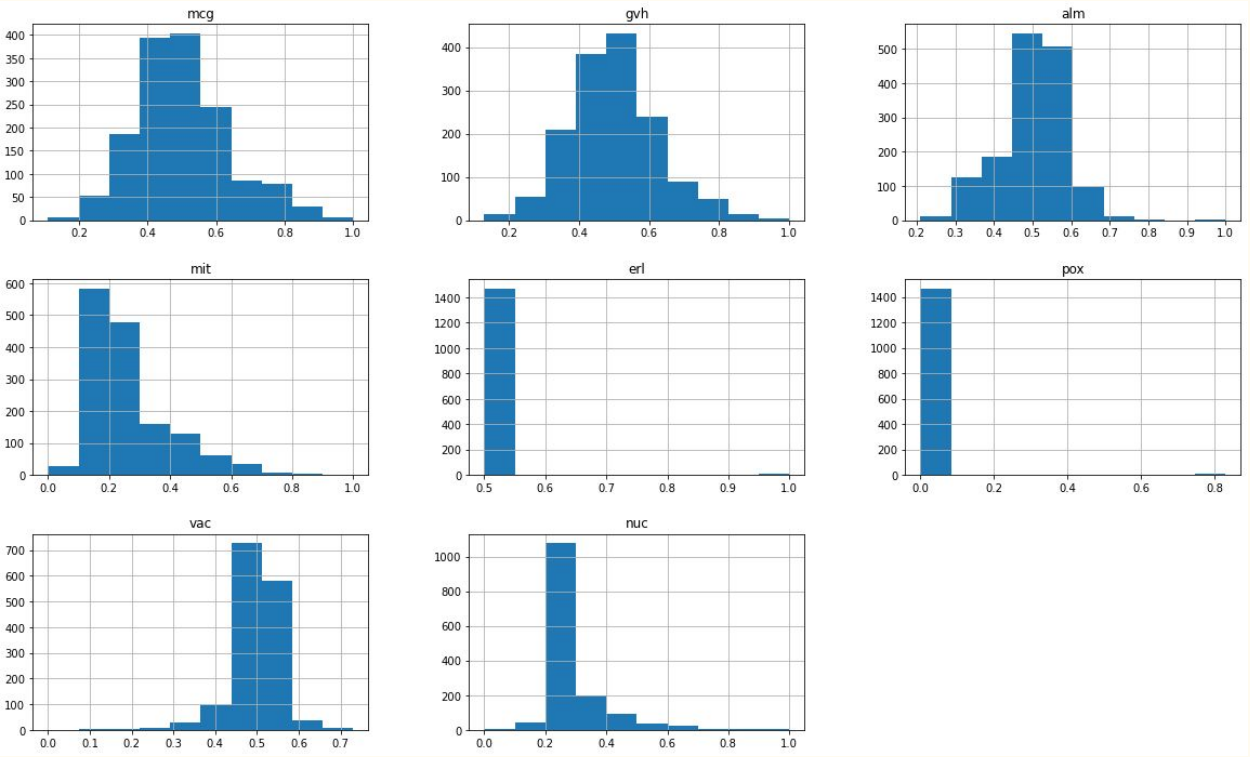
Significado dos atributos (amostra)

Índice	Atributo	Descrição
0	mcg	Método de McGeoch para reconhecimento de sequência de sinal.
1	gvh	Método de von Heijne para reconhecimento de sequência de sinal.
2	alm	Pontuação do programa de previsão da região de abrangência da membrana ALOM.
3	mit	Pontuação da análise discriminante do conteúdo de aminoácidos da região N-terminal (20 resíduos de comprimento) de proteínas mitocondriais e não mitocondriais.

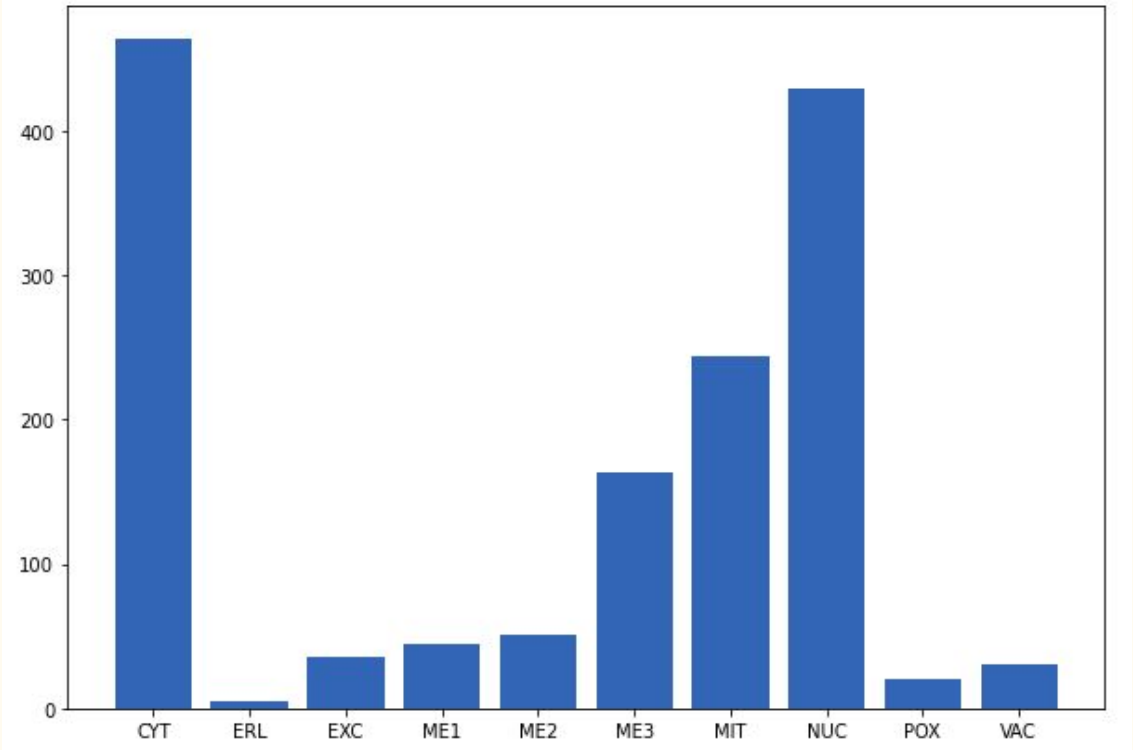
Significado das classes (amostra)

