

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
INSTITUTO DE MATEMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

PAULO RENATO CARVALHO DE AZEVEDO FILHO

QUAL A INFLUÊNCIA DO NÚMERO DE OBSERVAÇÕES E DO NÚMERO DE
ATRIBUTOS NA CLASSIFICAÇÃO DE GAUSSIANAS MULTIVARIADAS?

RIO DE JANEIRO
2023

PAULO RENATO CARVALHO DE AZEVEDO FILHO

QUAL A INFLUÊNCIA DO NÚMERO DE OBSERVAÇÕES E DO NÚMERO DE
ATRIBUTOS NA CLASSIFICAÇÃO DE GAUSSIANAS MULTIVARIADAS?

Trabalho de conclusão de curso de graduação apresentado ao Instituto de Computação da Universidade Federal do Rio de Janeiro como parte dos requisitos para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. Daniel Sadoc Menasche

RIO DE JANEIRO

2023

CIP - Catalogação na Publicação

PP667c Pires, Romeu Inojosa Lustosa
Construindo redes eficientes e robustas
minimizando distâncias em grafos biconexos / Romeu
Inojosa Lustosa Pires. -- Rio de Janeiro, 2021.
45 f.

Orientador: Daniel Sadoc Menasché.
Trabalho de conclusão de curso (graduação) -
Universidade Federal do Rio de Janeiro, Instituto
de Matemática, Bacharel em Ciência da Computação,
2021.

1. Grafos. 2. Algoritmos. I. Menasché, Daniel
Sadoc, orient. II. Título.

PAULO RENATO CARVALHO DE AZEVEDO FILHO

QUAL A INFLUÊNCIA DO NÚMERO DE OBSERVAÇÕES E DO NÚMERO DE
ATRIBUTOS NA CLASSIFICAÇÃO DE GAUSSIANAS MULTIVARIADAS?

Trabalho de conclusão de curso de graduação apresentado ao Instituto de Computação da Universidade Federal do Rio de Janeiro como parte dos requisitos para obtenção do grau de Bacharel em Ciência da Computação.

Aprovado em 5 de setembro de 2023

BANCA EXAMINADORA:

Prof. Daniel Sadoc Menasché,
Orientador
Ph.D. (UFRJ)

Prof. João Antonio Recio da Paixão
D.Sc. (UFRJ)

Prof. João Ismael D. Pinheiro
D.Sc. (UFRJ)

Prof. Pedro Henrique Cruz Caminha
D.Sc. (UFRJ)

RESUMO

Em um problema de classificação binária, o principal indicador de qualidade do classificador é a sua taxa de erro esperada, que idealmente deve ser o menor possível. Neste artigo, analisamos o efeito conjunto de mudanças na cardinalidade do conjunto de dados, no número de recursos ativos, etc. sobre a taxa de erro esperada de um classificador. Uma de nossas principais preocupações foi investigar o que acontece à medida que a cardinalidade do conjunto de dados cresce indefinidamente.

Para esse fim, adotamos uma abordagem experimental baseada em dados simulados. Nossos resultados mostram que aumentar a cardinalidade do conjunto de dados pode mitigar um eventual dano causado pela falta de recursos importantes. No entanto, também indica que a diminuição da taxa de erro esperada é muito lenta à medida que a cardinalidade do conjunto de dados cresce. Consequentemente, um possível trade-off entre cardinalidade e dimensionalidade por si só parece ser, pelo menos, controverso nesse contexto.

Palavras-chave: Classificadores; SVM; Erro de Bayes; Desempenho.

LISTA DE ILUSTRAÇÕES

Figura 1 – $\sigma_3 = [1, 1, 4], \rho_3 = [-0.8, 0, 0]$ e $\sigma_2 = [1, 1], \rho_2 = [-0.8]; n = 1024$	36
Figura 2 – $\sigma_3 = [1, 1, 4], \rho_3 = [0.8, 0, 0]$ e $\sigma_2 = [1, 1], \rho_2 = [0.8]; n = 1024$	36
Figura 3 – $\sigma_2 = [1, 1], \rho_2 = [0.8]$ e $\sigma_2 = [1, 1], \rho_2 = [-0.8]$	38
Figura 37 – Cenário 4 - indicadores 4/4	39
Figura 4 – Cenário 1 - indicadores 1/4	40
Figura 5 – Cenário 1 - indicadores 2/4	41
Figura 6 – Cenário 1 - indicadores 3/4	42
Figura 7 – Cenário 1 - indicadores 4/4	43
Figura 8 – Cenário 1 - $\hat{L}(\hat{h}^{(D)})$ vs σ_3	44
Figura 9 – Cenário 1 - $\hat{L}(\hat{h}^{(D)})$ vs n	45
Figura 10 – Cenário 1 - $L(\hat{h}^{(D)})$ vs σ_3	46
Figura 11 – Cenário 1 - $L(\hat{h}^{(D)})$ vs n	47
Figura 12 – Cenário 1 - $\hat{L}(\hat{h}^{(D)}, D')$ vs σ_3	48
Figura 13 – Cenário 1 - $\hat{L}(\hat{h}^{(D)}, D')$ vs n	49
Figura 14 – Cenário 2 - Indicadores 1/4	50
Figura 15 – Cenário 2 - Indicadores 2/4	51
Figura 16 – Cenário 2 - Indicadores 3/4	52
Figura 17 – Cenário 2 - Indicadores 4/4	53
Figura 18 – Cenário 2 - $\hat{L}(\hat{h}^{(D)})$ vs ρ_{12}	54
Figura 19 – Cenário 2 - $\hat{L}(\hat{h}^{(D)})$ vs n	55
Figura 20 – Cenário 2 - $L(\hat{h}^{(D)})$ vs ρ_{12}	56
Figura 21 – Cenário 2 - $L(\hat{h}^{(D)})$ vs n	57
Figura 22 – Cenário 2 - $\hat{L}(\hat{h}^{(D)}, D')$ vs ρ_{12}	58
Figura 23 – Cenário 2 - $\hat{L}(\hat{h}^{(D)}, D')$ vs n	59
Figura 24 – Cenário 3 - indicadores 1/4	60
Figura 25 – Cenário 3 - indicadores 2/4	61
Figura 26 – Cenário 3 - indicadores 3/4	62
Figura 27 – Cenário 3 - indicadores 4/4	63
Figura 28 – Cenário 3 - $\hat{L}(\hat{h}^{(D)})$ vs $\rho_{13} = \rho_{23}$	64
Figura 29 – Cenário 3 - $\hat{L}(\hat{h}^{(D)})$ vs n	65
Figura 30 – Cenário 3 - $L(\hat{h}^{(D)})$ vs $\rho_{13} = \rho_{23}$	66
Figura 31 – Cenário 3 - $L(\hat{h}^{(D)})$ vs n	67
Figura 32 – Cenário 3 - $\hat{L}(\hat{h}^{(D)}, D')$ vs $\rho_{13} = \rho_{23}$	68
Figura 33 – Cenário 3 - $\hat{L}(\hat{h}^{(D)}, D')$ vs n	69
Figura 34 – Cenário 4 - indicadores 1/4	70
Figura 35 – Cenário 4 - indicadores 2/4	71

Figura 36 – Cenário 4 - indicadores 3/4	72
Figura 38 – Cenário 4 - $\hat{L}(\hat{h}^{(D)})$ vs $\rho_{13} = \rho_{23}$	81
Figura 39 – Cenário 4 - $\hat{L}(\hat{h}^{(D)})$ vs n	82
Figura 40 – Cenário 4 - $L(\hat{h}^{(D)})$ vs $\rho_{13} = \rho_{23}$	83
Figura 41 – Cenário 4 - $L(\hat{h}^{(D)})$ vs n	84
Figura 42 – Cenário 4 - $\hat{L}(\hat{h}^{(D)}, D')$ vs $\rho_{13} = \rho_{23}$	85
Figura 43 – Cenário 4 - $\hat{L}(\hat{h}^{(D)}, D')$ vs n	86

SUMÁRIO

1	MOTIVAÇÃO E COMPROMISSO	9
1.1	METODOLOGIA	10
1.1.1	Descrevendo o problema a ser investigado	10
1.2	CONTRIBUIÇÕES	11
2	FORMULAÇÃO DO PROBLEMA E RESULTADOS FOR- MAIS	13
2.1	FORMULAÇÃO DO PROBLEMA	13
2.1.1	Contexto	13
2.1.2	Como Avaliar a Eficiência de um Classificador	13
2.1.3	Gaussiana Bi-variada.	14
2.1.4	Classificador SVM	15
2.1.5	Precisão na Medição do Erro à Medida Que a Cardinalidade Aumenta	16
2.2	MODELO ANALÍTICO	17
2.2.1	Erro de Bayes Correspondente a um Tamanho de Amostra Infinito ($n = \infty$).	17
2.2.2	Um Bissetor Simples Para $n = 2$	17
2.2.3	O Problema Geral de Classificação 3D	18
2.2.4	Cenários	19
2.3	SIMULAÇÃO	21
3	SIMULADOR	23
3.1	SLACGS	23
3.1.1	Relatórios Exportados	23
3.1.2	Imagens Exportadas	23
3.1.3	Funções de Erro	24
3.2	DEMO	24
3.2.1	1. Baixar e Instalar	24
3.2.2	2. Configurar Serviço de Relatório	25
3.2.3	3. Cenários de Experimento	25
3.2.4	4. Funções de Demonstração	25
3.3	CLASSE MODEL	26
3.3.1	Conteúdo	26
3.3.2	Construtor	26
3.4	CLASSE SIMULATOR	27

3.4.1	Conteúdo	27
3.4.2	Argumentos para a Simulação	27
3.4.3	Critérios de Parada de Convergência de Erro	28
3.4.4	Testes Após Simulação	28
3.4.5	Método run()	28
3.5	FUNÇÕES DO MÓDULO SIMULATOR.PY	30
3.5.1	Função de erro do Classificador h	30
3.5.2	Função de Erro Empírico	31
3.5.3	Função de Erro Teórico	32
3.6	CLASSE REPORT	33
3.6.1	Construtor	33
4	RESULTADOS	35
4.1	QUANDO X_3 CONTRIBUI PARA AUMENTAR O PODER DE DISCRIMINAÇÃO DE (X_1, X_2)?	35
4.1.1	Um modelo para entender o comportamento do risco de Bayes em duas dimensões	37
4.1.2	Estendendo para o caso de três dimensões	37
4.1.3	Comparando os cenários com dois e três atributos	38
5	TRABALHOS RELACIONADOS	87
6	CONCLUSÃO E TRABALHOS FUTUROS	89
	REFERÊNCIAS	90

1 MOTIVAÇÃO E COMPROMISSO

Suponhamos que os elementos de uma população podem ser divididos em grupos, sendo que a cada grupo corresponde um rótulo. Além disso, para cada elemento podem ser medidos alguns de seus atributos. E, supostamente, há uma relação (desconhecida) entre os atributos e o rótulo. O problema de classificação consiste, em essência, em obter uma regra que permita atribuir um rótulo a determinada observação não rotulada a partir de seus atributos. Essa regra é construída a partir de uma amostra de elementos para os quais tanto os atributos como os rótulos são conhecidos. Neste contexto, o número de observações usadas para treinamento do algoritmo de classificação, bem como o número de atributos em cada observação, são dois dos elementos chaves que influenciam na tarefa de obter uma regra de classificação.

Existe uma ampla literatura sobre o impacto no número de observações e do número de atributos no desempenho dos classificadores. Entretanto, a maioria dos trabalhos considera conjuntos de dados muito grandes, ou então foca em resultados assintóticos.

Neste trabalho, visamos entender o papel conjunto do número de observações e do número de atributos em tarefas de classificação. Para tal, consideramos um dos cenários mais simples possíveis, a saber, focando na classificação de observações advindas de modelos gaussianos bi-variados. Dada a simplicidade do modelo gerativo, somos capazes de extrair observações que eventualmente podem ser generalizadas para outros contextos. Em particular, focamos nas seguintes perguntas:

- Quando é possível trocar atributos por observações? Ou seja, na impossibilidade de se colher atributos adicionais, quando é possível alcançar o mesmo desempenho de classificação colhendo-se observações adicionais?
- Quando um atributo adicional contribui para aumentar o desempenho de classificação?
- Quais as diferenças essenciais entre classificação com 1, 2 ou 3 atributos? Em particular, quando passamos de 1 para 2 atributos, ou de 2 atributos para 3 atributos, o que faz com que o atributo adicional tenha uma contribuição expressiva em termos da taxa esperada de acertos do classificador?

Implementamos um simulador para gerar amostras gaussianas, e usamos o Support Vector Machine (SVM) para classificar as amostras. Então, provemos algumas respostas (parciais) para as perguntas acima. Em particular, identificamos cenários nos quais com poucas observações é vantajoso trabalhar com apenas 2 atributos, ao invés de 3, para evitar o problema do *overfitting*. Além disso, identificamos que com 3 atributos existe um ponto

Tabela 1 – probabilidades de erro SVM (multiplicado por 100) para $\rho = 0$

n	1 feature	$E_1(n)$, $\delta=1$	$E_2(n; \delta)$, 2 features					
			$\delta=2$	$\delta=3$	$\delta=4$	$\delta=5$	$\delta=6$	$\delta=7$
2	20.50	14.99	23.29	26.72	28.20	28.93	29.36	29.63
4	18.83	13.91	21.73	24.47	25.62	26.20	26.52	26.73
8	17.63	11.98	18.95	21.02	21.82	22.21	22.42	22.54
16	16.87	10.24	16.28	17.92	18.55	18.86	19.01	19.11
32	16.39	9.18	14.74	16.22	16.78	17.05	17.20	17.29
64	16.15	8.57	13.96	15.39	15.94	16.20	16.34	16.43
129	16.01	8.24	13.57	14.99	15.53	15.79	15.93	16.02
256	15.94	8.06	13.38	14.79	15.33	15.59	15.73	15.82
512	15.90	7.96	13.28	14.69	15.23	15.49	15.63	15.72
1024	15.88	7.91	13.23	14.64	15.18	15.44	15.58	15.67
n^*			12	28	46	70	98	130
Linear regression of error probability: $E(n) = \alpha + \beta/n$.								
α	0.159	0.080	0.132	0.146	0.151	0.154	0.155	0.156
β	0.142	0.333	0.468	0.522	0.542	0.552	0.556	0.558
n^*			12	30	52	78	106	136

de corte tal que, acima de terminado número de observações, é sempre vantajoso contar com os 3 atributos.

1.1 METODOLOGIA

Para ilustrar a relação entre o tamanho da amostra, o número de atributos e a probabilidade de erro assintótico do classificador treinado, consideramos um conjunto de dados simples com 1 ou 2 atributos. A distribuição condicional de cada amostra dada sua classe é admitiremos como uma distribuição gaussiana unidimensional ou bidimensional (por exemplo, como em (WILLETT; SWASZEK; BLUM, 2000)).

No caso bidimensional, temos um coeficiente de correlação ρ ; um dos atributos tem variância um (1) e o outro tem variância δ^2 , observando que para $\rho = 0$ o poder discriminatório do segundo atributo diminui à medida que δ aumenta. Embora esta carga de trabalho seja reconhecidamente simples, dados coletados na prática muitas vezes podem ser representados por meio de uma distribuição gaussiana bivariada. Além disso, este modelo simples já atende aos nossos propósitos, ou seja, mostrar que (a) dependendo do poder discriminatório do segundo atributo, pode valer a pena ignorá-lo e (b) amostras adicionais podem compensar a falta de um atributo. Como podemos ver na Tabela 1.

1.1.1 Descrevendo o problema a ser investigado

Temos um problema de classificação com duas classes, onde:

- O vetor aleatório (X, Y) segue uma distribuição de probabilidade conjunta P .

- X é um vetor de características d-dimensional.
- A resposta Y pode ser -1 ou +1.
- Consideramos apenas classificadores em um dicionário específico H (por exemplo, o conjunto de classificadores lineares).

Nosso objetivo é usar um conjunto de dados D com n pares (x_i, y_i) para escolher um classificador "bom" $h \in H$. Mas o que queremos dizer com um classificador "bom"? Quanto menor a taxa de erro esperada, mais eficiente será considerado um classificador. E, como estamos trabalhando com dados simulados, a melhor maneira de avaliar o desempenho esperado de um classificador é calculando sua perda teórica. Agora, investigaremos uma possível compensação entre a dimensionalidade e a cardinalidade no contexto de um problema de classificação binária.

A ideia é comparar duas situações A e B, onde:

- Na situação A, temos dimensionalidade d_A e cardinalidade n_A .
- Na situação B, temos dimensionalidade d_B e cardinalidade n_B .
- Todas os atributos presentes em A também estão presentes em B.

Se $d_A < d_B$, a menor dimensionalidade da situação A pode ser compensada por um aumento na cardinalidade do conjunto de dados, ou seja, fazendo $n_A > n_B$?

É claro que responder a essa pergunta não é apenas uma questão de comparar dimensionalidades e cardinalidades. Suponha que consideremos adicionar novos atributos ao conjunto de dados A para aumentar sua capacidade de discriminação. De fato, existem atributos que, quando adicionadas, melhoram significativamente o poder discriminatório do modelo, enquanto existem outros atributos cuja adição seria praticamente inútil para esse propósito. Então, não é apenas uma questão de quantos novos atributos são adicionados ao conjunto de dados. Depende da contribuição efetiva de cada novo atributo em termos de aumentar o poder discriminatório do modelo entre as duas populações envolvidas no problema. Nos exemplos deste TCC em que o poder discriminatório de cada atributo é o mesmo, tornam-se mais significativas quaisquer considerações sobre o número de características ativas em um contexto específico.

1.2 CONTRIBUIÇÕES

Nossa principal contribuição consiste num simulador para avaliar o desempenho do algoritmo de classificação SVM sob um conjunto de dados gaussiana. O código fonte gerado está disponível no Github no seguinte link:

- [<https://github.com/paulorenatoaz/slacgs>](https://github.com/paulorenatoaz/slacgs)

Além disso, com o simulador colhemos várias observações interessantes, para gaussianas bivariadas apresentadas ao decorrer deste trabalho.

2 FORMULAÇÃO DO PROBLEMA E RESULTADOS FORMAIS

2.1 FORMULAÇÃO DO PROBLEMA

O processo de classificação envolve duas etapas: o treinamento do classificador com amostras rotuladas e, em seguida, a inferência das classes de amostras não rotuladas. Nossos dois principais atributos de interesse são a cardinalidade do conjunto de dados, denotada por n , que corresponde ao número de amostras no conjunto de treinamento, e a dimensionalidade do conjunto de dados, denotada por d , que corresponde ao número de atributos em cada amostra. Um dos nossos objetivos é avaliar como n e d afetam simultaneamente o desempenho do classificador, medido pela sua probabilidade de erro esperada.

2.1.1 Contexto

Consideramos um problema de classificação binária, onde cada amostra é retirada de uma de duas classes, denotadas pelos rótulos -1 e $+1$. O conjunto de amostras compreende um conjunto de dados D com n pares ordenados (\mathbf{x}_i, y_i) , onde $\mathbf{x}_i \in \mathbb{R}^d$ e $y_i \in -1, +1$, para $i = 1, \dots, n$. O conjunto de dados D é utilizado para treinar um classificador que minimiza a probabilidade de erro empírico em D . Sejam X e Y as variáveis aleatórias correspondentes ao vetor de atributos e à classe alvo, admite-se que a distribuição conjunta de probabilidade para o par (X, Y) é desconhecida.

Um classificador é uma função $h : \mathbf{x} \rightarrow h(\mathbf{x})$ que atribui um rótulo $h(\mathbf{x}) \in -1, +1$ a qualquer vetor de atributos $\mathbf{x} \in \mathbb{R}^d$. A probabilidade de erro (ou perda) do classificador h é denotada por $L(h)$ e é dada por $L(h) = P(Y \neq h(X))$. Um dicionário H (também conhecido como conjunto de hipóteses ou viés do espaço de busca) é um conjunto de classificadores (por exemplo, classificadores lineares, quadráticos ou retangulares). O problema de aprendizado consiste em escolher um classificador do dicionário H como a "solução" para nosso problema de classificação. É bem conhecido que o classificador Bayesiano, que classifica cada entrada de acordo com $\text{argmax}_y P(Y = y | X = x)$, é globalmente ótimo (DEVROYE; GYÖRFI; LUGOSI, 2013; FARAGÓ; LUGOSI, 1993); o erro de Bayes é a probabilidade mínima de erro. Denotamos por $h^{(B)}$ e $L(h^{(B)})$, ou h^* e $L(h^*)$, o classificador de Bayes e sua probabilidade de erro.

2.1.2 Como Avaliar a Eficiência de um Classificador

Qual deve ser nosso objetivo ao enfrentar um problema de classificação binária? O senso comum indica que devemos procurar um classificador com a menor taxa de erro possível. Suponhamos que tenhamos um conjunto de dados reais D sobre um determinado

tópico. Nesse caso, qual é a taxa de erro que realmente importa? Vamos supor que os dados disponíveis sejam usados para calibrar um classificador cuja taxa de erro empírica seja a menor possível quando aplicada ao conjunto de dados D . Então, é claro que a melhor maneira de avaliar a eficiência desse classificador seria prever sua taxa de erro empírica sempre que aplicado a outro conjunto de dados independente D' . Se tivermos a sorte de obter um conjunto de dados com muitas observações, a solução usual seria dividir aleatoriamente o conjunto de dados disponível em treinamento e teste. Isso evitaria subestimar a taxa de erro real, usando os mesmos dados para calibrar o classificador e também para avaliar sua eficácia. Nesse contexto, é importante distinguir entre as taxas de erro empíricas e teóricas associadas a um determinado classificador: a taxa empírica é a taxa de treinamento (que tende a subestimar a taxa de erro real); enquanto a taxa de teste fornece uma estimativa mais confiável de sua taxa de erro esperada. Além disso, se não for possível obter um conjunto de dados com muitas observações, então uma técnica de reamostragem (como validação cruzada, bootstrap, etc.) poderia ser usada para evitar tal viés na estimativa da taxa de erros de classificação. Agora, e se trabalharmos com dados simulados provenientes de um modelo probabilístico conhecido? Nesse caso, para um determinado classificador, é sempre possível usar a teoria da probabilidade para calcular (eventualmente por métodos numéricos) sua taxa de erro teórica, sem ter que avaliar sua eficácia por meio de dados de teste. Como este é um trabalho metodológico, escolhemos trabalhar com dados simulados, o que facilita muito a tarefa de avaliar a eficácia de cada classificador de maneira imparcial. Felizmente, conclusões baseadas nesse tipo de abordagem também são aplicáveis a situações concretas em que trabalhamos com dados reais. Por outro lado, quando enfrentamos um problema de classificação binária, quais são nossos "botões de controle"? Entre eles, podemos mencionar:

- Cardinalidade do conjunto de dados n , ou seja, o número de observações disponíveis.
- A dimensionalidade do problema d , ou seja, o número de atributos disponíveis.
- O poder discriminatório de cada atributo, sozinho ou na presença das outros atributos.
- Dicionário H , do qual escolheremos nosso classificador.

Então, valeria a pena perguntar como uma escolha particular de combinação desses parâmetros de entrada afetará a taxa esperada de erros de classificação em nosso modelo matemático.

2.1.3 Gaussiana Bi-variada.

Para avaliar o poder de generalização dos classificadores considerados, temos pelo menos três alternativas:

1. Amostrar D a partir de um modelo probabilístico conhecido, por exemplo, como em (NAKKIRAN et al., 2020; D'ASCOLI; SAGUN; BIROLI, 2020) e utilizar a própria amostra de treino D para avaliar o desempenho empírico de classificação de h , como em (2.3).
2. Avaliar o desempenho de h utilizando equações do modelo linear probabilístico como (2.4), (2.6) ou (2.17)
3. Avaliar o desempenho de h em um dataset de teste D' , gerado a partir do mesmo modelo probabilístico utilizado para amostrar D .

Em particular, consideremos um problema de classificação de duas dimensões, no qual amostras de cada classe são retiradas de uma distribuição Gaussiana bivariada. Admitimos prioridades iguais para as duas classes, ou seja, $P(Y = +1) = P(Y = -1) = 0,5$. A distribuição condicional do vetor de atributos $X = (X_1, X_2)$ dado Y é dada por

$$(X_1, X_2) \sim \begin{cases} \mathcal{N}((+1, +1), \Sigma), & \text{if } y = +1, \\ \mathcal{N}((-1, -1), \Sigma), & \text{if } y = -1, \end{cases} \quad (2.1)$$

onde $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ denota uma distribuição Gaussiana bivariada com vetor médio $\boldsymbol{\mu} \in \mathbb{R}^2$ e matriz de covariância Σ ,

$$\Sigma = \begin{pmatrix} 1 & \rho\delta \\ \rho\delta & \delta^2 \end{pmatrix} \quad (2.2)$$

com $\delta \geq 1$ e $|\rho| \leq 1$. Note que δ é o desvio padrão condicional de atributo X_2 , dado Y . Se $\delta > 1$, o atributo X_2 tem menor poder discriminatório do que o atributo X_1 : quanto maior o valor de δ , menos informativo é o atributo X_2 .