2	MIT	mitocondrial
3	ME3	proteína de membrana, sem sinal N-terminal
4	ME2	proteína de membrana, sinal não clivado
5	ME1	proteína de membrana, sinal clivado
6	EXC	extracelular

Pré-Processamento (histograma dos atributos)



Pré-Processamento (histograma das classes)



Questão 1

Etapas

- Implementação FCM-DFCV
- Avaliação das partições
- Análise do melhor resultado

- Implementação FCM-DFCV

1. Inicialização aleatória da matriz de pertinência
2. Cálculo dos protótipos
3. Cálculo das distâncias
4. Cálculo da função objetivo
5. Cálculos do grau de pertinência

- Implementação FCM-DFCV

Problemas de implementação (Indeterminação)

1. **Cálculo dos protótipos:**
   Quando nenhuma amostra pertence ao cluster, será mantido o valor anterior do protótipo.

2. **Cálculo dos pesos de relevância:**
   Quando não existe amostra no cluster ou a distância do protótipo para a amostra é próximo de zero. Além disso, pode ocorrer overflow no cálculo do lambda. Tratamos no numpy verificando NaN (Not a Number). Caso ocorra NaN utilizamos a matriz M anterior.

3. **Cálculo do grau de pertinência:**
   Com distância próxima à 0, consideramos o valor 1 na pertinência em uma classe e 0 nas outras.

- Avaliação das partições

Valor de **m** utilizado	J - Função de custo	F-Measure	ARI	PE	MPC	w (aleatoriedade na inicialização)
1.1	1.10493	0.182480	0.01019	0.000825	0.99967	93
1.6	0.449938	0.181859	0.013080	0.11267	0.96255	22
2.0	0.064732	0.17879	0.00887	0.080185	0.98184	98

- Análise do melhor resultado

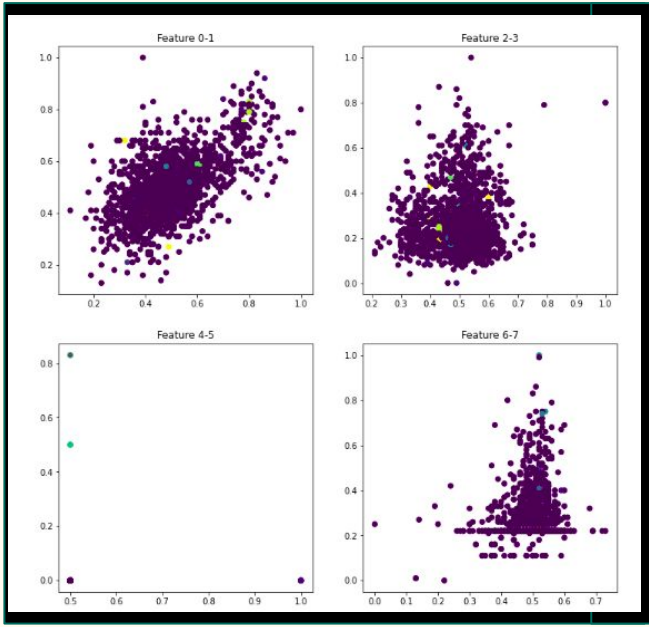
i) Protótipos

Protótipos	Feature 0	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7
Protótipo 0	0.495432	0.495256	0.501227	0.258262	0.5	0.0	0.499054	0.275499
Protótipo 1	0.662582	0.649665	0.492881	0.290595	0.99886	0.0	0.526019	0.25268
Protótipo 2	0.394642	0.383293	0.497410	0.151234	0.999973	0.0	0.530750	0.362294
Protótipo 3	0.571249	0.521402	0.460401	0.203001	0.5	0.829541	0.519781	0.40331
Protótipo 4	0.454812	0.526925	0.466934	0.174350	0.5	0.0	0.526506	0.772470
Protótipo 5	0.478467	0.557906	0.509643	0.636973	0.5	1.7e-06	0.514823	0.221632
Protótipo 6	0.467212	0.472401	0.525431	0.287125	0.5	0.584104	0.487815	0.230274
Protótipo 7	0.600729	0.587460	0.472981	0.487699	0.5	0.0	0.532463	0.220378
Protótipo 8	0.782893	0.760238	0.417798	0.272557	0.5	0.0	0.524617	0.220728
Protótipo 9	0.510468	0.511318	0.512704	0.232428	0.5	0.826378	0.519828	0.221666