Note que consideramos conjuntos de dados equilibrados, com o mesmo número de amostras em cada classe.

2.1.4 Classificador SVM

Nosso dicionário H é formado por todos os classificadores lineares, ou seja, linhas retas. Uma abordagem ingênua para o aprendizado consiste em procurar pela linha que minimize a taxa de erro empírica, por exemplo, por meio de uma busca em grade. No entanto, o problema pode admitir múltiplas soluções, exigindo uma estratégia fundamentada para desempatar (BLANCHARD; BOUSQUET; MASSART, 2008). Por isso, consideramos máquinas de vetores de suporte (SVM) lineares, que têm uma série de propriedades desejáveis, incluindo:

1. uma estratégia fundamentada para determinar o melhor discriminador, maximizando a distância, conhecida como *margem*, entre a linha de separação e as duas classes de pontos a serem separados (BLANCHARD; BOUSQUET; MASSART, 2008);

2. um custo computacional polinomial (BOTTOU; LIN, 2007; LIST; SIMON, 2009) e
3. convergência para uma solução ótima, sob condições brandas (HSIEH et al., 2008; GLASMACHERS, 2010; VERT; VERT; SCHÖLKOPF, 2006).

O SVM linear depende de um único parâmetro, o custo de classificação incorreta, que aparece como termo de regularização na formulação de Lagrange e é denotado por γ (R... , 2021). Ao longo deste trabalho, deixamos $\gamma = 1$, seu valor padrão; experimentamos com outros valores, mas os resultados permaneceram aproximadamente os mesmos.

2.1.5 Precisão na Medição do Erro à Medida Que a Cardinalidade Aumenta

Outro aspecto que não deve ser negligenciado é a precisão com que podemos estimar a taxa de erro do nosso procedimento de classificação. E, é claro, também é interessante investigar como cada um dos parâmetros de entrada afeta essa precisão. É isso que vamos discutir agora. Em particular, investigaremos o que acontece com a precisão da estimativa da taxa de erro à medida que a cardinalidade do conjunto de dados n tende ao infinito.

Assim, a partir de agora, seja H um dicionário específico. Além disso, suponha que tenhamos um conjunto de dados D , que é representativo do fenômeno em estudo e formado pelos pares (x_i, y_i) , $i = 1, 2, \dots, n$, onde $x_i \in \mathbb{R}^d$ e $y_i \in \{-1, +1\}$, para todo $i = 1, 2, \dots, n$.

Então, para cada classificador h em H , podemos calcular sua taxa de erro empírica $\hat{L}(h)$, ou seja, sua proporção de observações mal classificadas no conjunto de dados D :

$$\hat{L}(h) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(y_i \neq h(x_i)) \quad (2.3)$$

Seja $\hat{h}^{(D)}$ o melhor classificador empírico no dicionário H , ou seja, o classificador em H cuja taxa de erro empírica é mínima. Sob essas condições, podemos definir quatro taxas de erro:

- $\min_{h \in H} L(h) =$ a menor taxa de erro teórico alcançável em H
- $L(\hat{h}^{(D)}) =$ a taxa de erro teórica do classificador $\hat{h}^{(D)}$
- $\hat{L}(\hat{h}^{(D)}) =$ a taxa de erro empírica do classificador $\hat{h}^{(D)}$
- $\hat{L}(\hat{h}^{(D)}, D') =$ a taxa de erro do classificador $\hat{h}^{(D)}$ utilizando um dataset teste D'

É um fato conhecido que essas perdas obedecem necessariamente a uma ordem:

$$0 \leq \hat{L}(\hat{h}^{(D)}) \leq L(h^*) \leq \min_{h \in H} L(h) \leq L(\hat{h}^{(D)}),$$

onde $L(h^*)$ é a taxa de erro Bayesiana (também conhecida como risco Bayesiano), ou seja, a perda teórica mínima geralmente alcançável para esse problema de classificação particular.

2.2 MODELO ANALÍTICO

2.2.1 Erro de Bayes Correspondente a um Tamanho de Amostra Infinito ($n = \infty$).

Começamos considerando o caso em que apenas o atributo x_1 está disponível. Temos um problema de classificação unidimensional, no qual buscamos o melhor ponto de corte c . A probabilidade de erro, sob a configuração da Seção 2.1.3, é dada por

$$L_1(c) = (1 + \Phi(c - 1) - \Phi(c + 1))/2 \quad (2.4)$$

onde $\Phi(\cdot)$ é a função de distribuição acumulada de uma Gaussiana com média zero e variância um. O separador ótimo é $\tilde{c} = 0$, e

$$L_1(0) = L_1(h_1^*) = 1 - \Phi(1) = 0.1586553. \quad (2.5)$$

A seguir, consideramos o caso em que ambos os atributos estão disponíveis. Dado um classificador caracterizado pela reta $x_2 = a + bx_1$:

$$L_2(a, b) = \frac{1}{2} \left(1 - \Phi \left(\frac{|a + b - 1|}{\sqrt{S}} \right) \right) + \frac{1}{2} \left(1 - \Phi \left(\frac{|a - b + 1|}{\sqrt{S}} \right) \right), \quad (2.6)$$

onde $S = (\rho\delta - b)^2 + \delta^2(1 - \rho^2)$. Pode-se verificar que a solução ótima é dada por uma reta reta através da origem, $\tilde{a} = 0$, com uma inclinação de $\tilde{b} = \delta(\delta - \rho)/(\delta\rho - 1)$. Como o classificador $h^{(B)}$ pertence ao dicionário formado por retas em \mathbb{R}^2 ,

$$L(h^{(B)}) = L(h_2^*) = L_2 \left(0, \frac{\delta(\delta - \rho)}{\delta\rho - 1} \right) = 1 - \Phi \left(\frac{1}{\delta} \sqrt{\frac{\delta^2 - 2\rho\delta + 1}{1 - \rho^2}} \right). \quad (2.7)$$

Interessante notar que, ao tomar a derivada parcial da expressão acima com relação a ρ e igualando-a a zero, conclui-se que o limite superior de $L(h^{(B)})$ é alcançado em $\rho = 1/\delta$,

$$0 \leq L(h^{(B)}; \rho) \leq L(h^{(B)}; 1/\delta) = 1 - \Phi(1). \quad (2.8)$$

O limite inferior do erro é alcançado para $\rho = 1$ e $\rho = -1$, correspondendo a cenários em que X_2 é uma função determinística de X_1 . À medida que ρ aumenta de -1 para $1/\delta$, o erro aumenta; à medida que ρ é ainda mais aumentado, o erro diminui para zero para valores suficientemente grandes de n .

2.2.2 Um Bissetor Simples Para $n = 2$

Os resultados da seção anterior são resultados assintóticos para um tamanho de amostra infinito. A seguir, consideramos o oposto extremo: um conjunto de dados composto por duas amostras, uma em cada classe. Denotamos por $h_1^{(SVM)}$ e $h_2^{(SVM)}$ o classificador obtido usando SVM, com um atributo (X_1) e dois atributos, respectivamente. Então,

as soluções SVM correspondem a bissetores simples como as fronteiras de separação. Os erros esperados são dados por

$$E(L(h_1^{(SVM)})) = \int_{c=-\infty}^{+\infty} L_1(c) \frac{1}{\sqrt{\pi}} e^{-c^2} dc \approx 0.2070336 \quad (2.9)$$

e

$$E(L(h_2^{(SVM)})) = \int_{\mathbf{w} \in \mathbb{R}^2} \int_{\mathbf{v} \in \mathbb{R}^2} L_2(g(\mathbf{v}, \mathbf{w})) \phi_{+1}(\mathbf{v}) \phi_{-1}(\mathbf{w}) d\mathbf{v} d\mathbf{w} \quad (2.10)$$

onde as expectativas são tomadas sobre conjuntos de treinamento aleatórios com duas observações cada, extraídas de gaussianas bivariadas, $\mathbf{v} = (v_1, v_2)$, $\mathbf{w} = (w_1, w_2)$, $\phi_t(\cdot)$ é a função de densidade de probabilidade da gaussiana bivariada correspondente à classe t , e $g(\mathbf{v}, \mathbf{w})$ é uma função que mapeia um par de pontos em sua interseção e inclinação do bissetor perpendicular.

$$g(\mathbf{v}, \mathbf{w}) = \left(\frac{w_1 - v_1}{w_2 - v_2} \cdot \frac{v_1 + w_1}{2} + \frac{v_2 + w_2}{2}, -\frac{w_1 - v_1}{w_2 - v_2} \right). \quad (2.11)$$

Comparando $L_1(0)$ a $L(h_1^{(SVM)})$, (2.5) a (2.9), nós avaliamos o impacto do tamanho finito da amostra quando apenas um atributo está disponível e comparando o desempenho da função de perda $L(h^{(B)})$ ao da função de perda $L(h_2^{(SVM)})$, (2.7) a (2.10), quando dois atributos estão disponíveis.

Para $\rho = 0$, os valores das funções de perda correspondentes são apresentados nas linhas $n = 1,024$ e $n = 2$ da Tabela 1, cujos valores concordam em até três dígitos decimais com as expressões (2.9)-(2.10) e (2.5)-(2.7), respectivamente. Comparando (2.9) com (2.10), cujos valores correspondem aos elementos da primeira linha na Tabela 1, nota-se que quando o tamanho da amostra é pequeno e $\delta \geq 2$, é benéfico usar menos atributos. Por outro lado, comparando (2.5) a (2.7), cujos valores correspondem aos elementos da última linha na Tabela 1, nota-se que quando o tamanho da amostra é pequeno e $1 \leq \delta \leq 7$, é benéfico usar mais atributos.

2.2.3 O Problema Geral de Classificação 3D

Admitindo que as matrizes de covariância condicional são iguais, o problema de classificação 3D pode ser formulado da seguinte forma:

$$(X_1, X_2, X_3) \sim \begin{cases} \mathcal{N}((\mu_1+, \mu_2+, \mu_3+), \Sigma), & \text{se } Y = +1 \\ \mathcal{N}((\mu_1-, \mu_2-, \mu_3-), \Sigma), & \text{se } Y = -1 \end{cases} \quad (2.12)$$

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 \end{bmatrix} \quad (2.13)$$

Também admitiremos:

$$P(Y = +1) = P(Y = -1) = \frac{1}{2} \quad (2.14)$$

e:

$$\mu_+ = (1, 1, 1), \mu_- = (-1, -1, -1) \quad (2.15)$$

Agora, suponha que a equação do plano de separação ótimo seja:

$$X_3 = d^* + e^* X_1 + f^* X_2 \quad (2.16)$$

Para minimizar $P(\text{Erro})$ em relação a d^* , e^* e f^* , devemos igualar a zero as suas derivadas parciais em relação a esses coeficientes. Então, pode-se mostrar que:

- $d^* = 0$ (devido à simetria geral em relação à origem $(0, 0, 0)$)
- Tanto e^* quanto f^* dependem dos seis parâmetros que definem a matriz Σ , a saber: $\sigma_1, \sigma_2, \sigma_3, \rho_{12}, \rho_{13}, \rho_{23}$.
- A probabilidade mínima de erro (risco de Bayes) é $P(\text{Erro}) = 1 - \phi\left(\frac{|1-e^*-f^*|}{\sqrt{\Delta}}\right)$, onde:

- $\Delta = A^2 + B^2 + \lambda_{33}^2$, com: $A = e^* \lambda_{11} + f^* \lambda_{21} - \lambda_{31}$, $B = f^* \lambda_{22} - \lambda_{32}$
- $\lambda_{11} = \sigma_1$, $\lambda_{12} = 0$, $\lambda_{13} = 0$
- $\lambda_{21} = \rho_{12} \sigma_2$, $\lambda_{22} = \sigma_2 \sqrt{1 - \rho_{12}^2}$, $\lambda_{23} = 0$
- $\lambda_{31} = \rho_{13} \sigma_3$, $\lambda_{32} = \frac{(\rho_{23} - \rho_{12} \rho_{13}) \sigma_3}{\sqrt{1 - \rho_{12}^2}}$, $\lambda_{33} = \sigma_3 \sqrt{1 - \rho_{13}^2} - \frac{(\rho_{23} - \rho_{12} \rho_{13})^2}{(1 - \rho_{12})^2}$
- (Todo isso vem da decomposição de Cholesky da matriz Σ .)

Entretanto essa formulação inclui muitos (seis) parâmetros e o problema de otimização matemática torna-se muito complexo.

2.2.4 Cenários

Para simplificar a matemática, consideraremos quatro cenários específicos. Em cada cenário, a ideia é reduzir o número de parâmetros livres, permitindo obter expressões analíticas para os coeficientes ótimos do plano, e^* e f^* , e o risco de Bayes.

1. As 3 atributos são condicionalmente independentes em pares, dado o rótulo, ou seja,

$$\rho_{12} = \rho_{13} = \rho_{23} = 0.$$

Aqui pode ser demonstrado que:

- a) $e^* = -\frac{\sigma_3^2}{2\sigma_1^2}$
b) $f^* = -\frac{\sigma_3^2}{2\sigma_2^2}$
c) Risco Bayesiano = $1 - \phi\left(\frac{\sqrt{1-e^*-f^*}}{\sigma_3}\right)$

2. X_1 e X_2 são condicionalmente correlacionados, (X_1, X_2) e X_3 são condicionalmente independentes

- Os atributos x_1 e x_2 estão condicionalmente correlacionadas, ou seja, $\rho_{12} = \rho$ pode ser $\neq 0$
- Os atributos x_1 e x_3 são condicionalmente independentes, ou seja, $\rho_{13} = 0$
- Os atributos x_2 e x_3 são condicionalmente independentes, ou seja, $\rho_{23} = 0$
- $\sigma_1 = \sigma_2 = \sigma$

Aqui, pode ser demonstrado que:

- a) $e^* = f^* = -\frac{\sigma_3^2}{2\sigma^2(1+\rho)}$
b) $P(\text{Error}) = 1 - \phi\left(\sqrt{1 - 2e^*\frac{1}{\sigma_3}}\right)$

3. X_1 e X_2 são condicionalmente independentes, (X_1, X_2) e X_3 são condicionalmente correlacionados

- Os atributos x_1 e x_2 são condicionalmente independentes, ou seja, $\rho_{12} = 0$
- Os atributos x_1 e x_3 são condicionalmente correlacionadas, ou seja, ρ_{13} pode ser $\neq 0$
- Os atributos x_2 e x_3 são condicionalmente correlacionadas, ou seja, ρ_{23} pode ser $\neq 0$
- $\rho_{13} = \rho_{23} = r$ (Restrição: $-\frac{\sqrt{2}}{2} < r < \frac{\sqrt{2}}{2}$)
- $\sigma_1 = \sigma_2 = \sigma$

Nesse cenário, pode ser demonstrado que:

- a) $e^* = f^* = \frac{\sigma_3}{(\sigma_3 - r\sigma)} \cdot \frac{\sigma}{(2r\sigma_3 - \sigma)}$
b) $P(\text{Erro}) = 1 - \phi\left(\frac{1-2e^*}{\sqrt{\Delta}}\right)$, onde $\Delta = 2(e\sigma - r\sigma_3)^2 + \sigma_3^2(1 - 2r^2)$

4. X_1 e X_2 condicionalmente correlacionados, (X_1, X_2) e X_3 condicionalmente correlacionados

- Todas os atributos igualmente dispersas ($\sigma_1 = \sigma_2 = \sigma_3 = \sigma$)

- Os atributos x_1 e x_2 são condicionalmente correlacionadas, isto é, $\rho_{12} = \rho$ pode ser $\neq 0$
- Os atributos x_1 e x_3 são condicionalmente correlacionadas, isto é, ρ_{13} pode ser $\neq 0$
- Os atributos x_2 e x_3 são condicionalmente correlacionadas, isto é, ρ_{23} pode ser $\neq 0$
- $\rho_{13} = \rho_{23} = r$
- Restrição: $-\sqrt{\frac{1+\rho}{2}} \leq r \leq \sqrt{\frac{1+\rho}{2}}$

Nesse cenário, pode ser demonstrado que:

$$\text{a)} \quad e^* = f^* = \frac{1-r}{2r-(1+\rho)}; \quad P(\text{Erro}) = 1 - \phi\left(\frac{1-2e^*}{\sqrt{\Delta}}\right), \quad \text{onde}$$

$$\text{b)} \quad \Delta = A^2 + B^2 + \lambda_{33}^2; \quad A = \sigma[e^*(1+\rho) - r]; \quad B = \sigma\left[\sqrt{1-\rho^2}e^* - \frac{r(1-\rho)}{\sqrt{1-\rho^2}}\right]; \\ \lambda_{33} = \sigma\sqrt{1-2r^2/(1+\rho)}$$

2.3 SIMULAÇÃO

Nossa intenção é comparar " x_1 , x_2 e x_3 presentes" com "apenas x_1 e x_2 presentes", em relação ao desempenho esperado do classificador, à medida que a cardinalidade do conjunto de dados n aumenta. Em particular, queremos observar como o limiar n^* se comporta para várias combinações de parâmetros diferentes, dado um cenário específico.

Com x_3 presente/ausente, definiremos um problema de classificação, caracterizado por: um cenário, uma combinação de parâmetros dentro desse cenário, e uma cardinalidade de conjunto de dados n . Para esse problema de classificação, simularemos um grande número de replicações, sempre encontrando o plano (ou linha) separador por meio de SVM linear.

Dado um classificador 3D, definido pelo plano: $X_3 = d + eX_1 + fX_2$, a correspondente probabilidade de erro esperado é calculada por:

$$\begin{aligned} P(\text{Erro}) &= \frac{1}{2}(1 - \Phi(\text{Distancia}(+))) + \frac{1}{2}(1 - \Phi(\text{Distancia}(-))) = \\ &= \frac{1}{2}\left(1 - \Phi\left(\frac{|d + e + f - 1|}{\sqrt{\Delta}}\right)\right) + \frac{1}{2}\left(1 - \Phi\left(\frac{|-d + e + f - 1|}{\sqrt{\Delta}}\right)\right). \end{aligned} \tag{2.17}$$

Sempre que x_3 estiver ausente, temos um problema de classificação 2D e procedemos de forma análoga como em (2.6).

Dessa forma, seremos capazes de traçar duas curvas (com e sem x_3), expressando $P(\text{Erro})$ como uma função de n . O limiar n^* , presente na Tabela 1 é a abscissa do ponto em que as duas curvas se intersectam.

Para cada um dos quatro cenários, deve-se:

1. Especificar as combinações de parâmetros a serem simuladas.
2. Executar essas simulações e obter as duas curvas de desempenho correspondentes.
3. Interpretar os resultados e conectar essas interpretações com a geometria do problema.

3 SIMULADOR

3.1 SLACGS

Um Simulador para Análise de Erro de Classificadores sob uma carga Gaussiana. desenvolvido a fim de avaliar o Trade Off entre Tamanhos de Amostras e atributos em Problemas de Classificação com amostras gaussianas.

Documentação: <<https://slacgs.netlify.app/>>

- Relatórios com resultados serão armazenados em uma Planilha Google para cada: Cenário de Experimento, Cenário de Experimento Personalizado e outro para as Simulações Personalizadas.
- As Planilhas são armazenadas em uma pasta do Google Drive chamada 'slacgs.demo.<user_email>' pertencente à conta de serviço do Google de slacgs e compartilhada com a conta do Google Drive do usuário.
- Além disso, imagens com visualização de dados serão exportadas para uma pasta local dentro da pasta local do usuário (<user>/slacgs/images/ ou /content/slacgs/images (para G-colab))

3.1.1 Relatórios Exportados

- Relatório de Erro: Contém principalmente resultados focados nas avaliações das Funções de Erro para cada dimensionalidade do modelo.
- Relatório de Comparação: Contém principalmente resultados focados em comparar o desempenho do Modelo para um par de dimensionalidades.
- Relatório do Cenário: Contém resultados de todas as simulações em um Cenário e links para os outros relatórios. (disponível apenas para comparação entre 2D e 3D)

3.1.2 Imagens Exportadas

- Plots de Dados do Cenário .gif: Contém um gif com todos os plots com os pontos de dados ($n = 1024, dims = \{2, 3\}$) gerados para todos os Modelos em um Cenário de Experimento.
- Plot de Dados da Simulação .png: Contém um plot com os pontos de dados ($n = 1024, dims = \{2, 3\}$) gerados para um Modelo em uma Simulação.

- Plot de Erro da Simulação .png: Contém um plot com os valores de erro (Teórico, Empírico com Dados de Treinamento, Empírico com Dados de Teste) gerados para um Modelo em uma Simulação.

3.1.3 Funções de Erro

- Erro Teórico: estimado usando teoria da probabilidade
- Erro Empírico com Dados de Treinamento: estimado usando abordagem empírica apenas com dados de treinamento
- Erro Empírico com Dados de Teste: estimado usando abordagem empírica com dados de treinamento e teste distintos entre si.

3.2 DEMO

1. Baixar e Instalar
2. Configurar/Iniciar Serviço de Relatório
3. Cenários de Experimento
4. Funções de Demonstração:

Executar uma Simulação de Experimento: executar uma simulação para um dos cenários de experimento

Adicionar uma Simulação a um Cenário de Experimento: adicionar resultados de simulação a uma das planilhas de cenário de experimento

Executar um Cenário Personalizado: executar um cenário personalizado e escrever os resultados em uma Planilha Google compartilhada com o usuário

Adicionar uma Simulação a um Cenário Personalizado: adicionar uma simulação a uma planilha de cenário personalizado

Executar uma Simulação Personalizada: executar uma simulação personalizada para qualquer dimensionalidade e cardinalidade

Executar Todas as Simulações de Experimento: executar todas as simulações em todos os cenários de experimento

3.2.1 1. Baixar e Instalar

```
pip install slacgs
```

3.2.2 2. Configurar Serviço de Relatório

```
from slacgs.demo import *

## opt-1: configuração do serviço de relatório com seu próprio arquivo de chave de
→ uma conta de serviço de nuvem do Google
path_to_google_cloud_service_account_api_key = 'path/to/key.json'
set_report_service_conf(path_to_google_cloud_service_account_api_key)

# opt-2 configuração do serviço de relatório para usar o servidor slacgs se você
→ tiver a senha de acesso
set_report_service_conf()
```

3.2.3 3. Cenários de Experimento

```
from slacgs.demo import print_experiment_scenarios

print_experiment_scenarios()
```

3.2.4 4. Funções de Demonstração

```
from slacgs.demo import *

## 1. Executar uma Simulação de Experimento ##
run_experiment_simulation()

## 2. Adicionar uma Simulação a uma Planilha de Cenário de Experimento ##
### Cenário 1
scenario_number = 1
params = [1, 1, 2.1, 0, 0, 0]
add_simulation_to_experiment_scenario_spreadsheet(params, scenario_number)

### Cenário 2
scenario_number = 2
params = [1, 1, 2, -0.15, 0, 0]
add_simulation_to_experiment_scenario_spreadsheet(params, scenario_number)

### Cenário 3
scenario_number = 3
params = [1, 1, 2, 0, 0.15, 0.15]
add_simulation_to_experiment_scenario_spreadsheet(params, scenario_number)
```

```

#### Cenário 4
scenario_number = 4
params = [1, 1, 2, -0.1, 0.15, 0.15]
add_simulation_to_experiment_scenario_spreadsheet(params, scenario_number)

## 3. Executar um Cenário Personalizado ##
scenario_list = [[1,1,3,round(0.1*r,3),round(0.2*r,3),round(0.3*r,3)] for r in
→ range(1, 11)]
run_custom_scenario(scenario_list)

## 4. Adicionar uma Simulação a um Cenário Personalizado ##
### Planilha 1
scenario_number = 1
params = [1, 1, 2.1, 0, 0, 0]
add_simulation_to_custom_scenario_spreadsheet(params, scenario_number)

## 5. Executar uma Simulação Personalizada ##
run_custom_simulation([1, 1, 3, 0.1, 0.2, 0.3])

## 6. Executar Todas as Simulações de Experimento ##
run_all_experiment_simulations()

```

3.3 CLASSE MODEL

A classe `Model` representa um modelo para o SLACGS

3.3.1 Conteúdo

- d_M : Dimensionalidade do modelo.
- $\sigma = \bigcup_{i=1}^d \sigma_i$: Lista de desvios padrão para cada atributo.
- $\rho = \bigcup_{i=1}^d \bigcup_{j=i+1}^d \rho_{ij}$: Lista de correlações entre cada par de atributos.
- Σ : Matriz de correlação $d_M \times d_M$ do modelo
- $\mathbf{N} = \bigcup_{i=1}^k 2^i$, onde k é o comprimento de \mathbf{N} : Lista de cardinalidades n_i do modelo.
- n_{MAX} : Limite superior para uma cardinalidade n_i em \mathbf{N}
- H : Dicionário de classificadores.

3.3.2 Construtor

O construtor da classe `Model` contém os seguintes parâmetros:

- **params**: Lista contendo o vetor de desvio padrão σ e o vetor de correlação ρ , formalmente $\sigma \cup \rho$.
- **max_n**: Cardinalidade máxima para o conjunto \mathbf{N} .
- **N**: Lista de cardinalidades do modelo.
- **dictionary**: Dicionário de classificadores.

3.4 CLASSE SIMULATOR

A classe **Simulator** representa um simulador para análise de erro de classificador em amostras Gaussianas, visando avaliar o trade-off entre amostras e atributos em problemas de classificação em amostras Gaussianas multivariadas geradas.

3.4.1 Conteúdo

Este Simulador para um Modelo SLACGS contém:

- **m**: um objeto Modelo para SLACGS.
- **d**: um vetor de dimensionalidades d_i a serem simuladas, sendo $d_i \leq d_M$.
- **d_{COMP}**: um par de dimensionalidades (d_a, d_b) a serem comparadas.
- **L**: uma lista de funções de erro L_i a serem estimadas para cada cardinalidade $n \in \mathbf{N}$ e dimensionalidade $d \in \mathbf{d}$. $L_i = L_i(\Sigma, n, d) = L_i(\Sigma_{d \times d}, n) = L_i(\hat{h}(x))$, onde $\Sigma_{d \times d}$ é uma secção das primeiras d linhas e colunas de $\Sigma_{d_M \times d_M}$
- **n_{test}**: o número de amostras de teste \mathbf{X}_{test} e \mathbf{y}_{test} a serem geradas para o conjunto de dados D_{test} .

3.4.2 Argumentos para a Simulação

Também contém argumentos para a simulação:

- **n_{AUG}**: limite superior para aplicar aumento de dados a um conjunto com n amostras. Definido por $\text{arg}_{\text{AUGMENTATION_UNTIL_N}}$.
- **cte_{AUG}(n)**: fator de aumento para multiplicar o número de conjuntos de dados gerados para uma cardinalidade de n amostras. É definido por $\sqrt{\frac{n_{\text{AUG}}}{n}}$, se $n < n_{\text{AUG}}$, caso contrário, é igual a 1.
- **r_{STEP}(n)**: número de rodadas (iterações) por *passo* de simulação para uma cardinalidade de n amostras. Definido por $\text{arg}_{\text{STEP_SIZE}} \times \text{cte}_{\text{AUG}}(n)$.

- $r(n)$: número máximo de rodadas (iterações) para uma cardinalidade de n amostras. Definido por $\arg_{MAX_STEPS} \times r_{STEP}(n)$.
- max_{STEPS} : número máximo de *passos* de simulação para cada cardinalidade $n \in \mathbf{N}$. Definido por \arg_{MAX_STEPS} .
- min_{STEPS} : número mínimo de *passos* de simulação para cada cardinalidade $n \in \mathbf{N}$. Definido por \arg_{MIN_STEPS} .
- ϵ : precisão para o critério de parada da simulação. Definida por $\arg_{PRECISION}$.

3.4.3 Critérios de Parada de Convergência de Erro

- $\Delta L_{train}(n, d) < \epsilon$: a diferença entre $E[L_{train}]$ calculada em dois *passos* consecutivos deve ser menor que ϵ para uma cardinalidade de n amostras e uma dimensionalidade de d atributos.
- $\Delta L_{test}(n, d) < \epsilon$: a diferença entre $E[L_{test}]$ calculada em dois *passos* consecutivos deve ser menor que ϵ para uma cardinalidade de n amostras e uma dimensionalidade de d atributos.
- $\Delta L_{theoretical}(n, d) < \epsilon$: a diferença entre $E[L_{theoretical}]$ calculada em dois *passos* consecutivos deve ser menor que ϵ para uma cardinalidade de n amostras e uma dimensionalidade de d atributos.

3.4.4 Testes Após Simulação

Após realizada a simulação para cada cardinalidade $n \in N$, alguns testes são realizados para verificar se a simulação deve continuar para uma cardinalidade adicional $n = max(N) * 2$, caso o limite superior de \mathbf{N} não tenha sido alcançado, ou seja $max(N) < n_{MAX}$:

- Teste \mathbf{N}^* : Encerrar simulação apenas se $L_{type}(\Sigma, n, d_i) > L_{type}(\Sigma, n, d_{i-i})$, para $d_i = max(\mathbf{d})$ e $type \in (\text{theoretical}, \text{test})$, caso contrario indica que ainda pode ser encontrado o ponto $\mathbf{N}^* = L_{type}(\Sigma, n, d_i) \cap L_{type}(\Sigma, n, d_{i-i})$
- Teste de convergência de L : Encerrar simulação apenas se $\forall L \in \mathbf{L}, L(\Sigma, n_i, d) - L(\Sigma, n_{i-1}, d) < \sqrt{\epsilon}$.

3.4.5 Método run()

O método que inicia a simulação contém o algorítmo 1

Algorithm 1 Método run da classe Simulator

```

1: para cada dimensionalidade  $d_i \in \mathbf{d}$  faça
2:   Calcule  $\min(L)_i$  e distância  $d_i$ 
3: fim para
4: enquanto 1 faça
5:   para cada cardinalidade  $n_j \in \mathbf{N}$  faça
6:      $r(n_j) \leftarrow \text{argMAX\_STEPS} \times r_{\text{STEP}}(n_j)$ 
7:     switch( $d_i, L_k$ )  $\leftarrow$  ativado,  $\forall d_i \in \mathbf{d}$  e  $\forall L_k \in \mathbf{L}$   $\triangleright$  Inicie os switches de controle
8:      $\mathbb{E}[L_k] \leftarrow 0, L_k(\Sigma, n_j, d_i) \forall k \forall i$   $\triangleright$  Inicie os estimadores de esperança
9:     para estado aleatório e de 1 até  $r(n_j)$  faça
10:    Gere o conjunto de dados  $D(\Sigma, e) = D_{train} \cup D_{test}$ 
11:    para cada dimensionalidade  $d_i$  em  $\mathbf{dims}$  faça
12:      se switch( $d_i$ ) está ativo então
13:        Treine a SVM com o conjunto de dados  $D_{train}$  em um espaço
14:         $d_i$ -dimensional, obtendo  $\hat{h}_e(x)$ 
15:        para cada  $L_k$  faça
16:          se switch( $d_i, L_k$ ) está ativo então
17:            Compute o erro  $L_k(\hat{h}_e(x))$  e atualize  $\mathbb{E}[L_k(\Sigma, n_j, d_i)]$ ,
18:            fim se
19:        fim para
20:      fim se
21:    fim para
22:     $s \leftarrow e/r_{step}(n_j)$ , onde  $s$  é o passo corrente
23:    se  $s > \min_{\text{STEPS}}$  e  $s \in \mathbb{N}$  então
24:      para cada dimensionalidade  $d_i$  e erro  $L_k$  faça
25:        se  $\mathbb{E}[L_k(\Sigma, n_j, d_i)] - \mathbb{E}_{s-1}[L_k(\Sigma, n_j, d_i)] < \epsilon$  então
26:          Desative switch( $d_i, L_k$ )
27:        fim se
28:         $\mathbb{E}_s[L_k(\Sigma, n_j, d_i)] \leftarrow \mathbb{E}[L_k(\Sigma, n_j, d_i)], \forall k \forall i$ 
29:      fim para
30:      se switch( $d_i$ ) estiver inativo  $\forall i$  então
31:        PARE Simulação para cardinalidade  $n_j$  e CONTINUE para  $n_{j+1}$ 
32:      fim se
33:    fim se
34:    Relate  $\mathbb{E}[L_k(\Sigma, n_j, d_i)], \forall i \forall k$ 
35:  fim para
36:  se  $\max(\mathbf{N}) = n_{MAX}$  então
37:    PARE a Simulação
38:  fim se
39:  se  $(L_k(\Sigma, n_j, d_i) < L_k(\Sigma, n_j, d_{i-1}), \text{ para } d_i = d_M \text{ e } k \in (\text{theoretical}, test))$  e
40:     $(L_k(\Sigma, n_j, d_i) - L_k(\Sigma, n_{j-1}, d_i) < \sqrt{\epsilon}, \text{ para } d_i = d_M \text{ e } \forall k)$  então
41:    PARE a Simulação
42:  senão
43:     $n_{j+1} \leftarrow 2 * \max(\mathbf{N})$ 
44:     $\mathbf{N} \leftarrow \mathbf{N} \cup n_{j+1}$  e CONTINUE a Simulação para nova cardinalidade  $n_{j+1}$ 
45:  fim se
46: fim enquanto
47: Estime  $\min(L)_i \forall d_i$ , caso não tenha sido calculado pelo método analítico
  
```

3.5 FUNÇÕES DO MÓDULO SIMULATOR.PY

3.5.1 Função de erro do Classificador h

A função `h_error_rate` calcula a probabilidade de erro para um classificador linear h , treinado pelo conjunto de dados D , gerado por uma distribuição Gaussiana. Os parâmetros da função são:

- `h_bias` (float): viés do classificador $h(\mathbf{x})$.
- `h_weights` (lista): pesos do classificador $h(\mathbf{x})$.
- `cov` (lista de listas): matriz de covariância das amostras gaussianas usadas para treinar o modelo.

A probabilidade mínima de erro para o classificador linear é dada por:

$$P(\text{Erro}) = \frac{1}{2}(1 - \Phi(\text{Distancia}(+))) + \frac{1}{2}(1 - \Phi(\text{Distancia}(-)))$$

$$P(\text{Erro}) = \frac{1}{2} \left(1 - \Phi \left(\frac{|1 + b - (\mathbf{w}_0 + \dots + \mathbf{w}_{d-1})|}{\sqrt{\Delta}} \right) \right) + \frac{1}{2} \left(1 - \Phi \left(\frac{|1 - b - (\mathbf{w}_0 + \dots + \mathbf{w}_{d-1})|}{\sqrt{\Delta}} \right) \right)$$

Onde:

- d : é a dimensionalidade do classificador h .
- \mathbf{w} : é o vetor de pesos do classificador h .
- b : é o viés do classificador h .
- λ : é o fator de Cholesky inferior-triangular de $\Sigma_{d \times d}$.
- $\delta = (\mathbf{w} | [-1]) \cdot \lambda^T = [\mathbf{w}_0, \dots, \mathbf{w}_{d-1}, -1] \cdot \lambda^T$.
- $\Delta = \sum_{\delta_i \in \delta} \delta_i^2 = \delta_0^2 + \dots + \delta_d^2$.
- $\Phi(\cdot)$ é a função de distribuição cumulativa de uma Gaussiana com média zero e variância um.
- $\text{Distance}(+)$ e $\text{Distance}(-)$: são as distâncias do hiperplano h para a média das classes positiva e negativa, respectivamente.

Os vetores de média para as classes são definidos como:

- $\mu_{(+)} = [\mu_{(+i)}]_{1 \times d}$ para $i \in [1, \dots, d]$ $\mu_{(+)} = [\mu_{+1}, \dots, \mu_{+d}] = [-1, \dots, -1]_{1 \times d}$
- $\mu_{(-)} = [\mu_{(-i)}]_{1 \times d}$ para $i \in [1, \dots, d]$ $\mu_{(-)} = [\mu_{-1}, \dots, \mu_{-d}] = [1, \dots, 1]_{1 \times d}$

A função retorna um float que representa a probabilidade de erro para o classificador $h(\mathbf{x})$ treinado com as amostras gaussianas geradas com a matriz de covariância $\Sigma_{d \times d}$ e médias $\mu_{(+)} = [1, \dots, 1]$ e $\mu_{(-)} = [-1, \dots, -1]$.

Algorithm 2 Cálculo da Probabilidade de Erro para um Classificador Linear

- 1: **função** H_ERROR_RATE((b, \mathbf{w}, Σ))
 - 2: Calcule λ a partir da decomposição de Cholesky de Σ
 - 3: Defina $([w_0, \dots, w_{d-1}, -1] \leftarrow \mathbf{w} \cup [-1])$
 - 4: Calcule Δ :
- $$\delta = (\mathbf{w} | [-1]) \cdot \lambda^T = [w_0, \dots, w_{d-1}, -1] \cdot \lambda^T.$$
- $$\Delta = \sum_{\delta_i \in \delta} \delta_i^2 = \delta_0^2 + \dots + \delta_d^2.$$
- 5: Calcule (Distancia(+)) e (Distancia(-)):
- $$\text{Distancia}(+) = \frac{|\sum [w_0, \dots, w_{d-1}, -1] + b|}{\sqrt{\Delta}}$$
- $$\text{Distancia}(-) = \frac{|\sum [w_0, \dots, w_{d-1}, -1] - b|}{\sqrt{\Delta}}$$
- 6: Calcule $P(\text{Erro}(+))$ e $P(\text{Erro}(-))$ usando a função de distribuição cumulativa da normal padrão (cdf):
- $$P(\text{Erro}(+)) = 1 - \Phi(\text{Distancia}(+))$$
- $$P(\text{Erro}(-)) = 1 - \Phi(\text{Distancia}(-))$$
- 7: Calcule a probabilidade de erro final:
- $$P(\text{Erro}) = \frac{P(\text{Erro}(+)) + P(\text{Erro}(-))}{2}$$
- 8: **retorne** $P(\text{Erro})$
 - 9: **fim função**
-

3.5.2 Função de Erro Empírico

A função `loss_empirical` calcula um erro empírico, denotado por $\hat{L}(h)$, dados:

- Um modelo SVM treinado contendo o melhor classificador empírico $\hat{h}^{(D)}$ no dicionário H para o conjunto de dados D com d atributos e n amostras.
- Um conjunto de teste $D_{\text{test}} = D_{(+)\text{test}} \cup D_{(-)\text{test}}$.

Para cada classificador h em H , podemos calcular sua taxa de erro empírico $\hat{L}(h)$, ou seja, sua proporção de observações classificadas incorretamente no conjunto de dados

D_{test} :

$$\hat{L}(h) = \frac{1}{n} \sum_{i=1}^n 1(y_i \neq h(x_i))$$

Seja $\hat{h}^{(D)}$ o melhor classificador empírico no dicionário H para o conjunto de dados D com d atributos e n amostras. Sob essas condições, podemos definir duas taxas de erro:

- $\hat{L}(\hat{h}^{(D)})$ = a taxa de erro empírico do classificador $\hat{h}^{(D)}$, quando $D_{test} = D_{train}$
- $\hat{L}(\hat{h}^{(D)}, D')$ = a taxa de erro empírico do classificador $\hat{h}^{(D)}$ usando um conjunto de dados de teste D' , quando $D_{test} \neq D_{train}$

Parâmetros:

- `clf` (`sklearn.svm.SVC`): Modelo SVM treinado contendo o melhor classificador empírico $\hat{h}^{(D)}$ no dicionário H para o conjunto de dados D com d atributos e n amostras.
- `dataset_test` (dicionário): um conjunto de teste $D_{test} = D_{(+)\text{test}} \cup D_{(-)\text{test}}$.

Retorna:

- `float`: taxa de erro empírico do classificador $\hat{h}^{(D)}$, $\hat{L}(\hat{h}^{(D)})$ quando $D_{test} = D_{train}$ ou $\hat{L}(\hat{h}^{(D)}, D')$ quando $D_{test} \neq D_{train}$.

Algorithm 3 Cálculo do Erro Empírico

```

1: função LOSS_EMPIRICAL(  $(\hat{h}^{(D_{train})}), (D_{test})$  )
2:   dims_to_remove  $\leftarrow$  dimensionalidades a serem removidas de  $D_{test}$  de acordo com
   o tamanho de w
3:    $\mathbf{X}_{test}, \mathbf{y}_{test} \leftarrow$  remover dimensões de  $D_{test}$  usando dims_to_remove
4:   score  $\leftarrow \frac{1}{n} \sum_{i=1}^n 1(y_i = \hat{h}^{(D_{train})}(x_i))$ 
5:    $\hat{L}(\hat{h}^{(D_{train})}, D_{test}) \leftarrow 1 - score$ 
6:   retorne  $\hat{L}(\hat{h}^{(D_{train})}, D_{test})$ 
7: fim função

```

3.5.3 Função de Erro Teórico

A função `loss_theoretical` utiliza a teoria da probabilidade para calcular o erro teórico, dados:

- Um modelo SVM treinado da biblioteca `sklearn` contendo o melhor classificador linear h no espaço de hipóteses H para o conjunto de dados de treinamento D com d atributos e n amostras.
- A matriz de covariância Σ do conjunto de dados D .

Podemos definir a taxa de erro teórico como:

$$L(\hat{h}^{(D)}) = \text{a taxa de erro teórico do classificador } \hat{h}^{(D)}$$

Parâmetros:

- `h_clf` (`sklearn.svm._classes.SVC`): Classificador SVM treinado contendo o melhor classificador empírico $\hat{h}^{(D)}$ no dicionário H para o conjunto de dados D com d atributos e n amostras.
- `cov` (lista de listas ou `np.ndarray`): Matriz de covariância $\Sigma_{d \times d}$ das amostras gaussianas usadas para treinar o modelo SVM.

Retorna:

- `float`: Erro teórico para o classificador $\hat{h}^{(D)}$.

Algorithm 4 Cálculo do Erro Teórico

```

1: função LOSS_THEORETICAL( $(\hat{h}, \Sigma)$ )
2:   coef_  $\leftarrow \hat{h}.\text{coef\_}[0]$ 
3:    $b \leftarrow (\hat{h}.\text{intercept\_}[0]/\text{coef\_}[\text{len}(\text{coef\_}) - 1])$ 
4:    $w \leftarrow [-\text{coef\_}[i]/\text{coef\_}[\text{len}(\text{coef\_}) - 1] \text{ for } i \text{ in range}(\text{len}(\text{coef\_}) - 1)]$ 
5:   retorne H_ERROR_RATE( $b, w, \Sigma$ )
6: fim função

```

3.6 CLASSE REPORT

A classe `Report` é responsável por gerar um relatório das simulações executadas.

3.6.1 Construtor

O construtor da classe `Report` aceita um objeto `Simulator` como argumento e inicializa os seguintes atributos:

- `sim` (`Simulator`): Objeto `Simulator`.
- `iter_N` (`dict`): Número de iterações para cada dimensionalidade e tipo de erro.
- `max_iter_N` (`list`): Número máximo de iterações para cada dimensionalidade.
- `loss_N` (`dict`): Erro para cada dimensionalidade e tipo de erro.
- `loss_bayes` (`dict`): Erro Bayes para cada dimensionalidade.
- `d` (`dict`): distância da origem ao ponto de interseção entre o elipsoide normalizado e a diagonal principal para cada dimensionalidade.

- `duration` (float): Duração da simulação.
- `time_spent` (dict): Tempo gasto para cada dimensionalidade e tipo de erro.
- `sim_tag` (dict): Atributos do objeto Simulator.
- `model_tag` (dict): Atributos do objeto Model.
- `loss_plot` (matplotlib.figure.Figure): Plotagem do erro.

4 RESULTADOS

4.1 QUANDO X_3 CONTRIBUI PARA AUMENTAR O PODER DE DISCRIMINAÇÃO DE (X_1, X_2) ?

Dado um problema de classificação com as características que estamos analisando, ou seja, duas distribuições tri-normais centradas respectivamente em $(1, 1, 1)$ e $(-1, -1, -1)$, ambas com matrizes de covariâncias iguais a

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 \end{bmatrix} \quad (4.1)$$

Sabemos que, dependendo dos valores dos 6 parâmetros (variâncias e correlações) que definem essa matriz Σ , existe um limiar n^* a partir do qual passa a ser vantajosa a presença da feature X_3 , dado que as duas outras features X_1 e X_2 já estão presentes.

Pergunta: Dada uma matriz Σ , caracterizada por uma particular combinação dos 6 parâmetros $\sigma_1, \sigma_2, \sigma_3, \rho_{12}, \rho_{13}, \rho_{23}$ que a definem, como avaliar o potencial de contribuição adicional de X_3 para a discriminação entre os dois grupos, uma vez que X_1 e X_2 já estejam presentes?

Para analisar essa questão foram criados 4 cenários, sendo que em cada um deles apenas 3 dos 6 parâmetros podem ser escolhidos livremente. Em cada um desses 4 cenários, dada a matriz Σ , pode ser calculado o quociente $\frac{Risco2}{Risco3}$, onde $Risco3$ e $Risco2$ são os riscos de Bayes (erro de Bayes) relativos, respectivamente, à situação em que todas as 3 features estão presentes e à situação em que somente X_1 e X_2 estão presentes.

Supostamente, quanto maior for o quociente $\frac{Risco2}{Risco3}$, menor deverá ser n^* .

Por outro lado, sabemos que as superfícies de nível correspondentes a uma determinada distribuição tri-normal são elipsóides centrados no respectivo centróide (vetor de médias). Ora, se os dois centróides são $(1, 1, 1)$ e $(-1, -1, -1)$, quanto mais a direção do eixo principal desses elipsóides se aproximar da direção da reta que passa por esses dois centróides, maior será a interseção entre as nuvens de pontos relativas aos dois grupos, ou seja, maior será a probabilidade esperada de erro do classificador, se todas as 3 features estiverem presentes. Como podemos observar nas figuras 1 e 2.

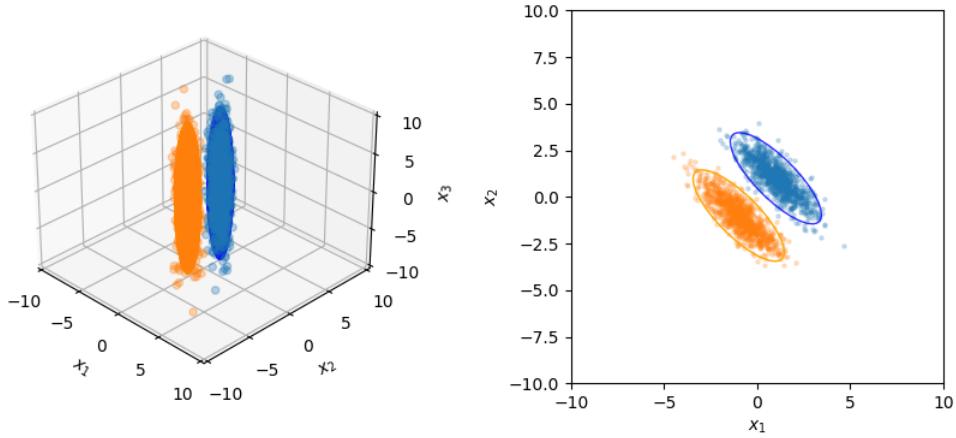


Figura 1 – $\sigma_3 = [1, 1, 4]$, $\rho_3 = [-0.8, 0, 0]$ e $\sigma_2 = [1, 1]$, $\rho_2 = [-0.8]$; $n = 1024$
 $, L_3(\hat{h}^{(D)}) = 0,000616, L_3(h^*) = 0,000756, L_3(\hat{h}^{(D)}) = 0,000929, \hat{L}_3(\hat{h}^{(D)}, D') = 0,000884$
 $\hat{L}_2(\hat{h}^{(D)}) = 0,000664, L_2(h^*) = 0,000782, L_2(\hat{h}^{(D)}) = 0,000895, \hat{L}_2(\hat{h}^{(D)}, D') = 0,000852$
 $\frac{Risco_2}{Risco_3} = 1,034513271, \frac{d_2}{d_3} = 1,003172086$
 $n^* = 3155$, para $\hat{L}_{2\cap 3}(h^{(D)})$; $n^* = 3477$, para $\hat{L}_{2\cap 3}(h^{(D)}, D')$

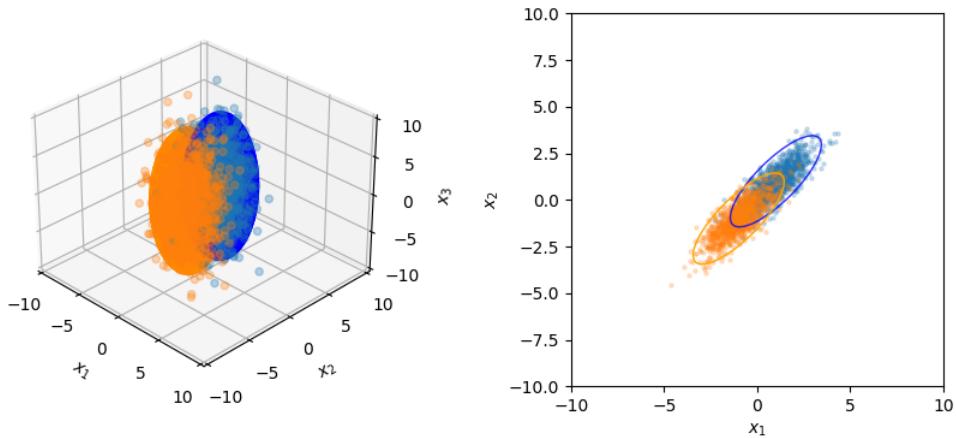


Figura 2 – $\sigma_3 = [1, 1, 4]$, $\rho_3 = [0.8, 0, 0]$ e $\sigma_2 = [1, 1]$, $\rho_2 = [0.8]$; $n = 1024$
 $, \hat{L}_3(\hat{h}^{(D)}) = 0,138705, L_3(h^*) = 0,139330, L_3(\hat{h}^{(D)}) = 0,140103, \hat{L}_3(\hat{h}^{(D)}, D') = 0,140113$
 $\hat{L}_2(\hat{h}^{(D)}) = 0,145647, L_2(h^*) = 0,145920, L_2(\hat{h}^{(D)}) = 0,146405, \hat{L}_2(\hat{h}^{(D)}, D') = 0,146841$
 $\frac{Risco_2}{Risco_3} = 1,047297879, \frac{d_2}{d_3} = 1,02773264$
 $n^* = 51$, para $\hat{L}_{2\cap 3}(h^{(D)})$; $n^* = 51$, para $\hat{L}_{2\cap 3}(h^{(D)}, D')$

4.1.1 Um modelo para entender o comportamento do risco de Bayes em duas dimensões

Suponha que somente X_1 e X_2 estão presentes. Consideremos uma elipse no \mathbb{R}^2 (digamos, centrada na origem $(0, 0)$) cuja equação seja

$$x^T \Sigma^{-1} x = 1.$$

A cada classe temos uma elipse, e para cada elipse temos um centróide. A reta que une os dois centróides $(1, 1)$ e $(-1, -1)$ é o conjunto de todos os vetores do \mathbb{R}^2 do tipo

$$(c, c),$$

cujas duas coordenadas são iguais. Essa reta corta a referida elipse em algum ponto que tem as duas coordenadas iguais e positivas,

$$(c_2, c_2) \cdot \Sigma^{-1} \begin{pmatrix} c_2 \\ c_2 \end{pmatrix} = 1.$$

Seja d_2 a distância desse ponto à origem,

$$d_2 = \sqrt{2}c_2.$$

Quanto maior for essa distância, menor será o poder de discriminação do classificador, com X_1 e X_2 presentes.