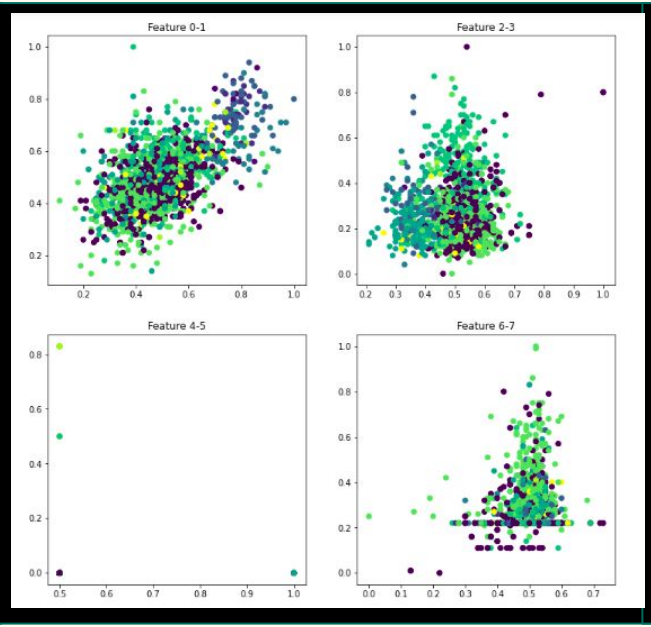
- Análise do melhor resultado

i) Protótipos

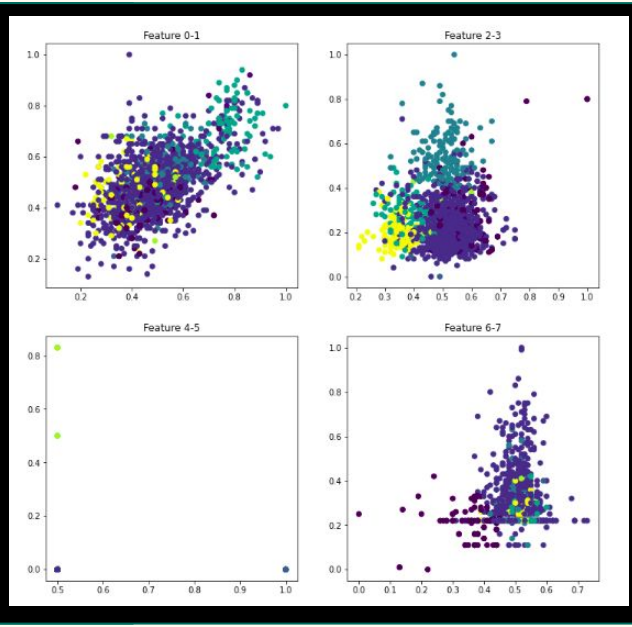
Partição com menor J



Classificação a priori



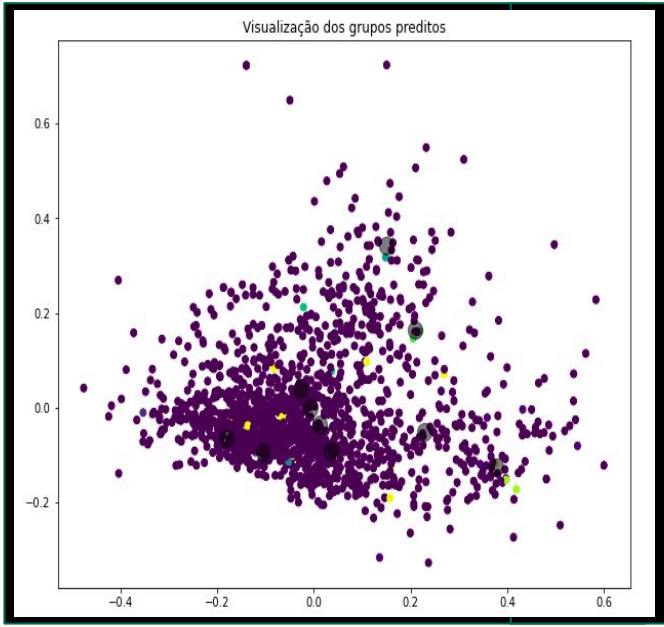
Partição com maior F-Measure



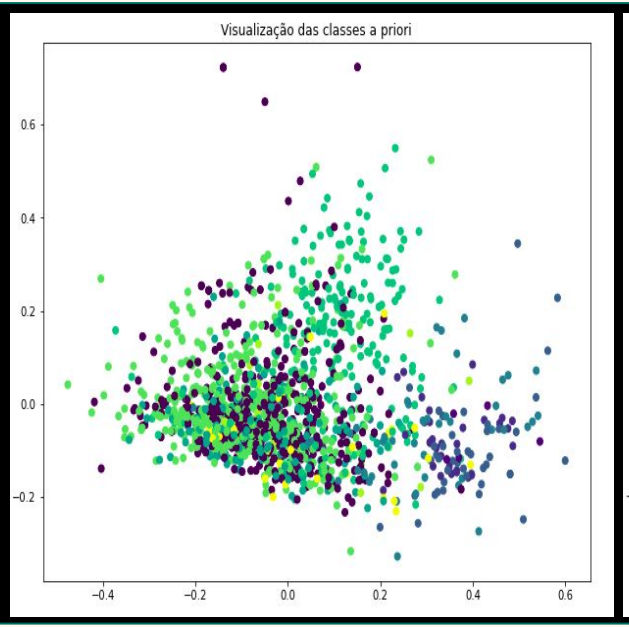
- Análise do melhor resultado

i) Protótipos (PCA)