4.1.2 Estendendo para o caso de três dimensões

Estendendo a discussão, consideremos um elipsóide em \mathbb{R}^3 (digamos, centrado na origem $(0, 0, 0)$) cuja equação seja

$$x^T \Sigma^{-1} x = 1.$$

lembmando que agora $x \in \mathbb{R}^3$.

Recordando, cada classes de pontos (por exemplo, laranja e azul) está associada a um elipsóide, que por sua vez contém um centróide. A reta que une os dois centróides é o conjunto de todos os vetores do \mathbb{R}^3 do tipo (c, c, c) , cujas três coordenadas são iguais. Essa reta corta o referido elipsóide em algum ponto que tem todas as três coordenadas iguais e positivas.

$$(c_3, c_3) \cdot \Sigma^{-1} \begin{pmatrix} c_3 \\ c_3 \\ c_3 \end{pmatrix} = 1.$$

Seja d_3 a distância desse ponto à origem,

$$d_3 = \sqrt{3}c_3.$$

Quanto maior for essa distância, menor será o poder de discriminação do classificador, com X_1 , X_2 e X_3 presentes.

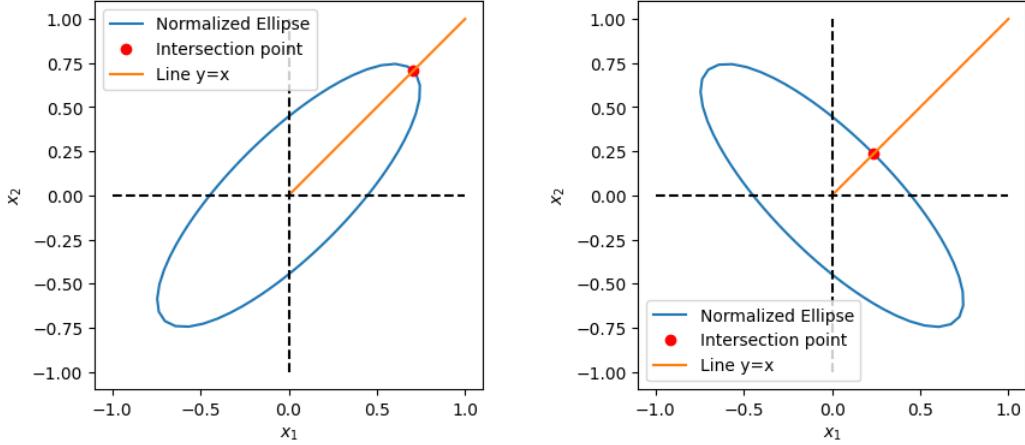


Figura 3 – $\sigma_2 = [1, 1]$, $\rho_2 = [0.8]$ e $\sigma_2 = [1, 1]$, $\rho_2 = [-0.8]$

4.1.3 Comparando os cenários com dois e três atributos

Como usar d_3 e d_2 para criar, a partir da matriz Σ , um indicador do potencial adicional de discriminação de X_3 dado que X_1 e X_2 já estão presentes? O indicador que estamos propondo é d_2/d_3 ,

$$U_3(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \frac{d_2}{d_3} = \sqrt{\frac{2}{3}} \frac{c_2}{c_3} \quad (4.2)$$

onde U_3 é uma métrica de utilidade do atributo 3.

Seja R_2 o risco de Bayes com 2 atributos, e R_3 o correspondente risco com 3 atributos. Empiricamente verificamos que a razão entre os riscos é uma função crescente e suave da função de utilidade acima, ou seja,

$$\frac{R_2}{R_3} = \varphi(U_3) \quad (4.3)$$

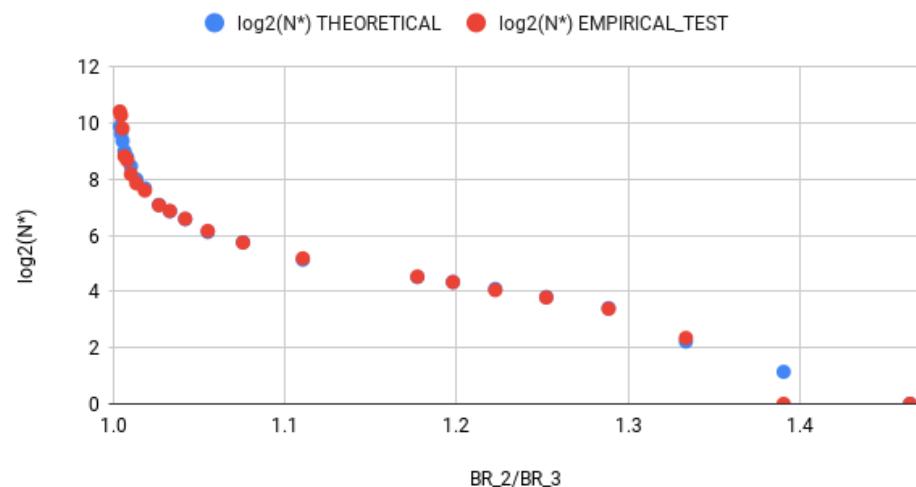
onde a função $\varphi(\cdot)$ é uma função crescente e suave, a ser experimentalmente obtida nas próximas seções.

Tabela 2 – Cenário 1 - indicadores

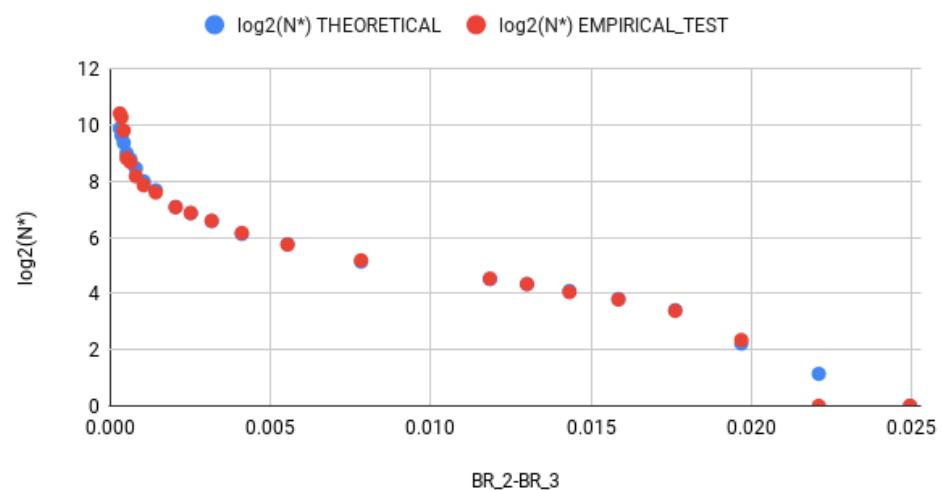
σ_1	σ_2	σ_3	ρ_{12}	ρ_{13}	ρ_{23}	$BR_2 - BR_3$	$\frac{BR_2}{BR_3}$	$\frac{d_2}{d_3}$	$d_2 - d_3$	$\log_2(n*)$	theoretical	empirical, D'
1	1	1.3	0	0	0	0.0249	1.4643	0.1215	1.1383	0.0000	0.0000	
1	1	1.4	0	0	0	0.0221	1.3907	0.1073	1.1202	1.1406	0.0000	
1	1	1.5	0	0	0	0.0197	1.3337	0.0955	1.1055	2.2187	2.3477	
1	1	1.6	0	0	0	0.0176	1.2887	0.0853	1.0933	3.4104	3.3854	
1	1	1.7	0	0	0	0.0158	1.2523	0.0767	1.0830	3.8092	3.7875	
1	1	1.8	0	0	0	0.0143	1.2226	0.0692	1.0744	4.0930	4.0501	
1	1	1.9	0	0	0	0.0130	1.1980	0.0628	1.0670	4.3382	4.3366	
1	1	2	0	0	0	0.0118	1.1773	0.0572	1.0607	4.5153	4.5309	
1	1	3	0	0	0	0.0055	1.0757	0.0267	1.0274	5.7580	5.7437	
1	1	4	0	0	0	0.0032	1.0420	0.0153	1.0155	6.5794	6.5922	
1	1	5	0	0	0	0.0020	1.0267	0.0098	1.0099	7.0890	7.0739	
1	1	6	0	0	0	0.0014	1.0185	0.0069	1.0069	7.6776	7.5998	
1	1	7	0	0	0	0.0011	1.0135	0.0050	1.0050	7.9957	7.8544	
1	1	8	0	0	0	0.0008	1.0104	0.0039	1.0039	8.4570	8.1806	
1	1	9	0	0	0	0.0006	1.0082	0.0030	1.0030	8.7796	8.6800	
1	1	10	0	0	0	0.0005	1.0066	0.0025	1.0025	8.9991	8.8188	

Tabela 3 – Cenário 1 - $\hat{L}(\hat{h}^{(D)})$: Erro Empírico de Treino

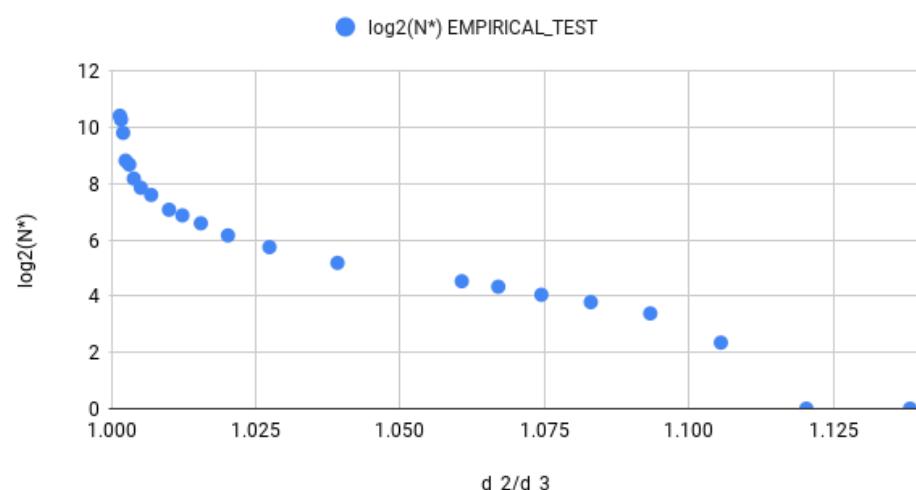
σ_3	$n = 2$	$n = 4$	$n = 8$	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$	$n = 512$	$n = 1024$	$\min(L)$
2 atributos	0	0.0295	0.0478	0.0593	0.0669	0.0723	0.0751	0.0769	0.0774	0.0780	0.0786
1.3	0	0.0073	0.0167	0.0278	0.0374	0.0447	0.0486	0.0511	0.0524	0.0530	0.0537
1.4	0	0.0078	0.0180	0.0294	0.0402	0.0476	0.0514	0.0540	0.0551	0.0558	0.0566
1.5	0	0.0082	0.0190	0.0312	0.0424	0.0498	0.0538	0.0563	0.0574	0.0583	0.0590
1.6	0	0.0086	0.0203	0.0326	0.0438	0.0516	0.0557	0.0583	0.0594	0.0604	0.0610
1.7	0	0.0089	0.0207	0.0336	0.0450	0.0533	0.0572	0.0601	0.0612	0.0621	0.0628
1.8	0	0.0090	0.0213	0.0348	0.0461	0.0548	0.0589	0.0615	0.0627	0.0636	0.0643
1.9	0	0.0090	0.0217	0.0357	0.0473	0.0559	0.0602	0.0629	0.0640	0.0649	0.0657
2	0	0.0094	0.0222	0.0366	0.0482	0.0568	0.0613	0.0640	0.0652	0.0661	0.0668
3	0	0.0095	0.0255	0.0417	0.0542	0.0632	0.0676	0.0703	0.0713	0.0725	0.0731
4	0	0.0097	0.0272	0.0440	0.0564	0.0658	0.0699	0.0725	0.0737	0.0747	0.0755
5	0	0.0095	0.0282	0.0453	0.0574	0.0668	0.0711	0.0738	0.0748	0.0759	0.0766
6	0	0.0093	0.0284	0.0460	0.0581	0.0675	0.0718	0.0745	0.0755	0.0765	0.0772
7	0	0.0094	0.0285	0.0463	0.0584	0.0679	0.0721	0.0749	0.0759	0.0769	0.0776
8	0	0.0095	0.0286	0.0466	0.0592	0.0680	0.0722	0.0752	0.0762	0.0771	0.0778
9	0	0.0095	0.0285	0.0468	0.0589	0.0682	0.0724	0.0754	0.0763	0.0773	0.0780
10	0	0.0095	0.0290	0.0470	0.0590	0.0683	0.0726	0.0755	0.0765	0.0774	0.0781

$\log_2(N^*)$ vs BR_2/BR_3

(a)

 $\log_2(N^*)$ vs BR_2-BR_3

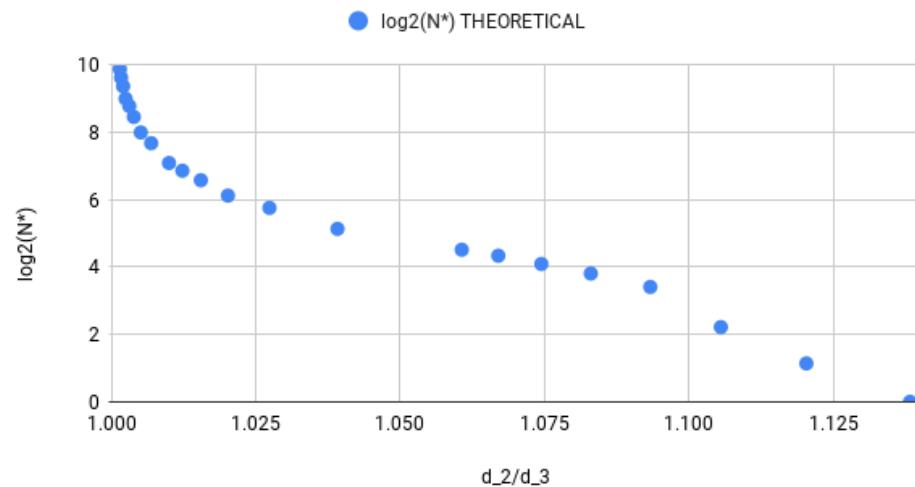
(b)

 $\log_2(N^*)$ vs d_2/d_3 | Empirical

(c)

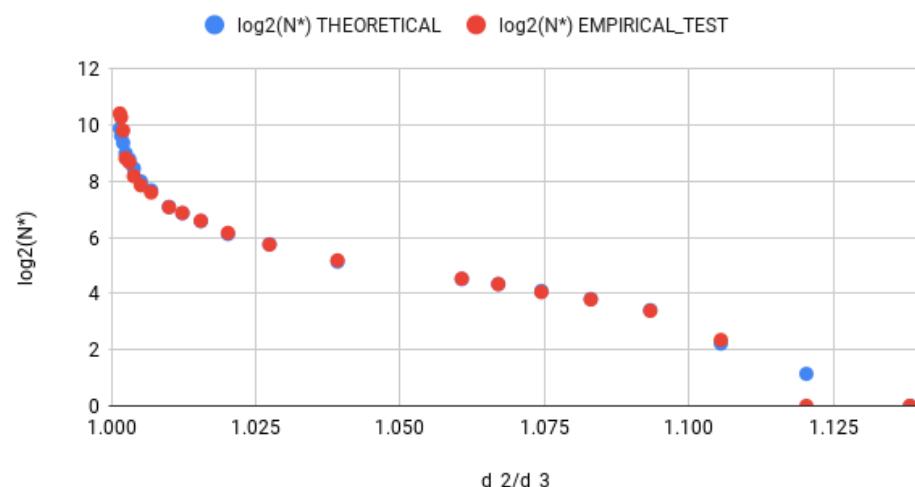
Figura 4 – Cenário 1 - indicadores 1/4

$\log_2(N^*)$ vs d_2/d_3 | Theoretical



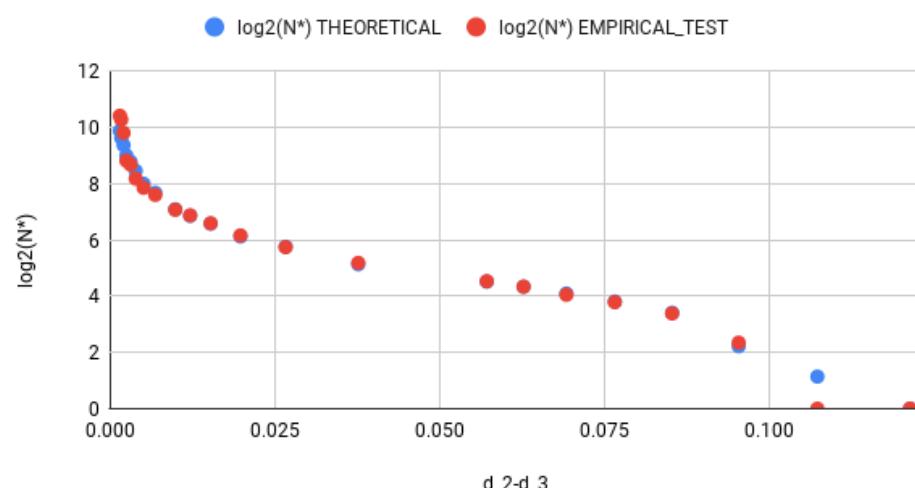
(a)

$\log_2(N^*)$ vs d_2/d_3



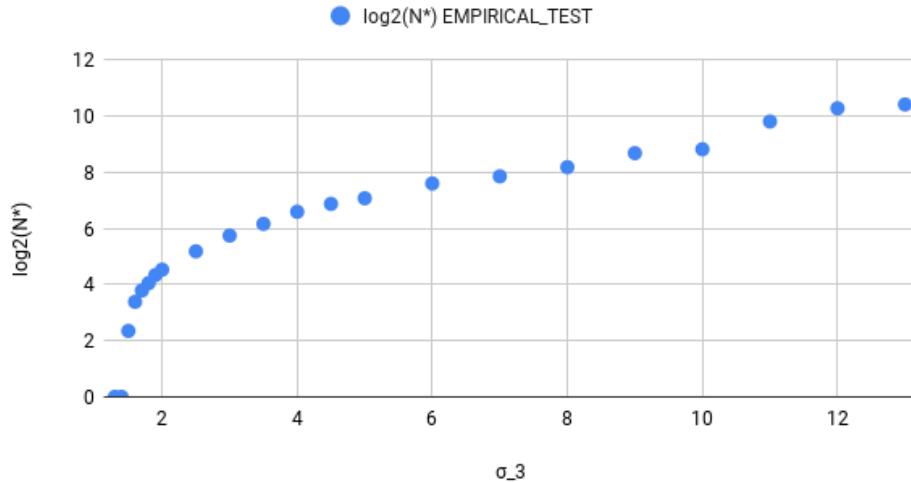
(b)

$\log_2(N^*)$ vs d_2-d_3

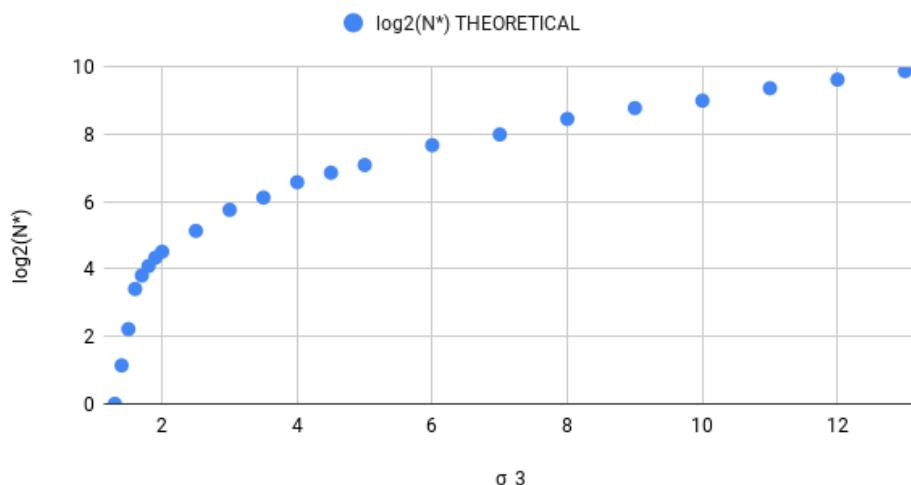


(c)

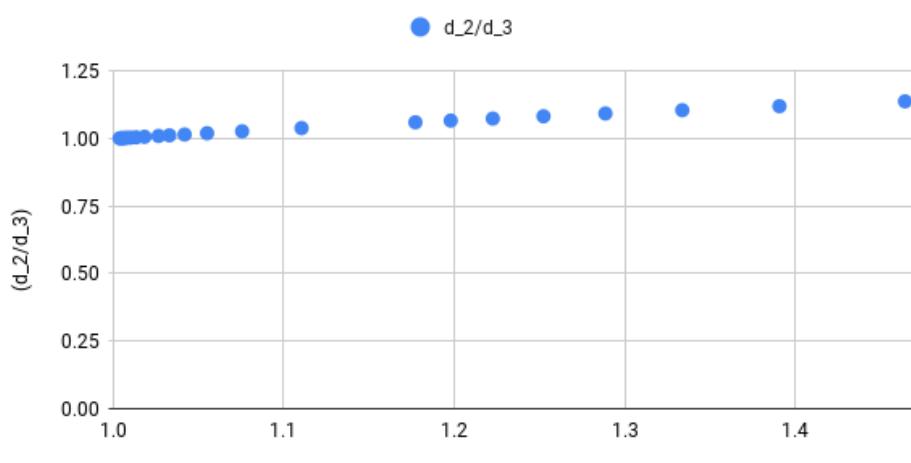
Figura 5 – Cenário 1 - indicadores 2/4

$\log_2(N^*)$ vs σ_3 | Empirical

(a)