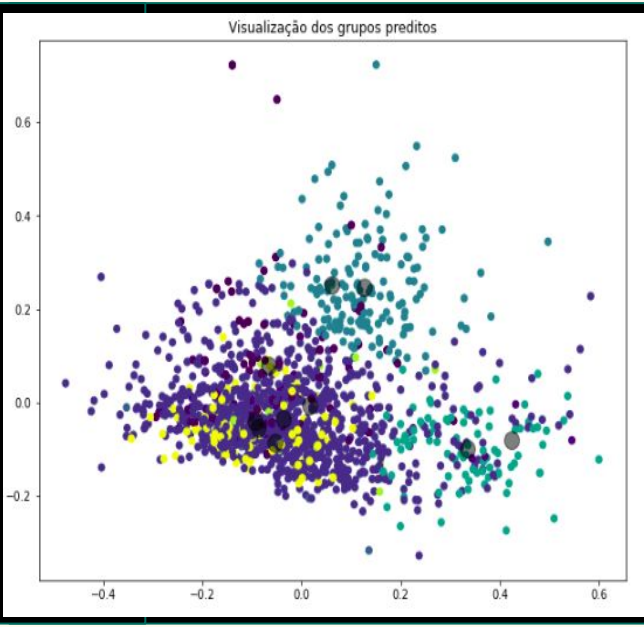
Partição com menor J



Classificação a priori



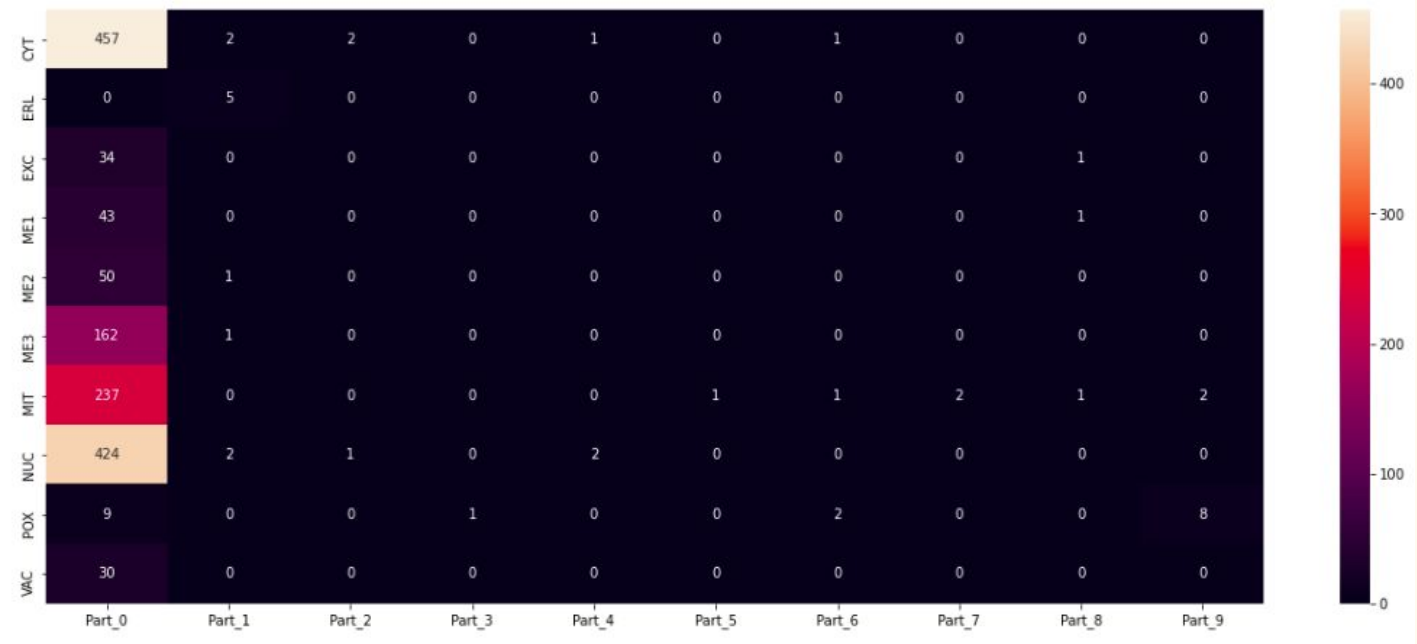
Partição com maior F-Measure



- Análise do melhor resultado

ii) Matriz de confusão

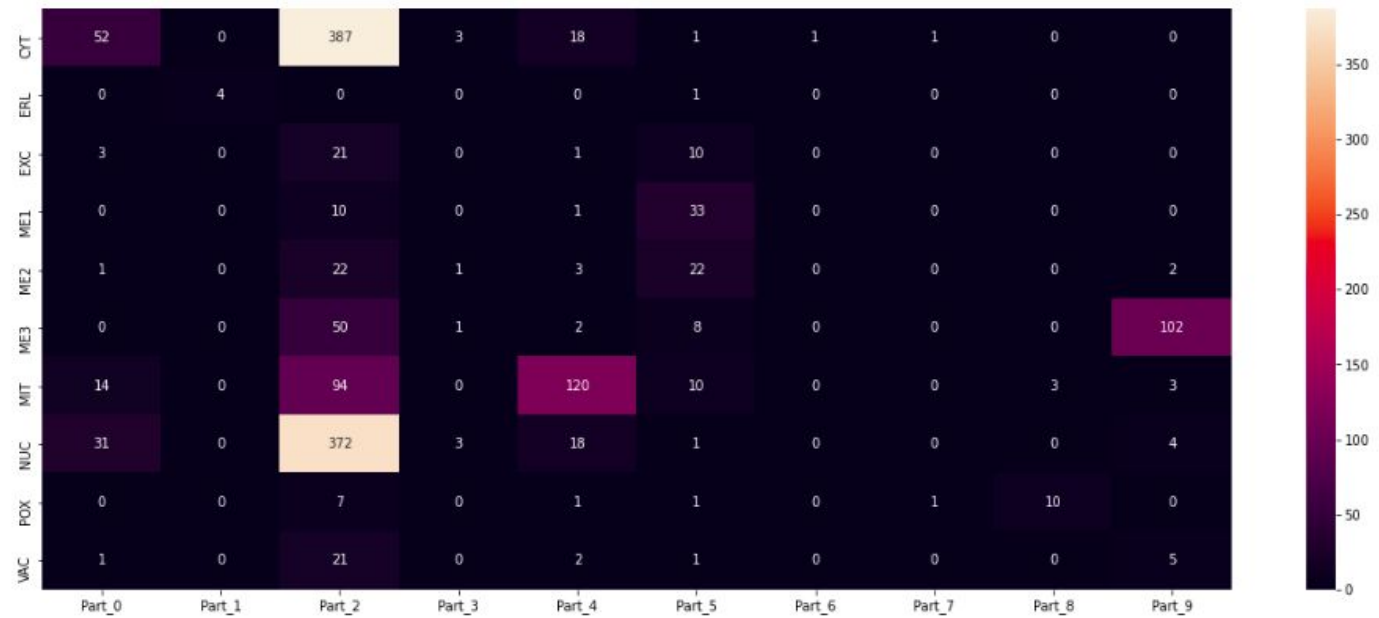
Partição com menor J



- Análise do melhor resultado

ii) Matriz de confusão

Partição com maior F-Measure



- Análise do melhor resultado

iii) Modified Partition Coefficient e Partition Entropy

Partição	Partition Entropy	Modified Partition Coefficient
Partição com menor J	0.000825	0.99967
Partição com maior F-Measure	0.11267	0.96255

- Análise do melhor resultado

iv) Índice de Rand Corrigido, F-Measure e o Erro de Classificação

Partição	F-Measure	Índice de Rand Corrigido	Erro de Classificação (OERC)	Acurácia
Partição com menor J	0.182480	0.01019	0.01347	0.3207
Partição com maior F-Measure	0.26887	0.1932	0.26415	0.4676

Questão 2

Classificadores

- Bayesiano Gaussiano
- k-Vizinhos
- Janela de Parzen
- Regressão Logística
- Voto Majoritário

Bayesiano Gaussiano (Fundamentação)

1. $$P(\omega_j|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_j)P(\omega_j)}{p(\mathbf{x})}$$