 $\log_2(N^*)$ vs σ_3 | Theoretical

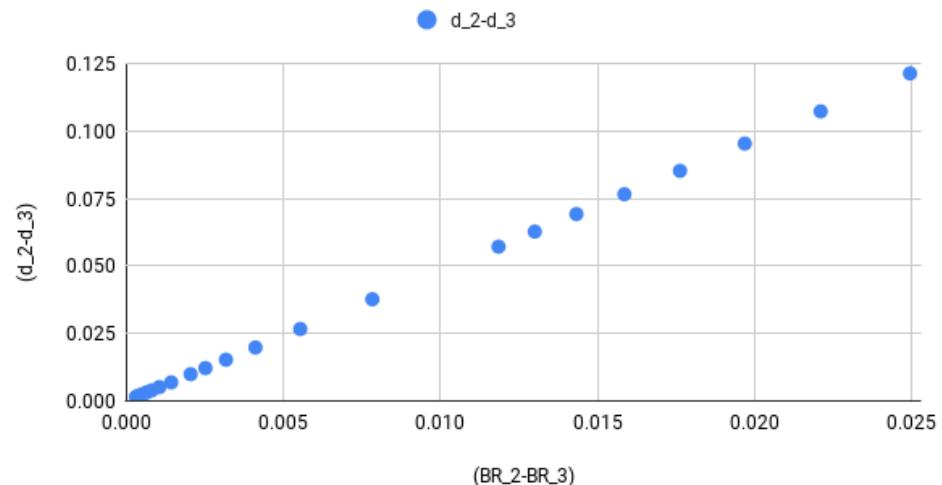
(b)

 (d_2/d_3) vs (BR_2/BR_3) 

(c)

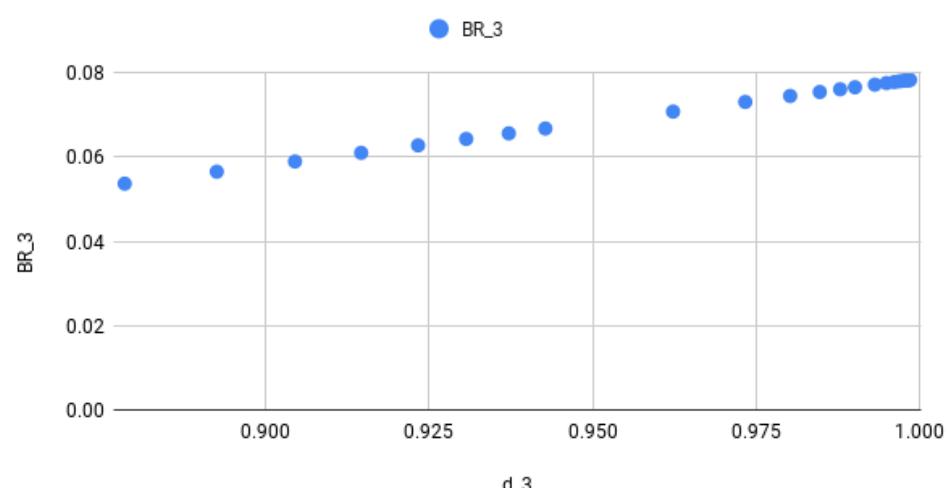
Figura 6 – Cenário 1 - indicadores 3/4

(d_2-d_3) vs (BR_2-BR_3)



(a)

BR_3 vs d_3



(b)

Figura 7 – Cenário 1 - indicadores 4/4

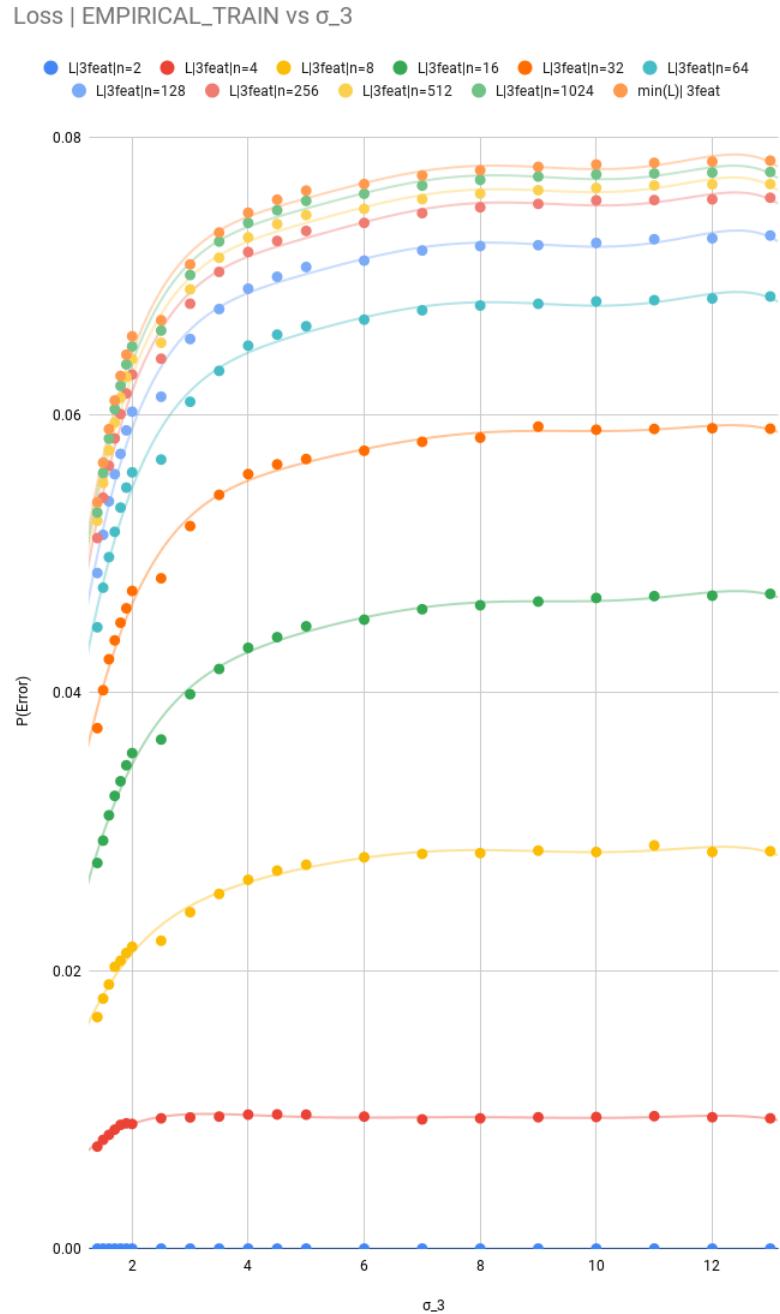
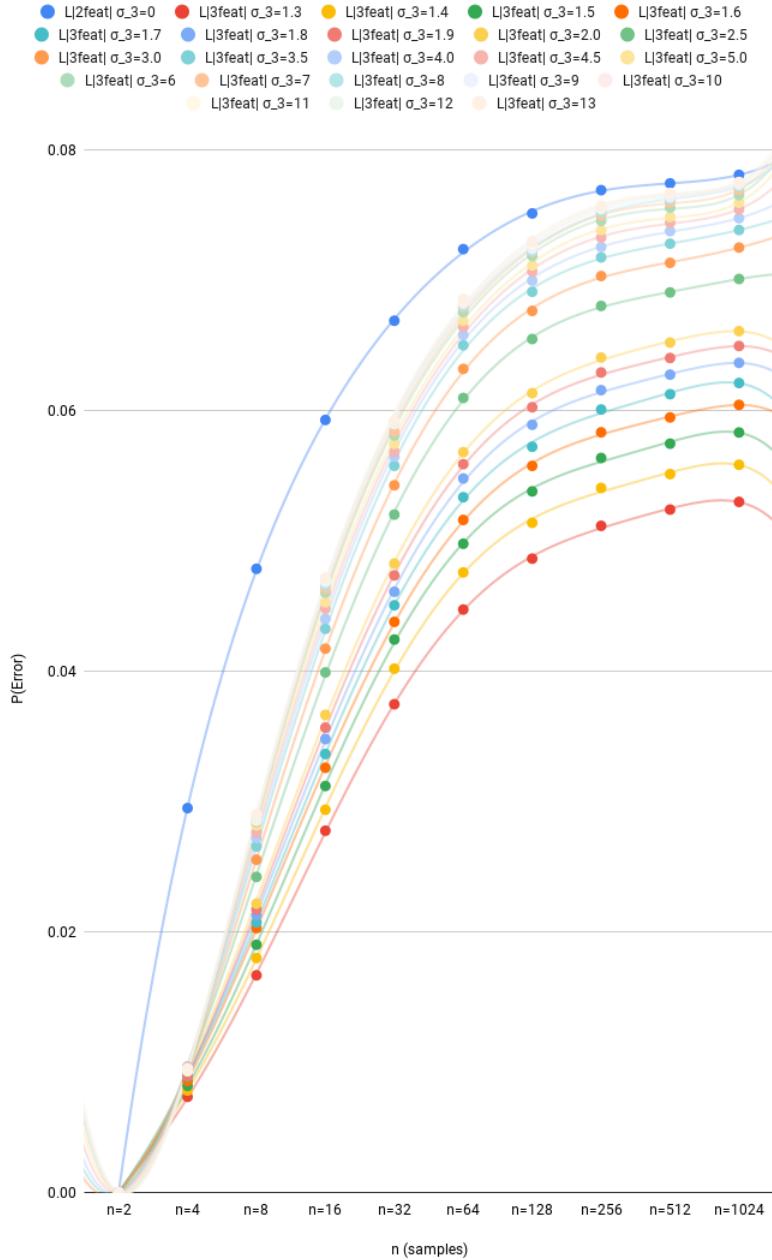


Figura 8 – Cenário 1 - $\hat{L}(\hat{h}^{(D)})$ vs σ_3

Loss | EMPIRICAL_TRAIN vs n

Figura 9 – Cenário 1 - $\hat{L}(\hat{h}^{(D)})$ vs n

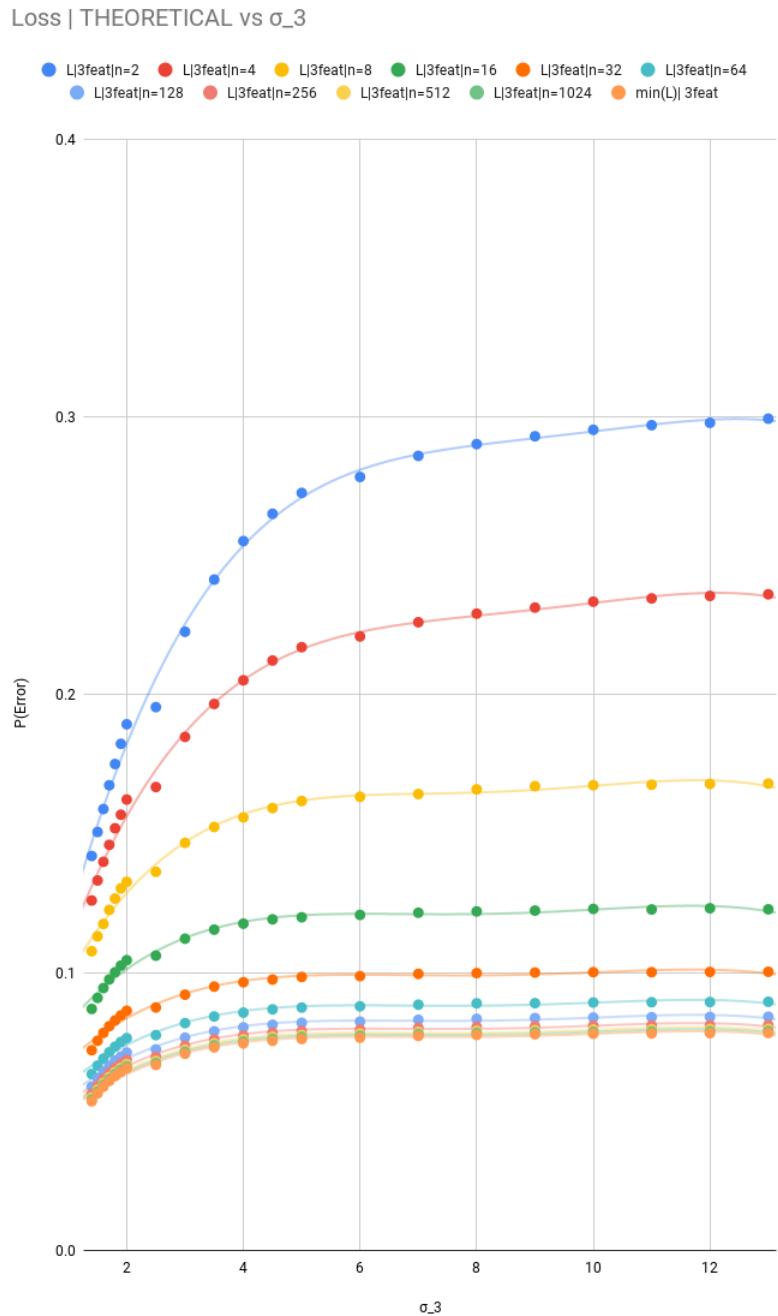
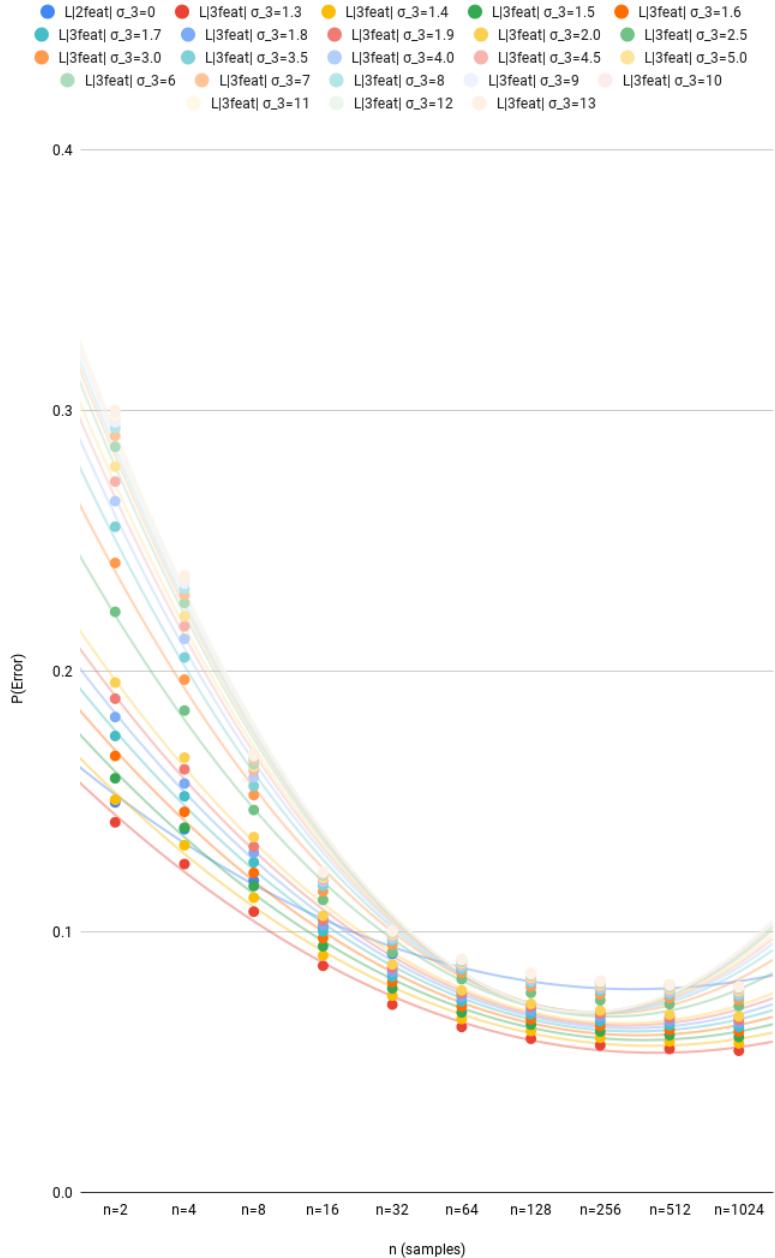


Figura 10 – Cenário 1 - $L(\hat{h}^{(D)})$ vs σ_3

Loss | THEORETICAL vs n

Figura 11 – Cenário 1 - $L(\hat{h}^{(D)})$ vs n

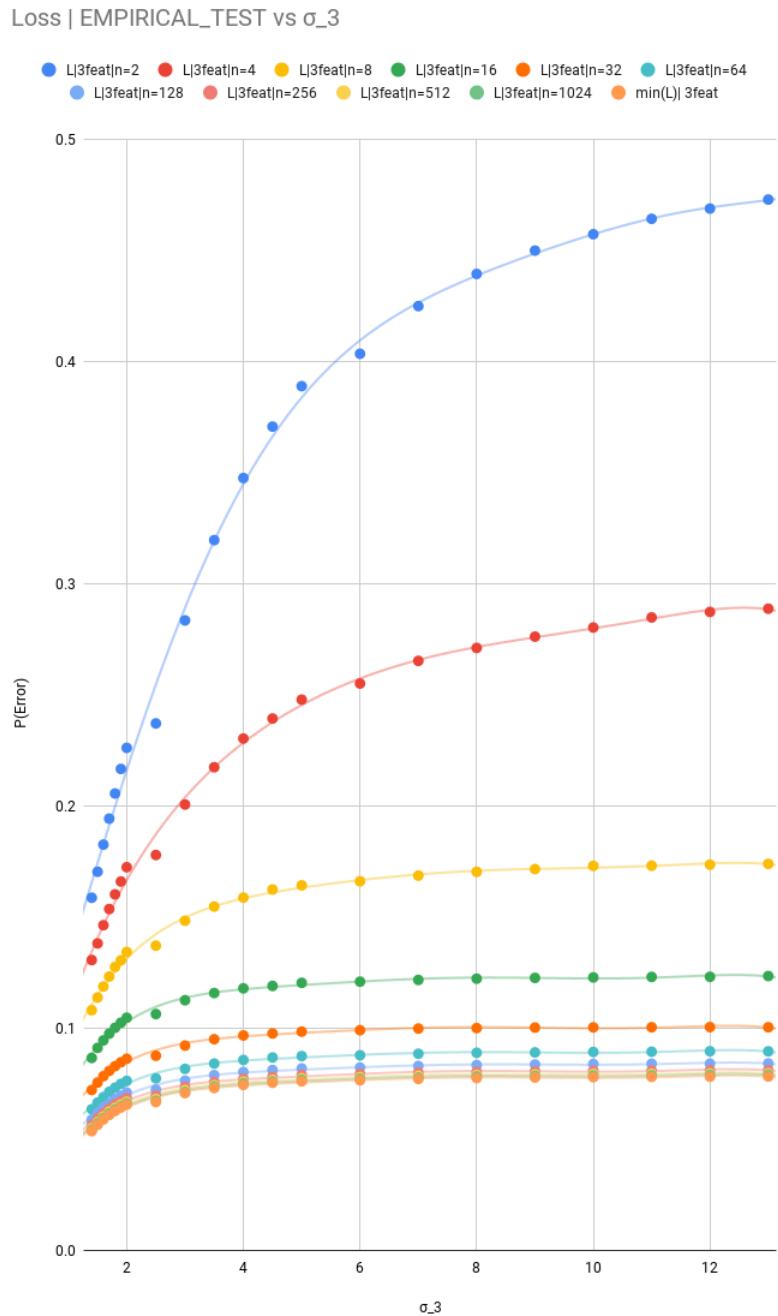
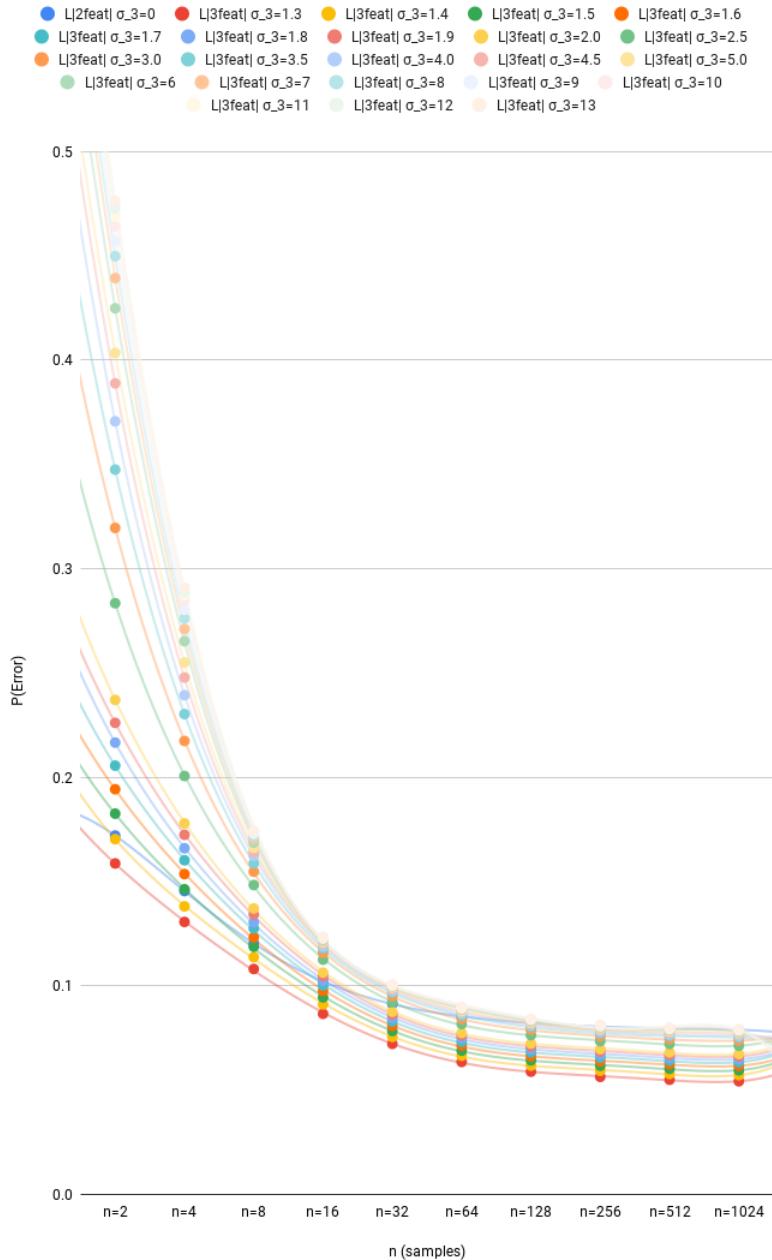
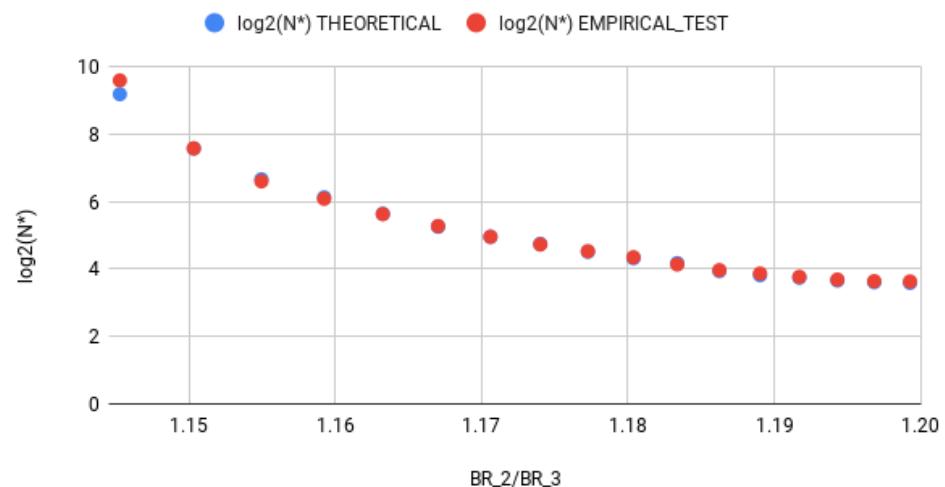


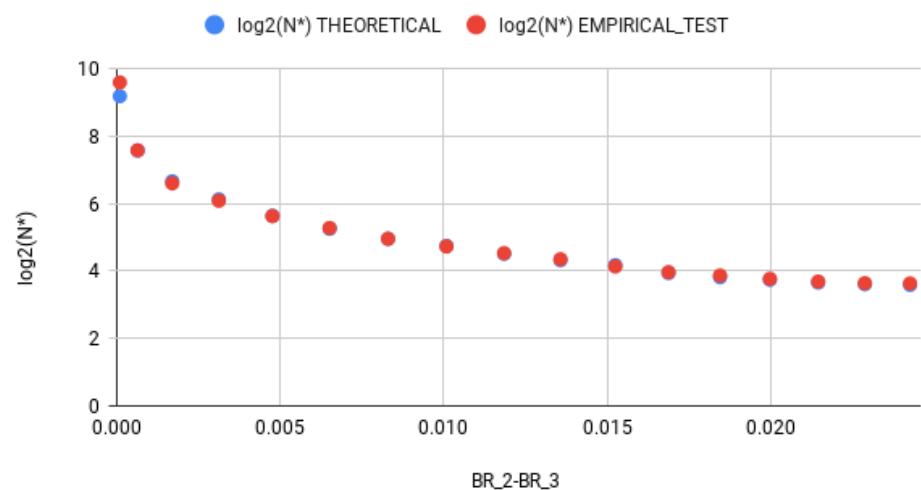
Figura 12 – Cenário 1 - $\hat{L}(\hat{h}^{(D)}, D')$ vs σ_3

Loss | EMPIRICAL_TEST vs n

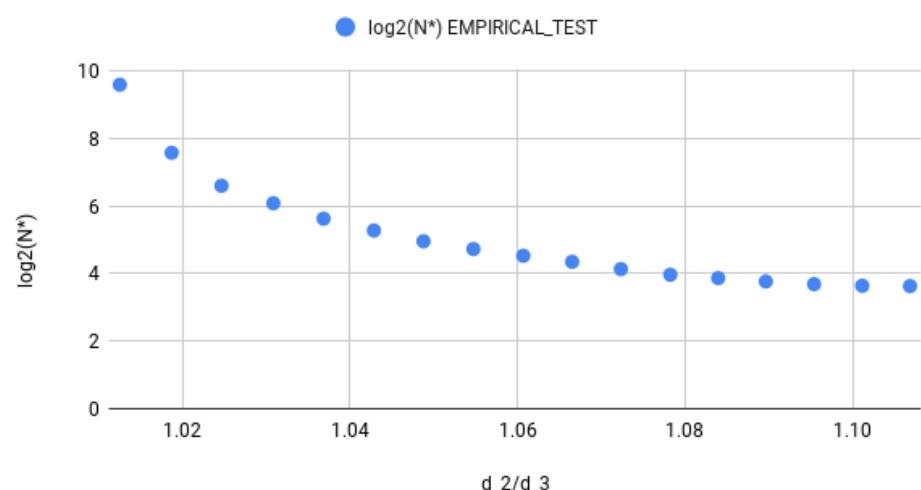
Figura 13 – Cenário 1 - $\hat{L}(\hat{h}^{(D)}, D')$ vs n

$\log_2(N^*)$ vs BR_2/BR_3

(a)

 $\log_2(N^*)$ vs BR_2-BR_3

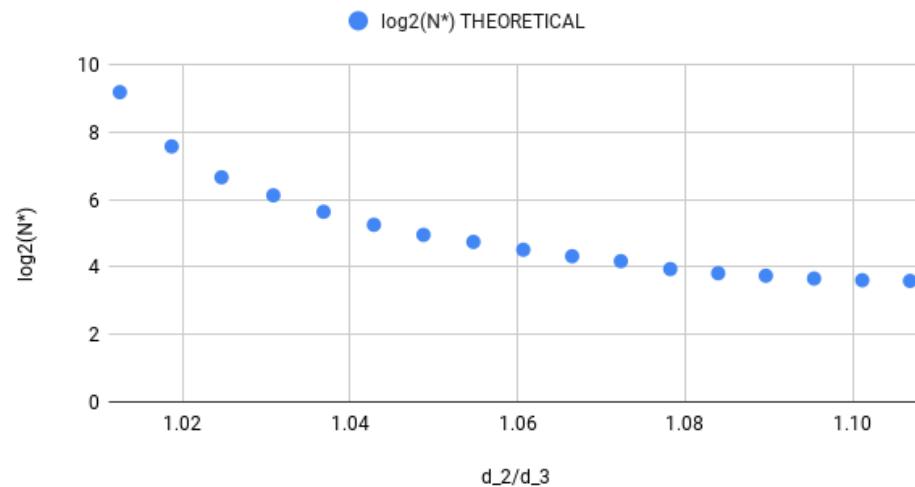
(b)

 $\log_2(N^*)$ vs d_2/d_3 | Empirical

(c)

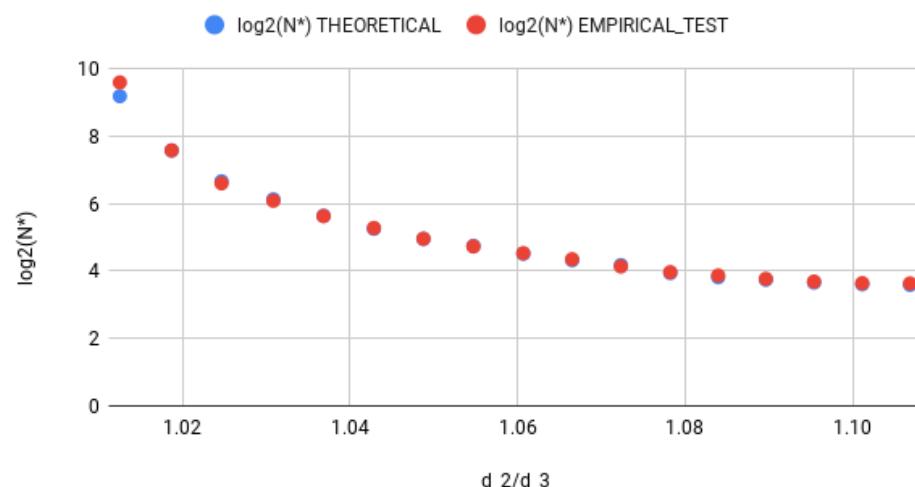
Figura 14 – Cenário 2 - Indicadores 1/4

$\log_2(N^*)$ vs d_2/d_3 | Theoretical



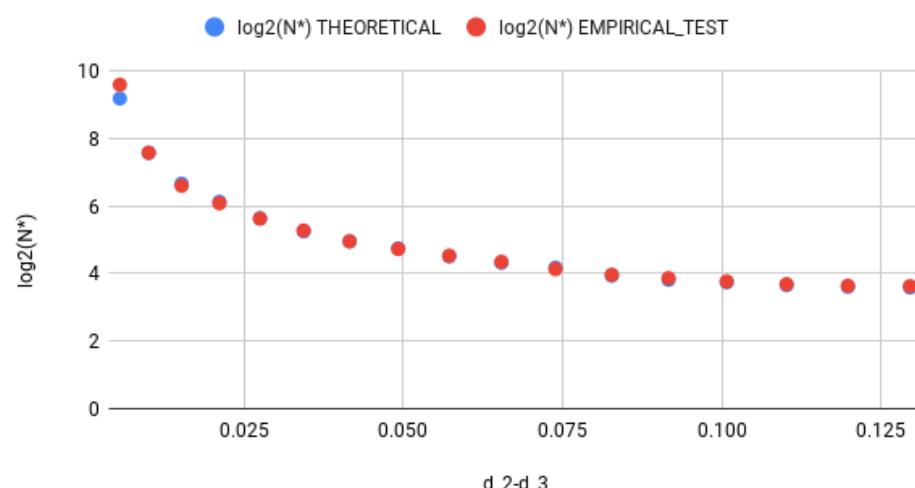
(a)

$\log_2(N^*)$ vs d_2/d_3



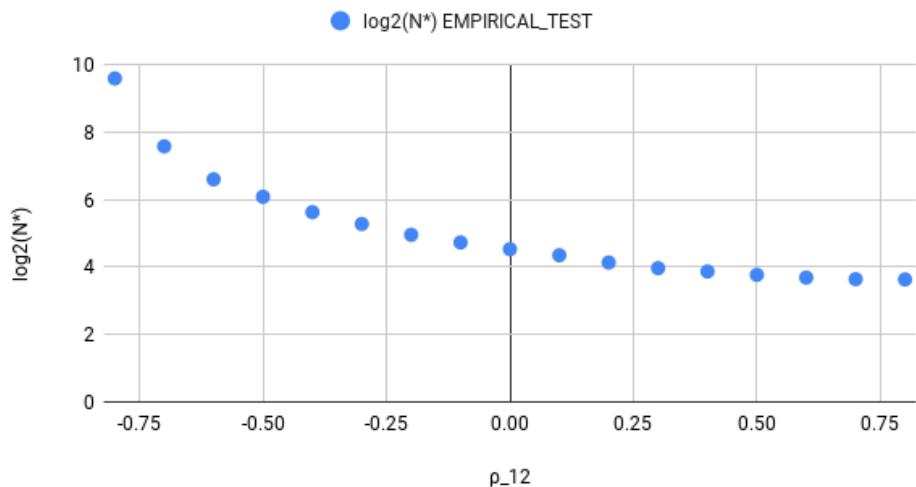
(b)

$\log_2(N^*)$ vs d_2-d_3

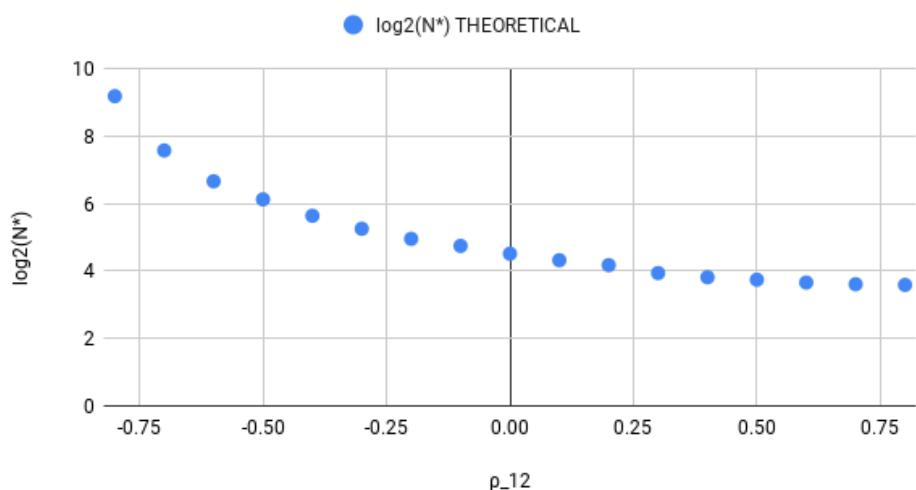


(c)

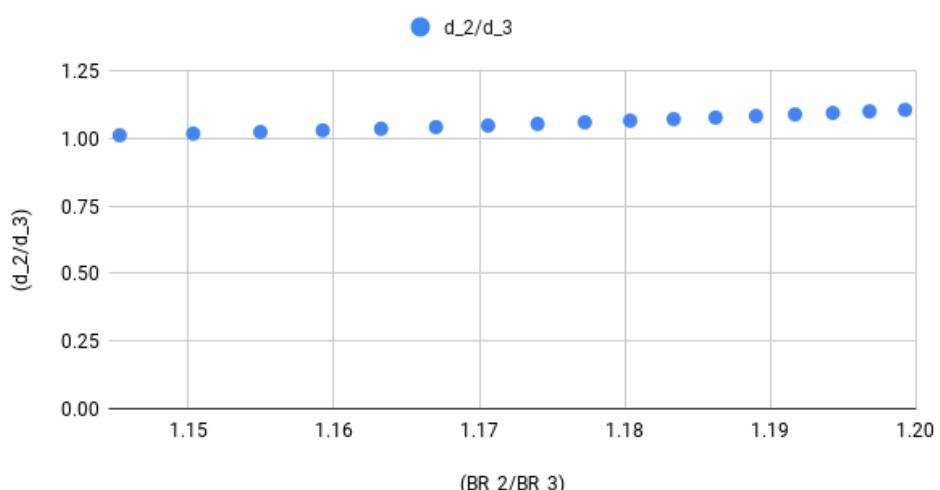
Figura 15 – Cenário 2 - Indicadores 2/4

$\log_2(N^*)$ vs ρ_{12} | Empirical

(a)

 $\log_2(N^*)$ vs ρ_{12} | Theoretical

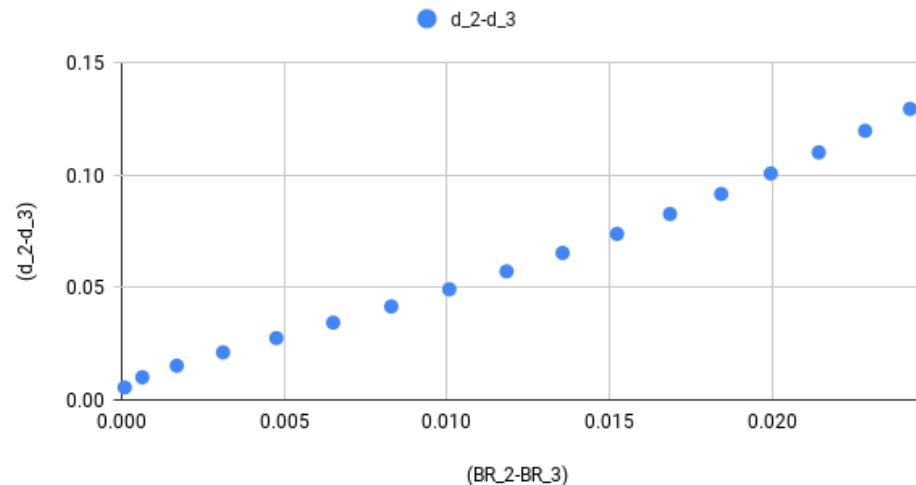
(b)

 (d_2/d_3) vs (BR_2/BR_3) 

(c)

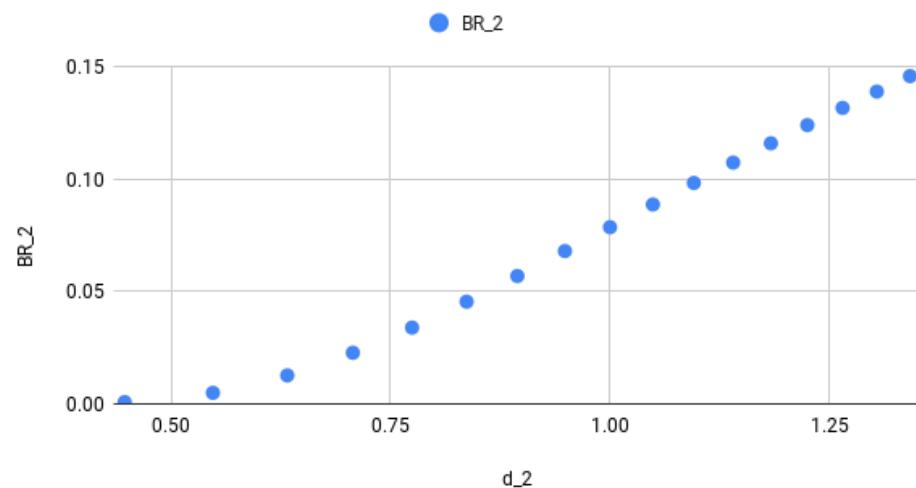
Figura 16 – Cenário 2 - Indicadores 3/4

(d_2-d_3) vs (BR_2-BR_3)