2. $$p(\mathbf{x}) = \sum_{j=1}^{c} p(\mathbf{x}|\omega_j)P(\omega_j)$$

3. $$posterior = \frac{likelihood \times prior}{evidence}$$

Bayesiano Gaussiano (Implementação)

$$P(\omega_l|\mathbf{x}_k) = \max_{i=1}^{10} P(\omega_i|\mathbf{x}_k) \text{ com } P(\omega_i|\mathbf{x}_k) = \frac{p(\mathbf{x}_k|\omega_i)P(\omega_i)}{\sum_{r=1}^{c} p(\mathbf{x}_k|\omega_r)P(\omega_r)} \; (1 \leq l \leq 10)$$

$$p(\mathbf{x}_k|\omega_i, \boldsymbol{\theta}_i) = (2\pi)^{-\frac{d}{2}} (|\boldsymbol{\Sigma}_i^{-1}|)^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu}_i)^{tr} \boldsymbol{\Sigma}_i^{-1} (\mathbf{x}_k - \boldsymbol{\mu}_i)\right\}, \text{ onde}$$

$$\boldsymbol{\theta}_i = \begin{pmatrix} \boldsymbol{\mu}_i \\ \boldsymbol{\Sigma}_i \end{pmatrix}, \boldsymbol{\Sigma}_i = diag(\sigma^2, \ldots, \sigma^2)$$

$$\boldsymbol{\mu}_i = \frac{1}{n} \sum_{k=1}^{n} \mathbf{x}_k, \mu_{ij} = \frac{1}{n} \sum_{k=1}^{n} x_{kj}$$

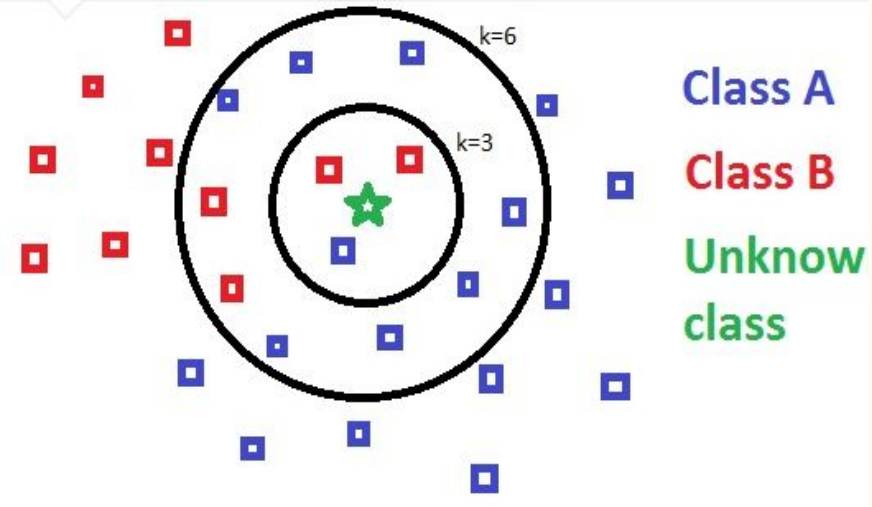
$$\sigma^2 = \frac{1}{d \times n} \sum_{k=1}^{n} ||\mathbf{x}_k - \boldsymbol{\mu}_i||^2 = \frac{1}{d \times n} \sum_{k=1}^{n} \sum_{j=1}^{d} (x_{kj} - \mu_{ij})^2 \; (1 \leq j \leq d)$$

Bayesiano Gaussiano (Desempenho)

Métrica	Média	Desvio Padrão	Intervalo de confiança
Acurácia	0.557	0.029	0.557 ± 0.025
Precisão	0.519	0.047	0.519 ± 0.041
Cobertura	0.541	0.055	0.541 ± 0.048
F-Measure	0.513	0.051	0.513 ± 0.045

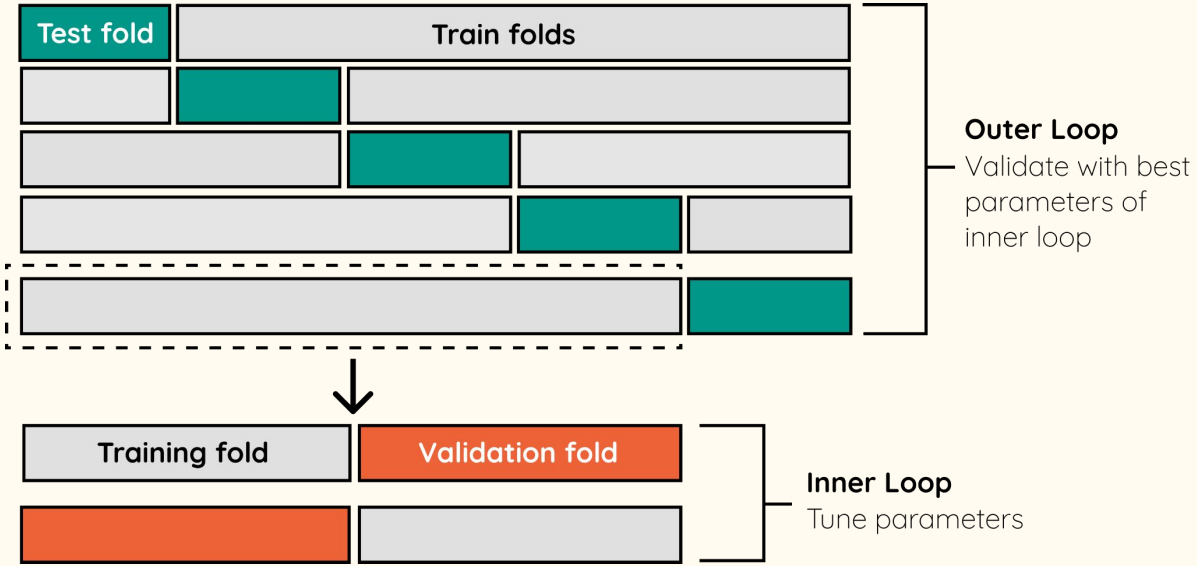
k-Vizinhos (Fundamentação)

1. Estima diretamente a probabilidade a posteriori
2. $P(x|wj)=(kj/nj) / V$
3. $P(wj)=nj/n$
4. $P(x)=k/(nV)$



k-Vizinhos (Hiperparâmetros)

1. Nested Cross-Validation
2. Externa: 5-fold estratificado
3. Interna: 5-fold
4. Evita superestimação



k-Vizinhos (Desempenho)

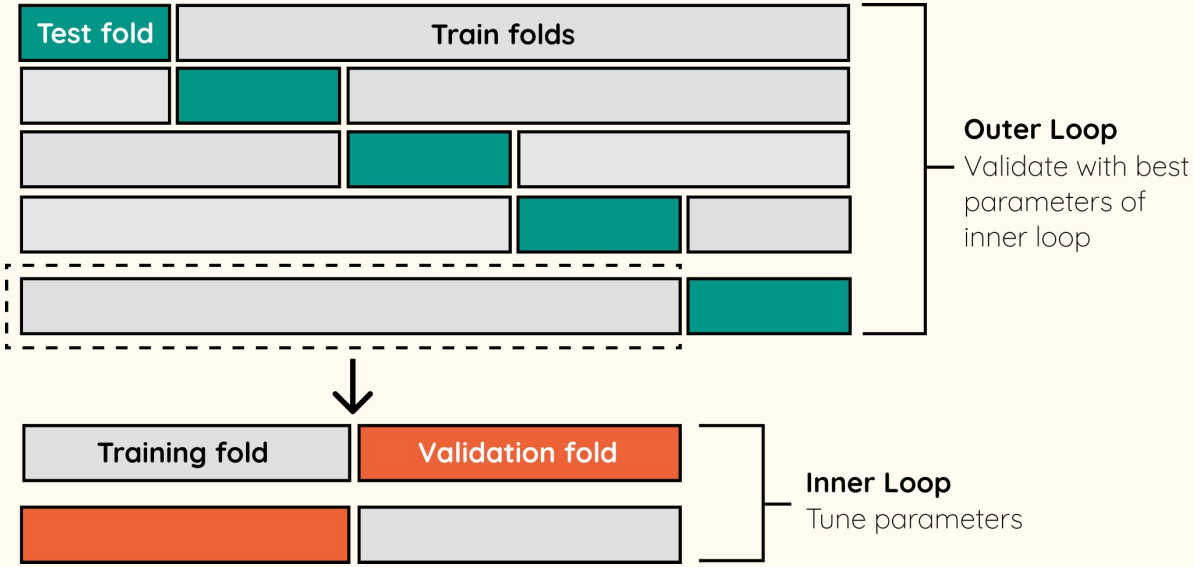
Métrica	Média	Desvio Padrão	Intervalo de confiança
Acurácia	0.573	0.026	0.573 ± 0.023
Precisão	0.573	0.026	0.573 ± 0.026
Cobertura	0.524	0.072	0.524 ± 0.063
F-measure	0.516	0.063	0.516 ± 0.055

Janela de Parzen (Fundamentação)

1. $$\hat{p}(\mathbf{x}) = \frac{1}{n}\frac{1}{h_1 \dots h_p}\sum_{i=1}^{n}\prod_{j=1}^{p} K_j\left(\frac{x_j - x_{ij}}{h_j}\right)$$

2. $$K(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$$

Janela de Parzen (Hiperparâmetros)



Janela de Parzen (Desempenho)