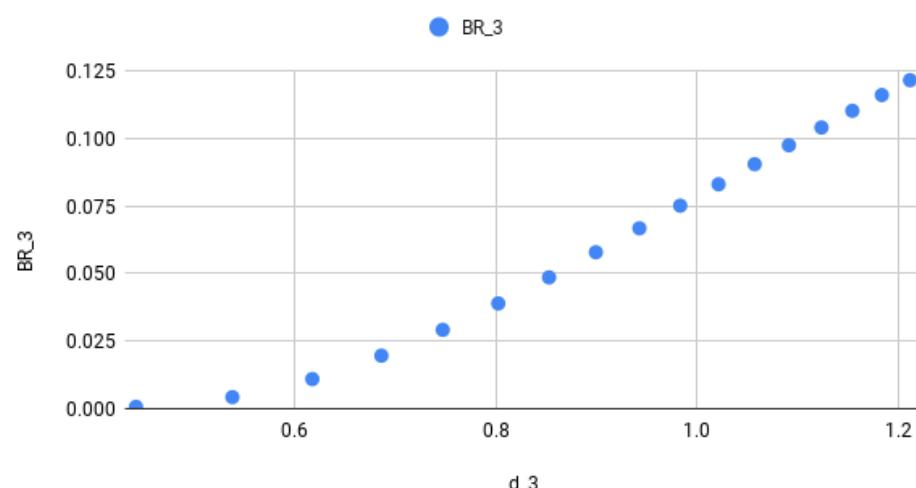
(a) 9

BR_2 vs d_2



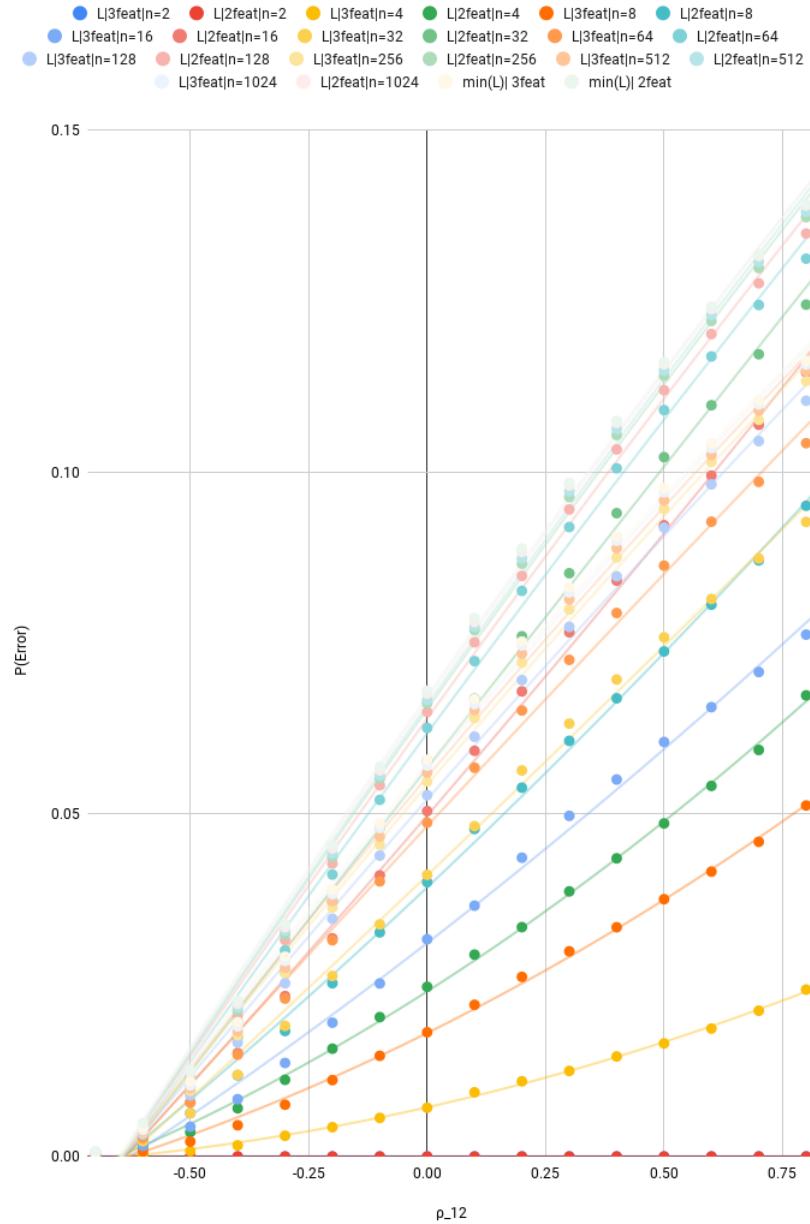
(b) 10

BR_3 vs d_3

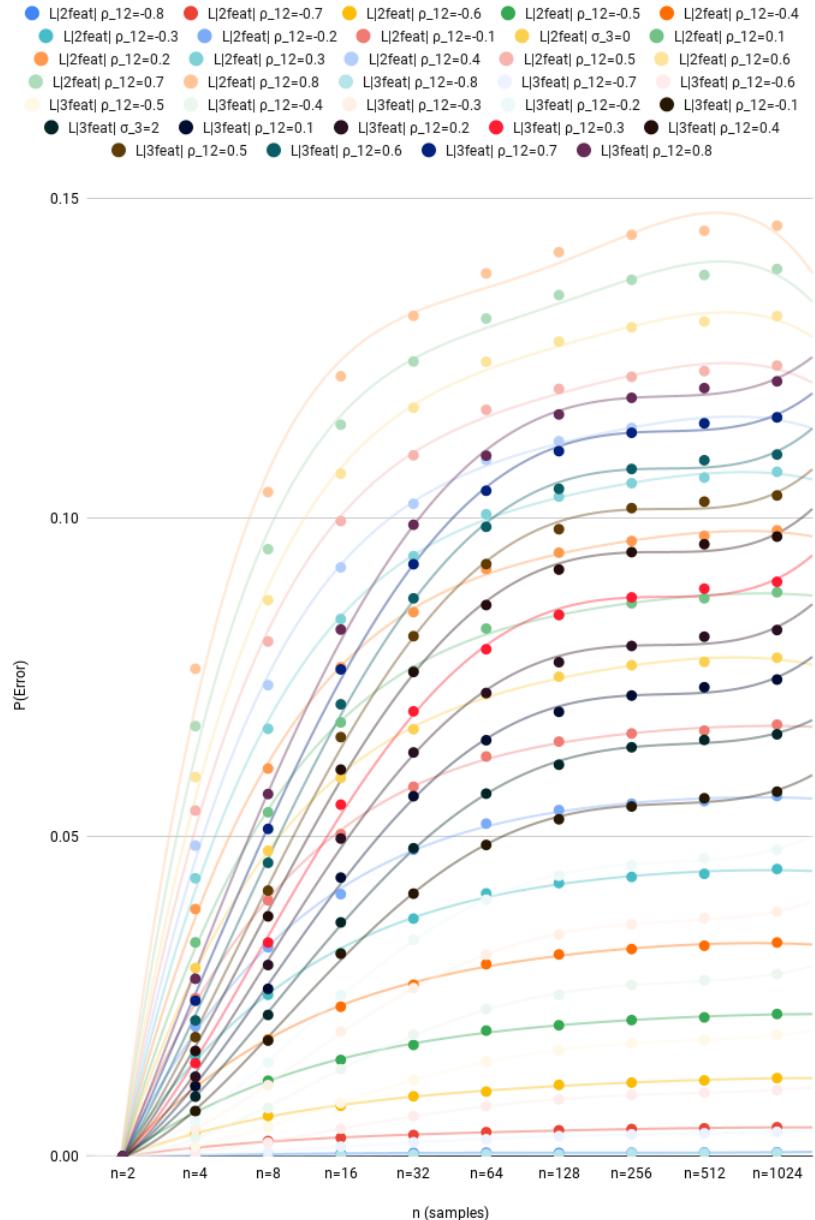


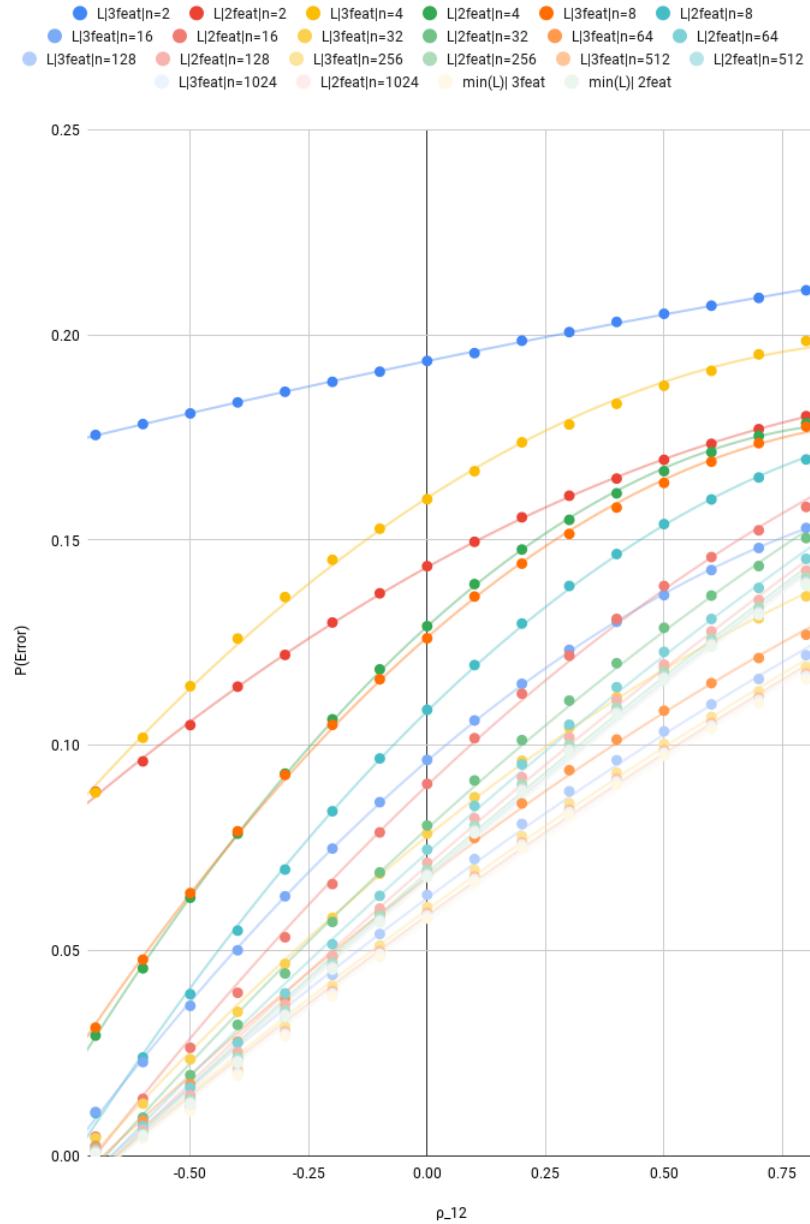
(c) 11

Figura 17 – Cenário 2 - Indicadores 4/4

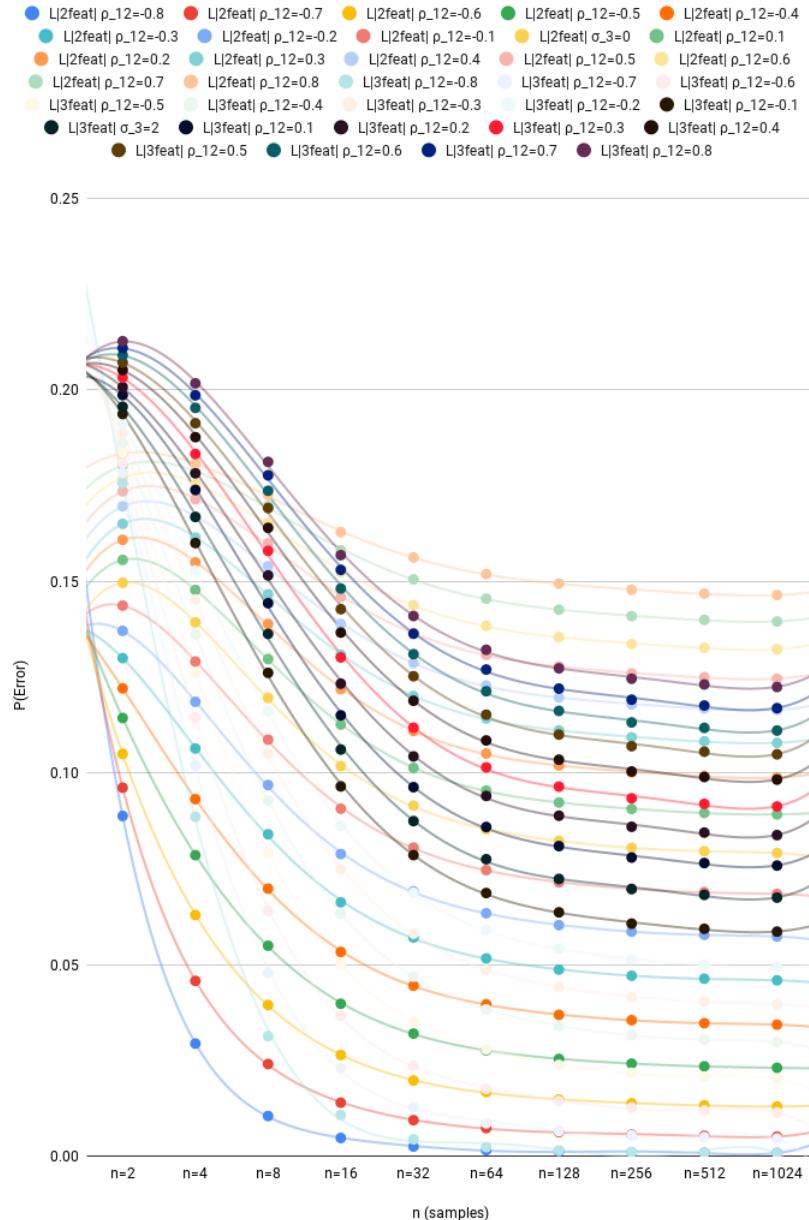
Loss | EMPIRICAL_TRAIN vs ρ_{12} Figura 18 – Cenário 2 - $\hat{L}(\hat{h}^{(D)})$ vs ρ_{12}

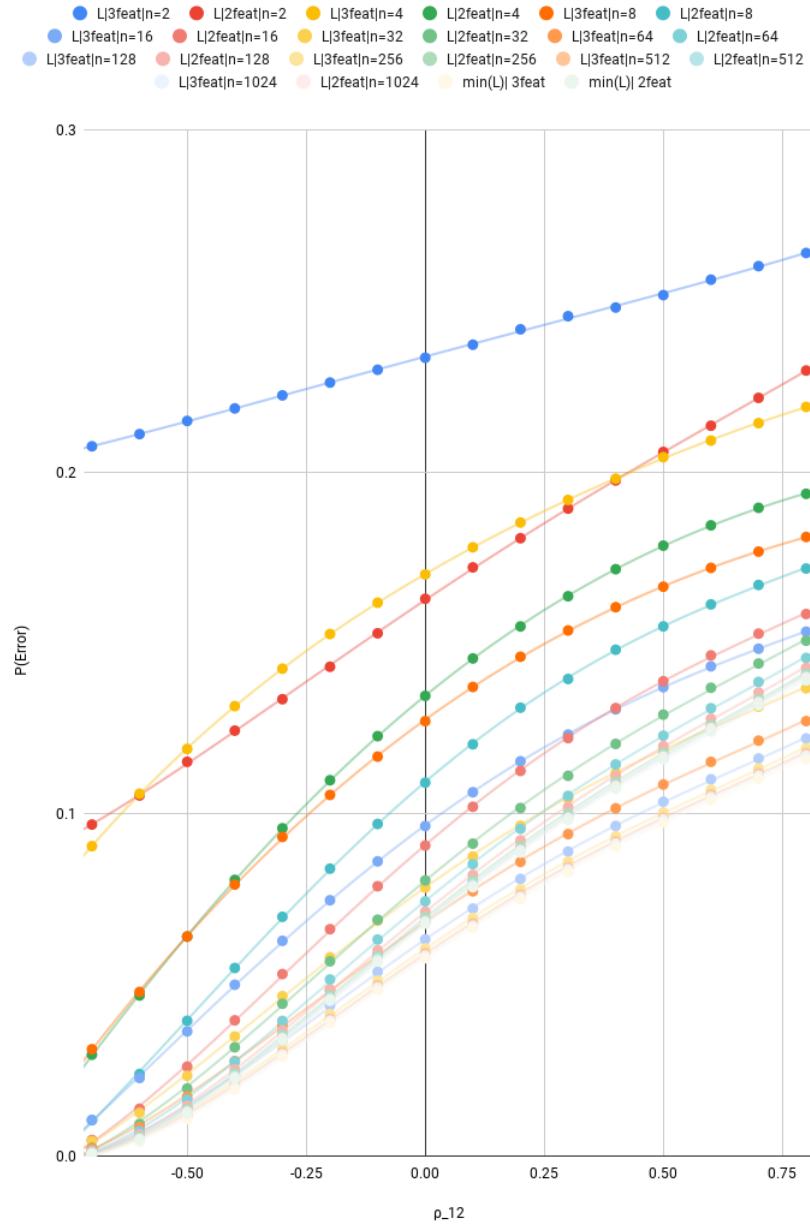
Loss | EMPIRICAL_TRAIN vs n

Figura 19 – Cenário 2 - $\hat{L}(\hat{h}^{(D)})$ vs n

Loss | THEORETICAL vs ρ_{12} Figura 20 – Cenário 2 - $L(\hat{h}^{(D)})$ vs ρ_{12}

Loss | THEORETICAL vs n

Figura 21 – Cenário 2 - $L(\hat{h}^{(D)})$ vs n

Loss | EMPIRICAL_TEST vs ρ_{12} Figura 22 – Cenário 2 - $\hat{L}(\hat{h}^{(D)}, D')$ vs ρ_{12}

Loss | EMPIRICAL_TEST vs n

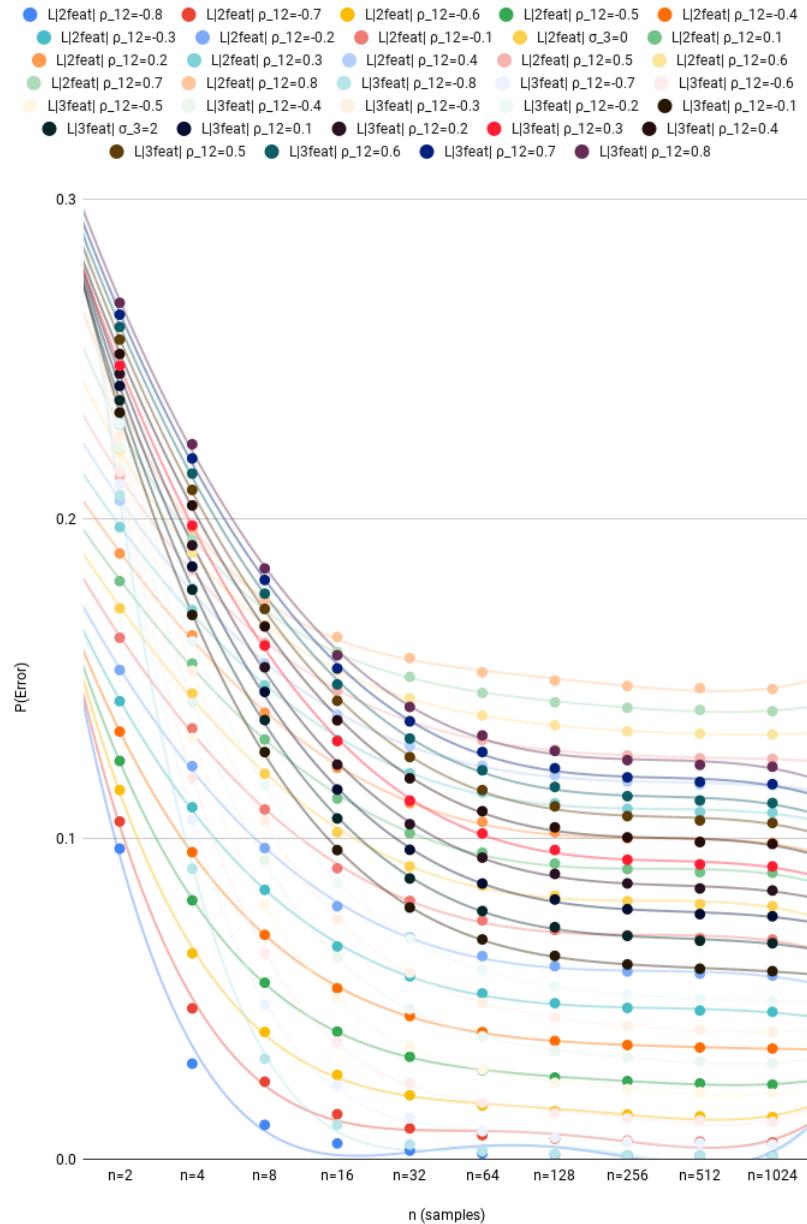
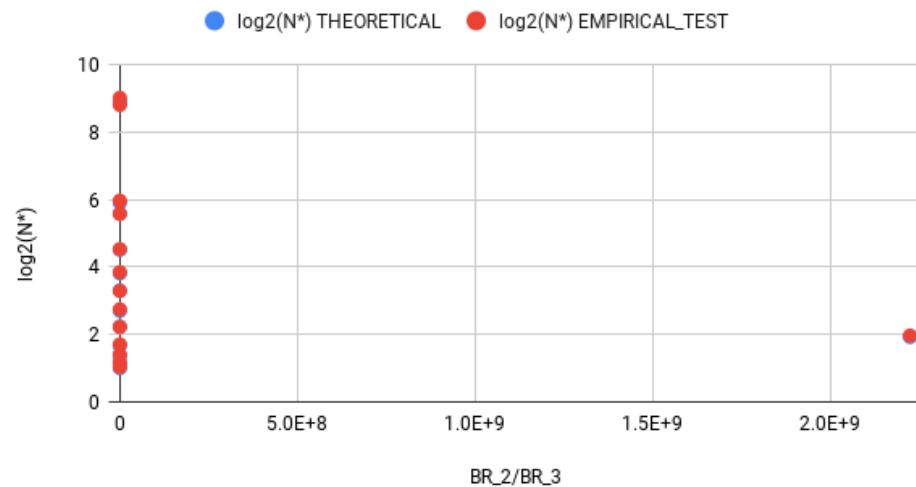
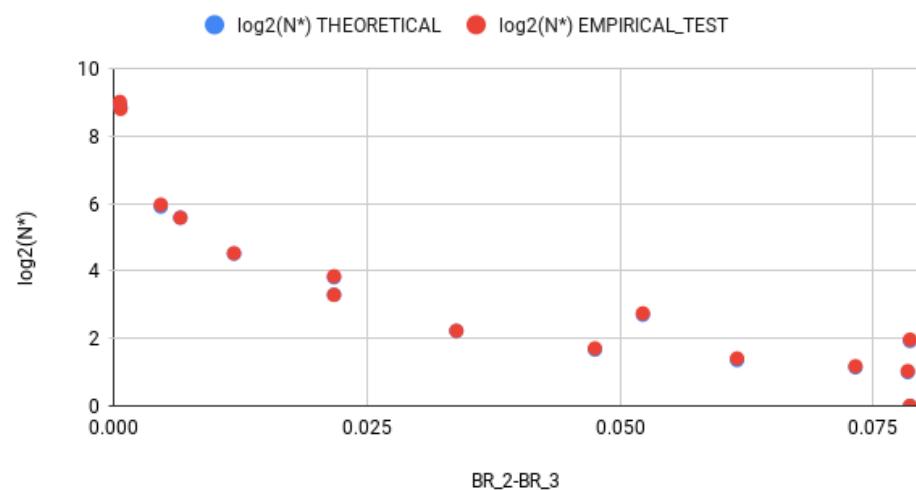


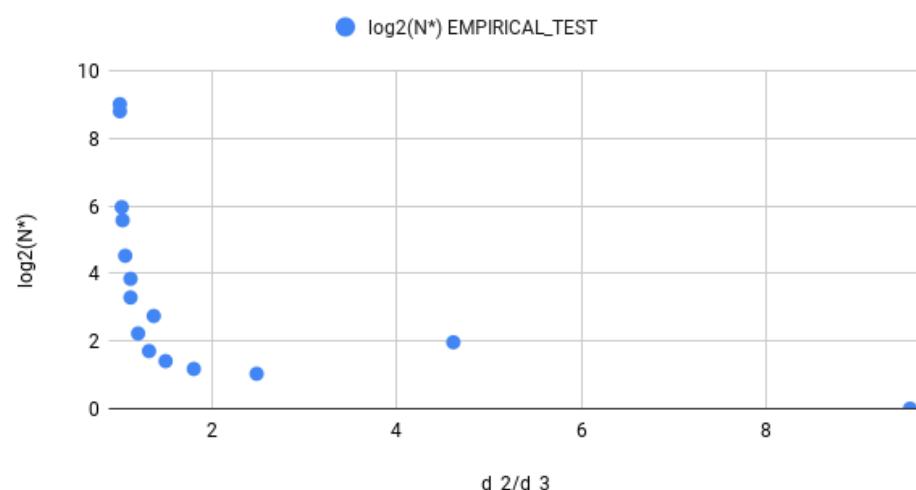
Figura 23 – Cenário 2 - $\hat{L}(\hat{h}^{(D)}, D')$ vs n

$\log_2(N^*)$ vs BR_2/BR_3

(a)

 $\log_2(N^*)$ vs BR_2-BR_3

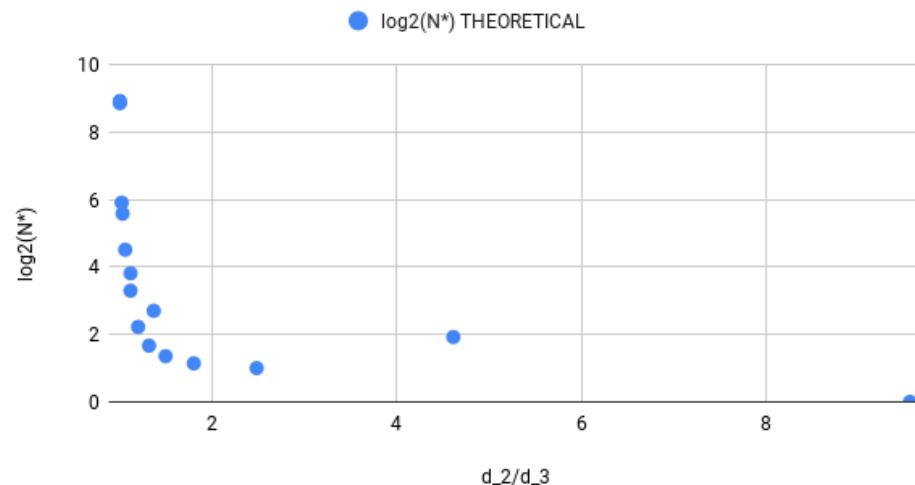
(b)

 $\log_2(N^*)$ vs d_2/d_3 | Empirical

(c)

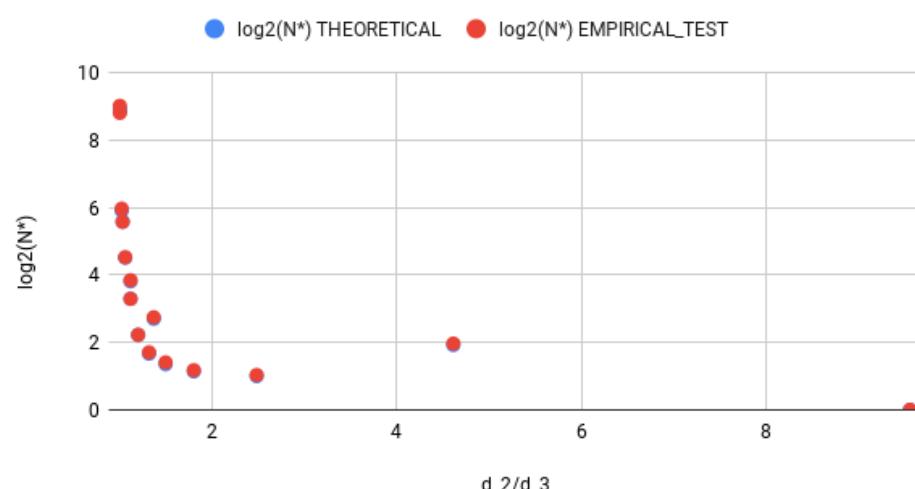
Figura 24 – Cenário 3 - indicadores 1/4

$\log_2(N^*)$ vs d_2/d_3 | Theoretical



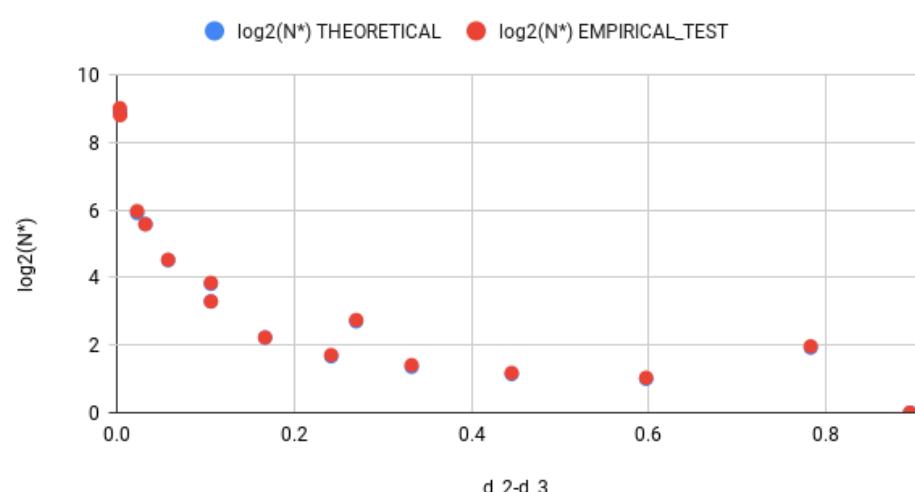
(a)

$\log_2(N^*)$ vs d_2/d_3



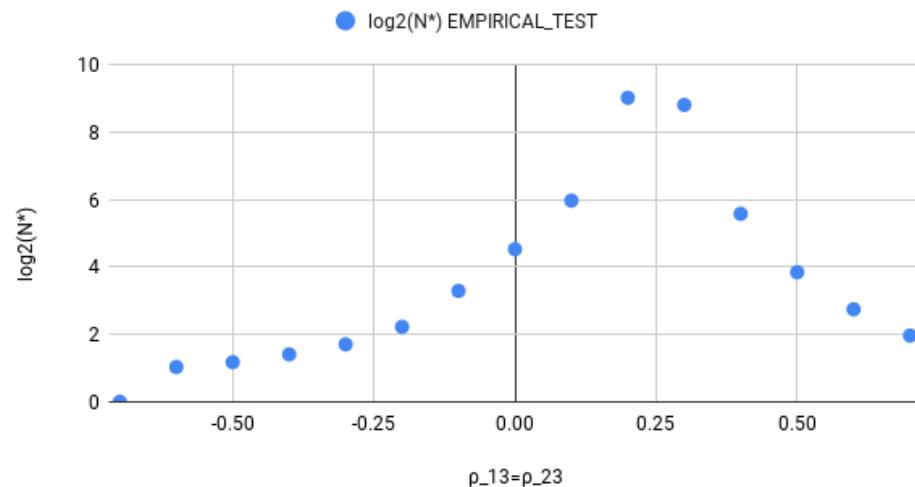
(b)

$\log_2(N^*)$ vs d_2-d_3

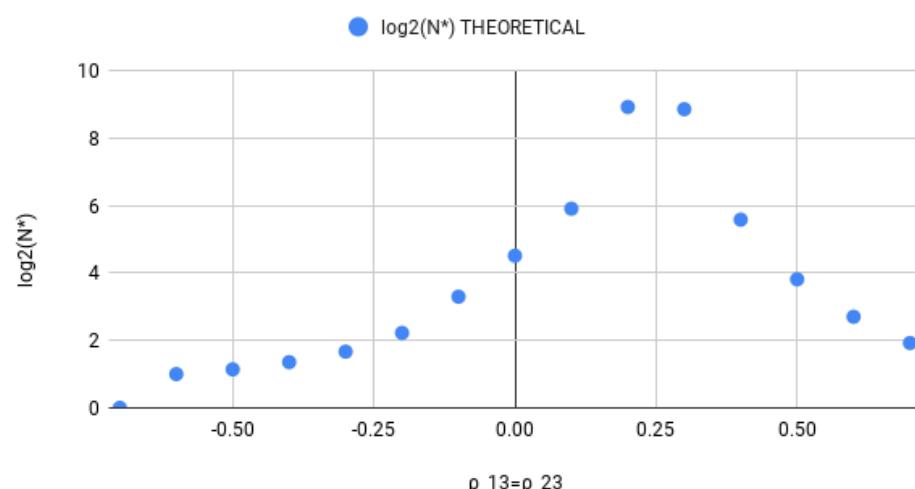


(c)

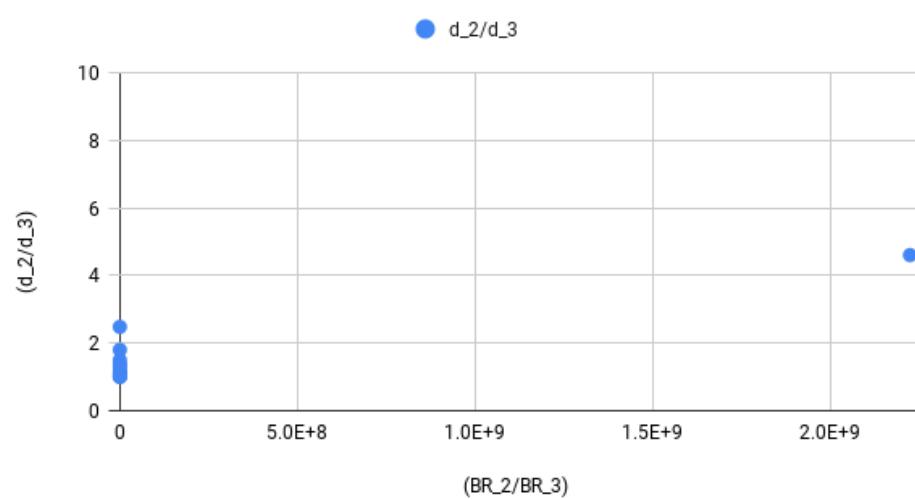
Figura 25 – Cenário 3 - indicadores 2/4

$\log_2(N^*)$ vs $\rho_{13}=\rho_{23}$ | Empirical

(a)