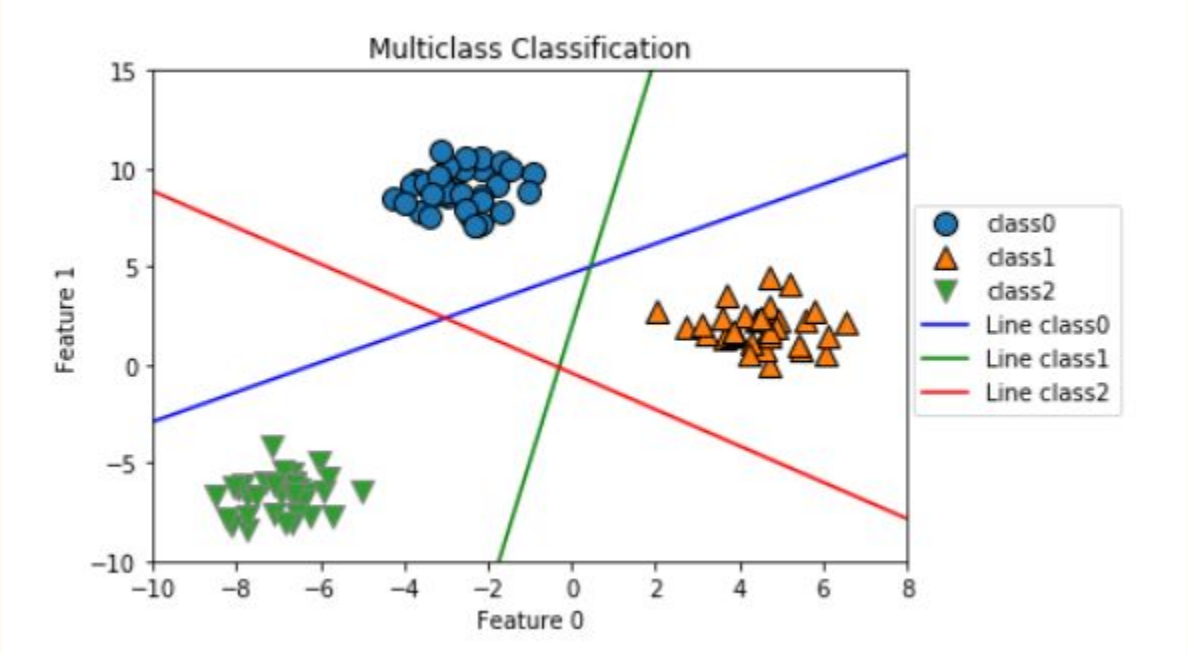
Métrica	Média	Desvio Padrão	Intervalo de confiança
Acurácia	0.575	0.035	0.575 ± 0.031
Precisão	0.614	0.045	0.614 ± 0.039
Cobertura	0.551	0.023	0.551 ± 0.020
F-Measure	0.559	0.024	0.559 ± 0.021

Regressão Logística (Fundamentação)

1. $$\frac{1}{1+\exp\{-\mathbf{x}^\top \boldsymbol{\theta}\}}$$

2. $$\Theta_{n+1} = \Theta_n - \alpha \frac{\partial}{\partial \Theta_n} J(\Theta_n)$$

Regressão Logística (One vs Rest)



Regressão Logística (Desempenho)

Métrica	Média	Desvio Padrão	Intervalo de confiança
Acurácia	0.542	0.025	0.542 ± 0.022
Precisão	0.407	0.021	0.407 ± 0.018
Cobertura	0.311	0.016	0.311 ± 0.014
F-Measure	0.331	0.015	0.331 ± 0.013

Voto Majoritário (Fundamentação)

1. Combinação de classificadores
2. Classe determinada pelo voto da maioria

Voto Majoritário (Desempenho)

Métrica	Média	Desvio Padrão	Intervalo de confiança
Acurácia	0.583	0.030	0.580 ± 0.026
Precisão	0.604	0.045	0.609 ± 0.039
Cobertura	0.538	0.041	0.537 ± 0.036
F-Measure	0.549	0.039	0.549 ± 0.034

Testes de Hipótese

- Teste de Friedman
- Teste de Nemenyi

Teste de Friedman

Metricas	Bayseano Gaussiano	KNN	Parzen	Regressao Logistica	Ensemble
Acuracia	0.557	0.573	0.575	0.542	0.583
Precisao	0.519	0.573	0.614	0.407	0.604
Cobertura	0.541	0.524	0.551	0.311	0.538
F-Measure	0.513	0.516	0.559	0.331	0.549
Ranqueamento	Bayseano Gaussiano	KNN	Parzen	Regressao Logistica	Ensemble
Acuracia	4	3	2	5	1
Precisao	4	3	1	5	2
Cobertura	2	4	1	5	3
F-Measure	4	3	1	5	2
Average	3.5	3.25	1.25	5	2

Teste de Friedman

$$\chi_F^2 = \frac{12N}{k(k+1)} \left[\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4} \right]$$

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2}$$

F Table for α = 0.05

/	df_1=1	2	3	4	5	6	7	8	9	10	12
df_2=1	161.4476	199.5000	215.7073	224.5832	230.1619	233.9860	236.7684	238.8827	240.5433	241.8817	243.9060
2	18.5128	19.0000	19.1643	19.2468	19.2964	19.3295	19.3532	19.3710	19.3848	19.3959	19.4125
3	10.1280	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123	8.7855	8.7446
4	7.7086	6.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	5.9988	5.9644	5.9117
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725	4.7351	4.6777
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990	4.0600	3.9999
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767	3.6365	3.5747
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881	3.3472	3.2839
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789	3.1373	3.0729
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204	2.9782	2.9130
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962	2.8536	2.7876
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964	2.7534	2.6866
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144	2.6710	2.6037
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458	2.6022	2.5342
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876	2.5437	2.4753

Teste de Nemenyi

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}$$

Valores de q_α :

k	2	3	4	5	6	7	8	9	10
Nemenyi	1.960	2.343	2.569	2.728	2.850	2.949	3.031	3.102	3.164

Ranqueamento médio

Bayseano Gaussiano	KNN	Parzen	Regressao Logistica	Ensemble
3.5	3.25	1.25	5	2

Perguntas?