 $\log_2(N^*)$ vs $\rho_{13}=\rho_{23}$ | Theoretical

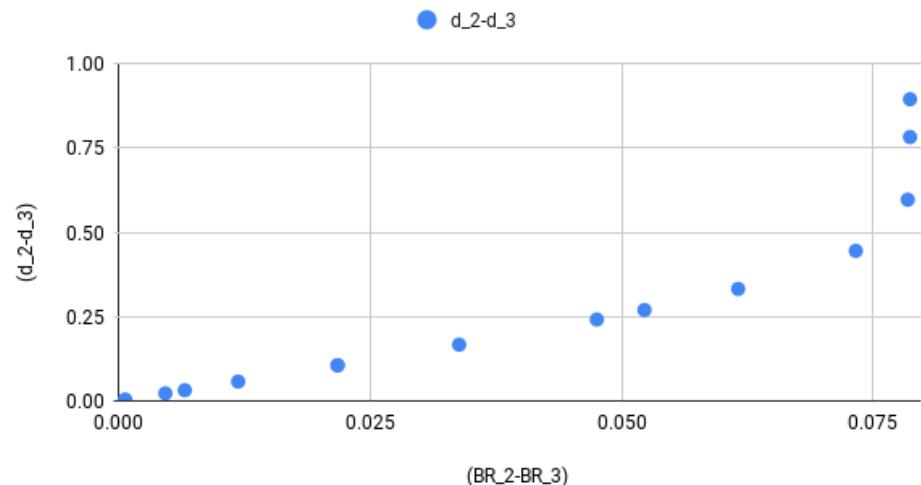
(b)

 (d_2/d_3) vs (BR_2/BR_3) 

(c)

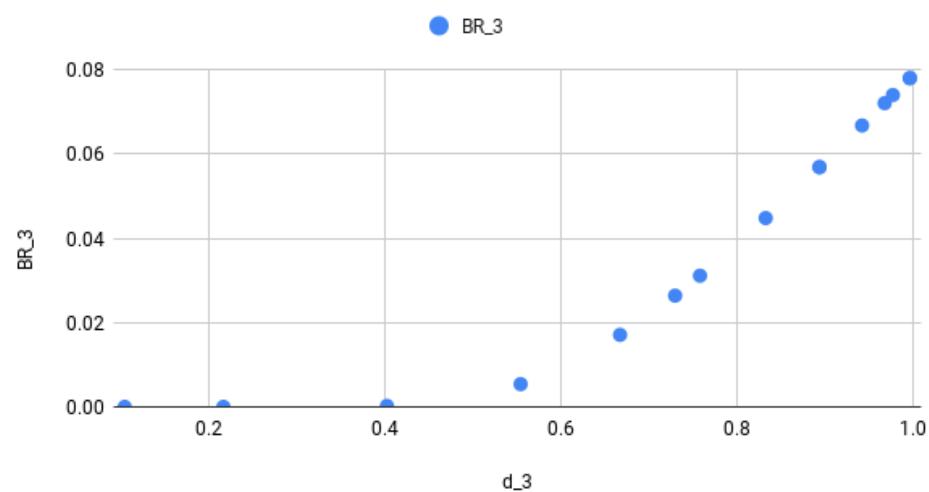
Figura 26 – Cenário 3 - indicadores 3/4

(d_2-d_3) vs (BR_2-BR_3)



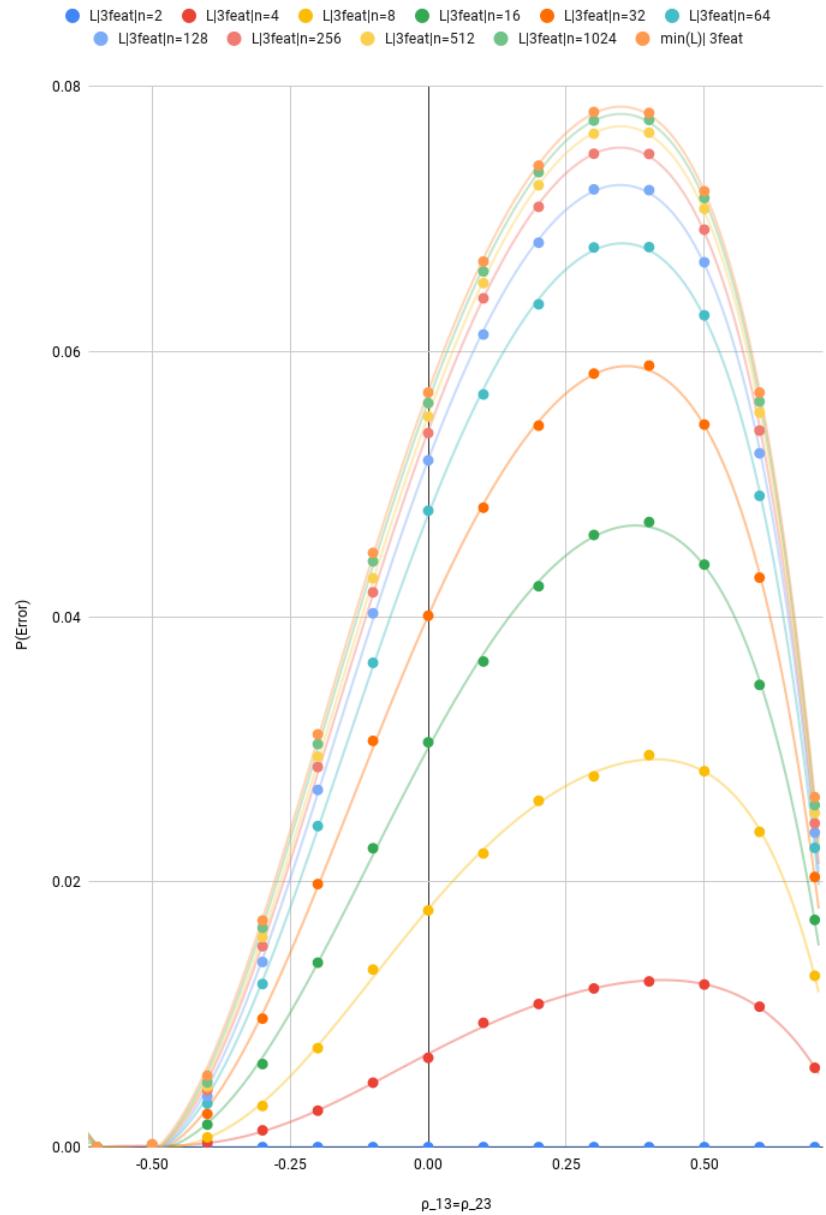
(a)

BR_3 vs d_3



(b)

Figura 27 – Cenário 3 - indicadores 4/4

Loss | EMPIRICAL_TRAIN vs $\rho_{13}=\rho_{23}$ Figura 28 – Cenário 3 - $\hat{L}(\hat{h}^{(D)})$ vs $\rho_{13} = \rho_{23}$

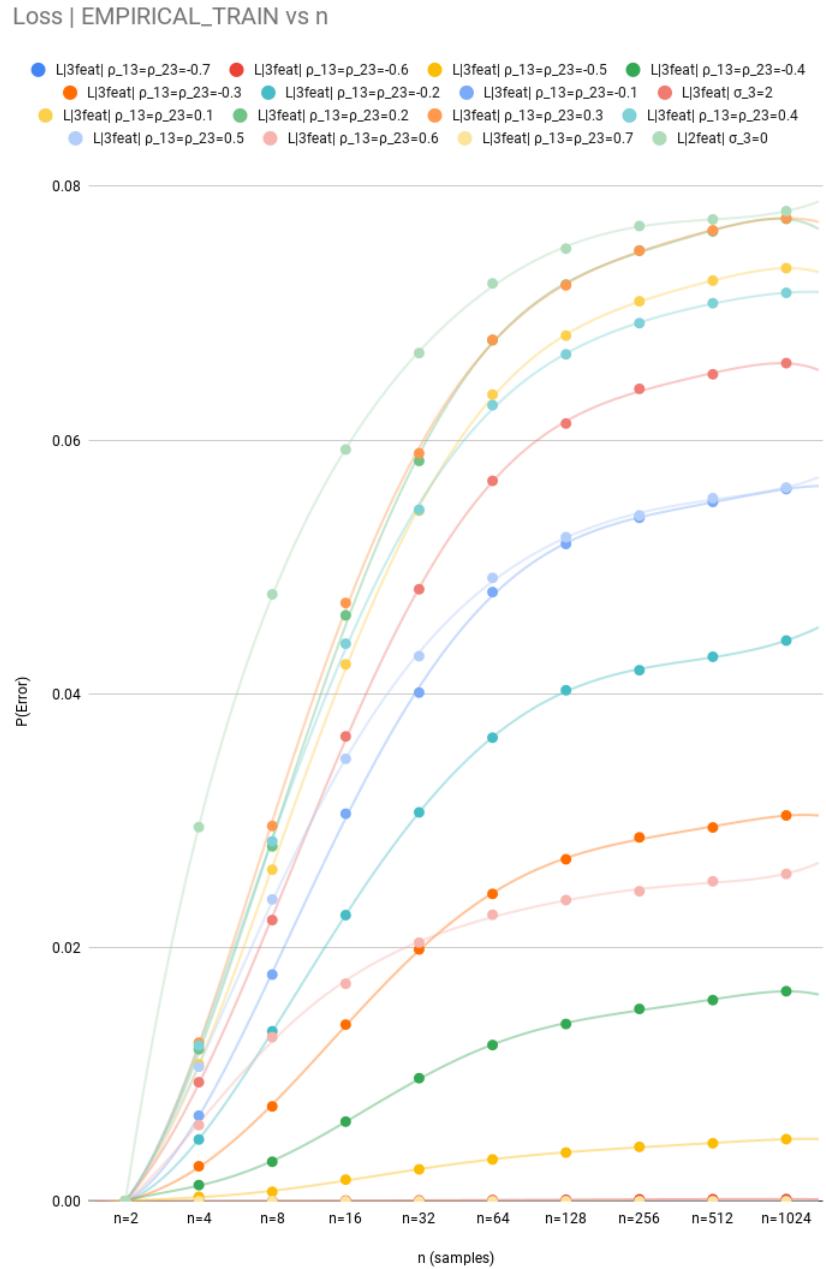
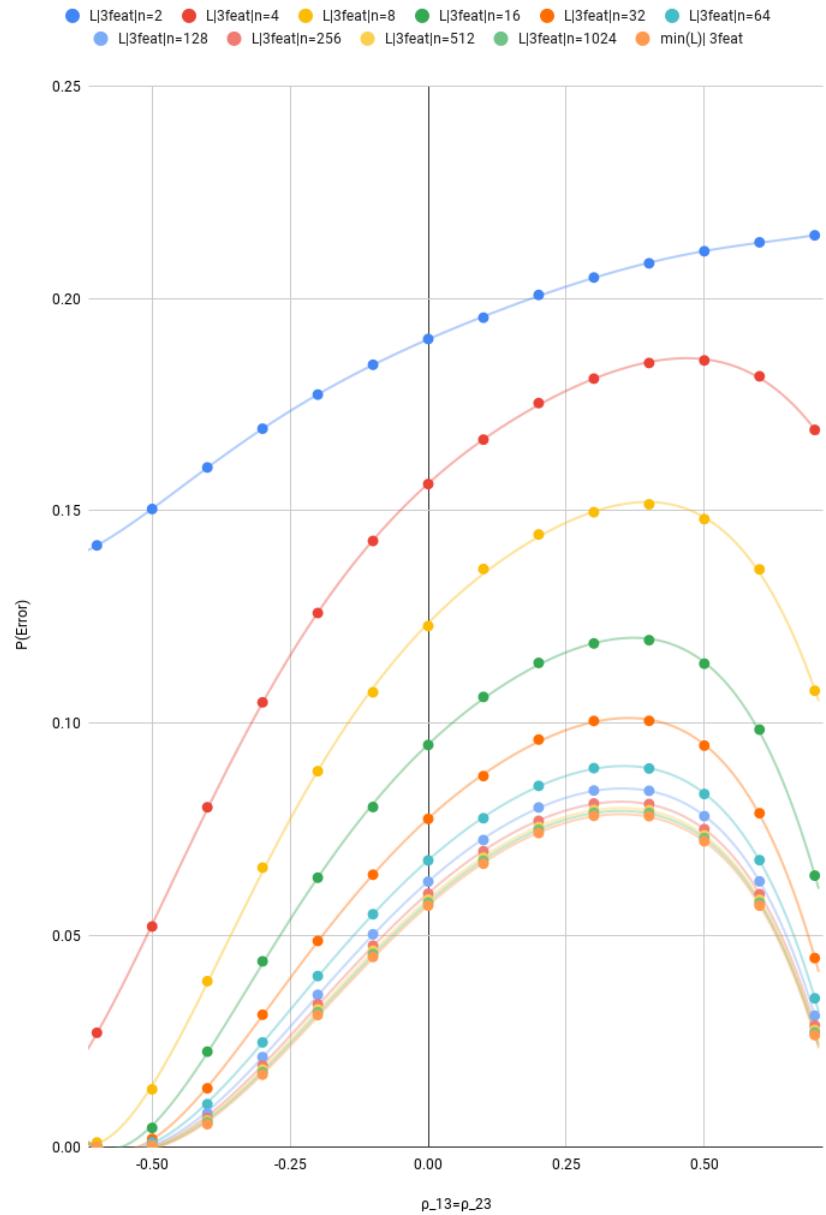
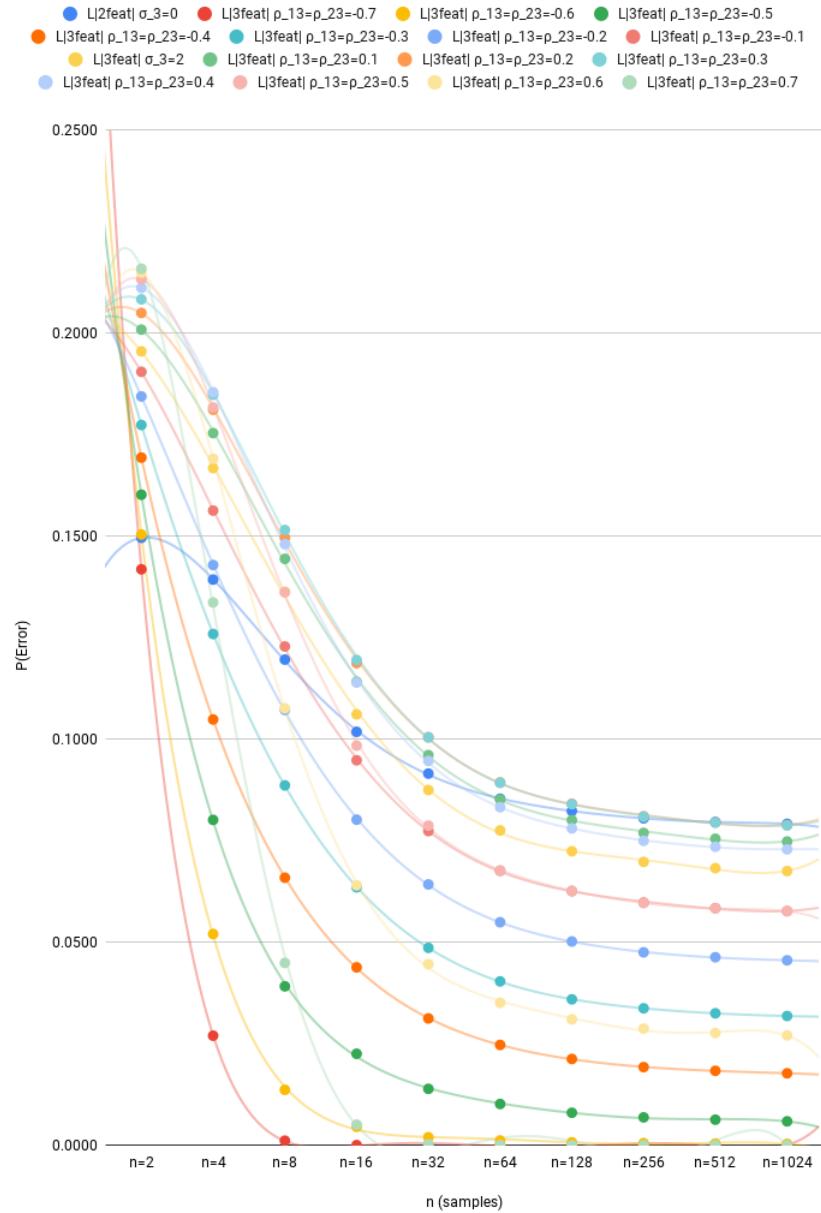


Figura 29 – Cenário 3 - $\hat{L}(\hat{h}^{(D)})$ vs n

Loss | THEORETICAL vs $\rho_{13}=\rho_{23}$ Figura 30 – Cenário 3 - $L(\hat{h}^{(D)})$ vs $\rho_{13} = \rho_{23}$

Loss | THEORETICAL vs n

Figura 31 – Cenário 3 - $L(\hat{h}^{(D)})$ vs n

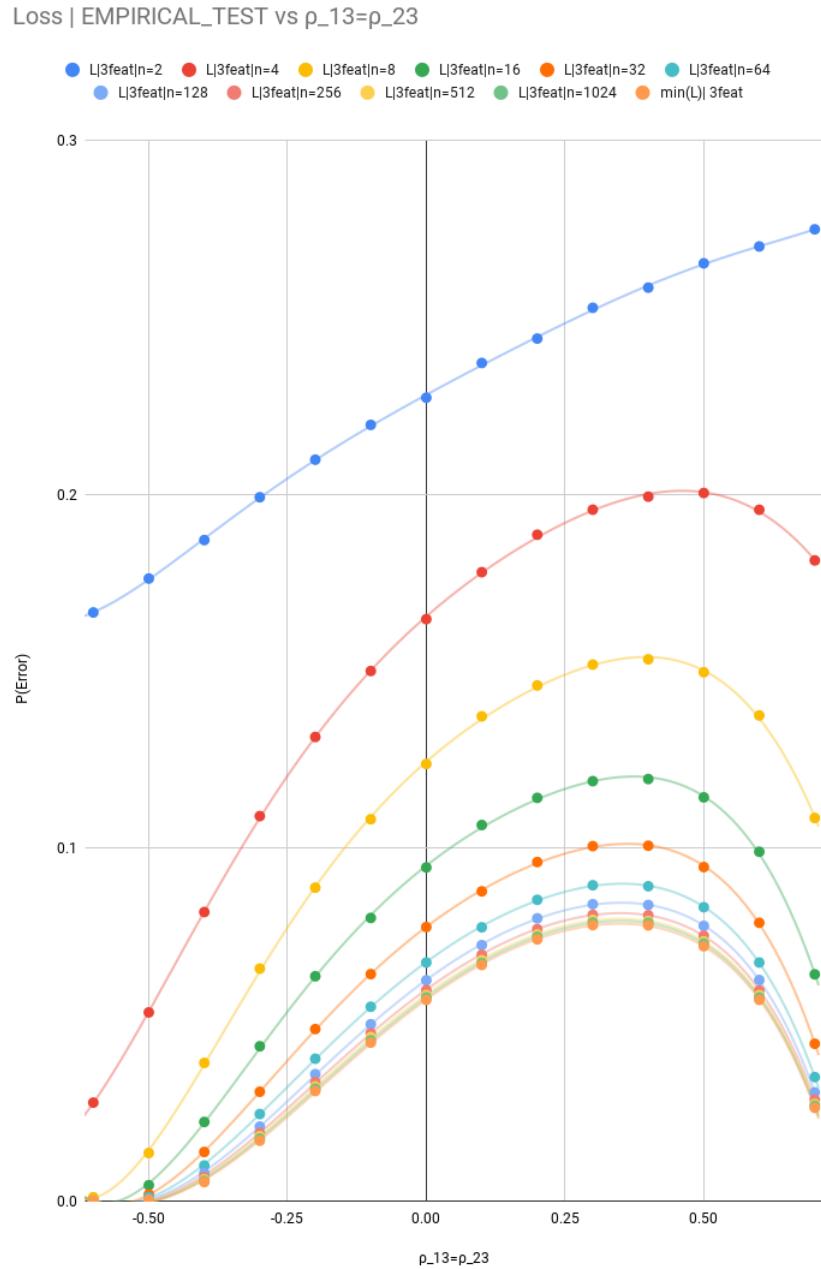
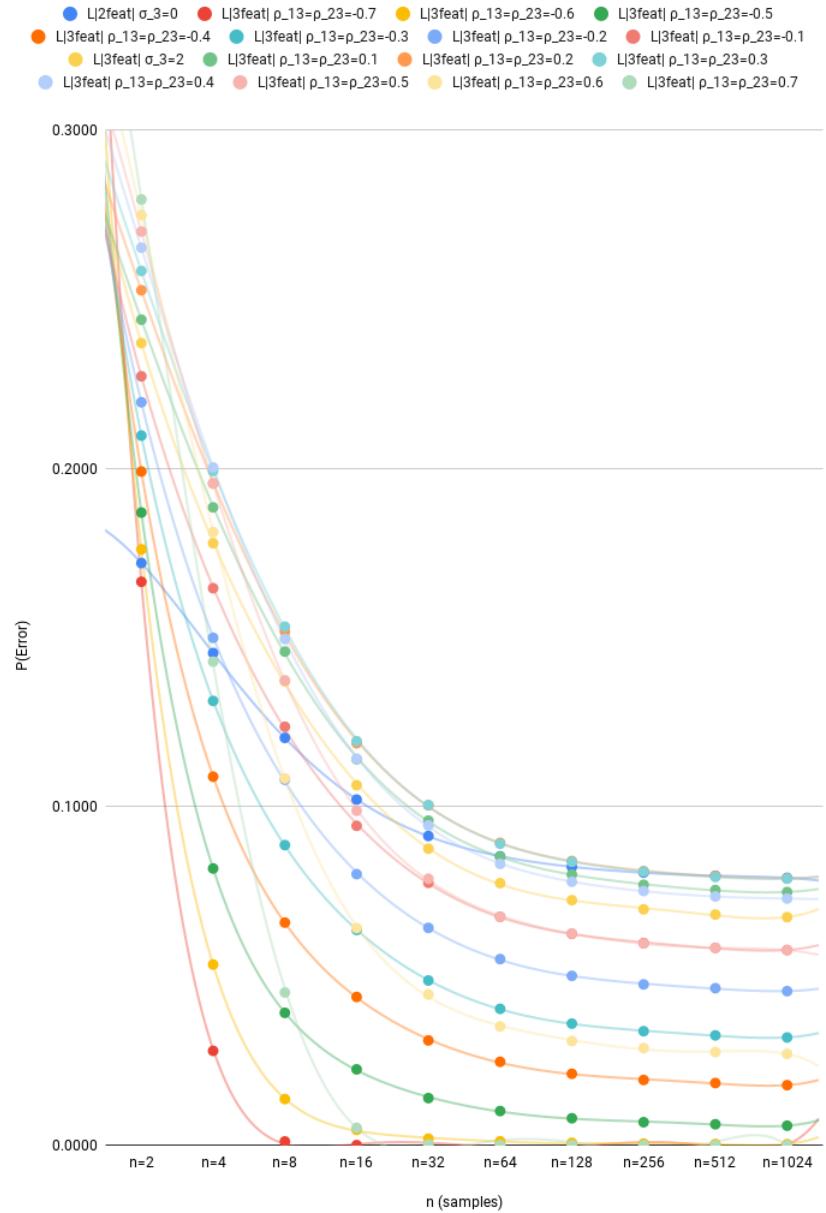
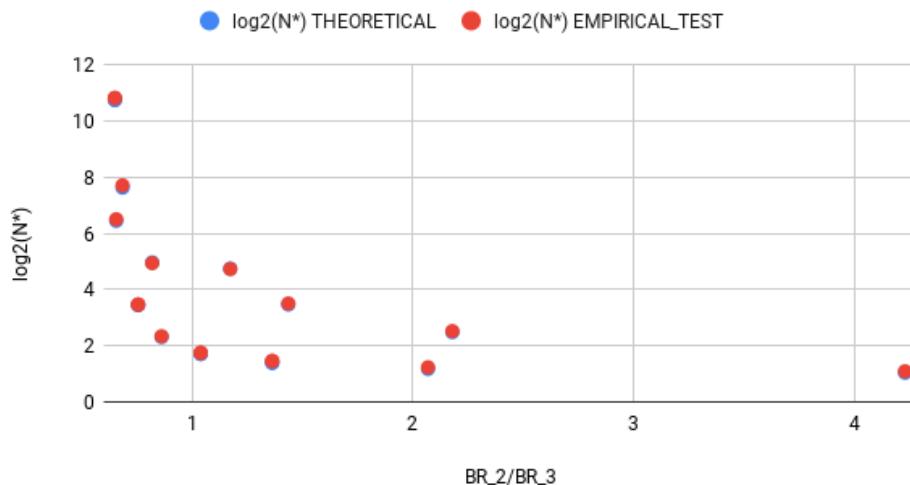


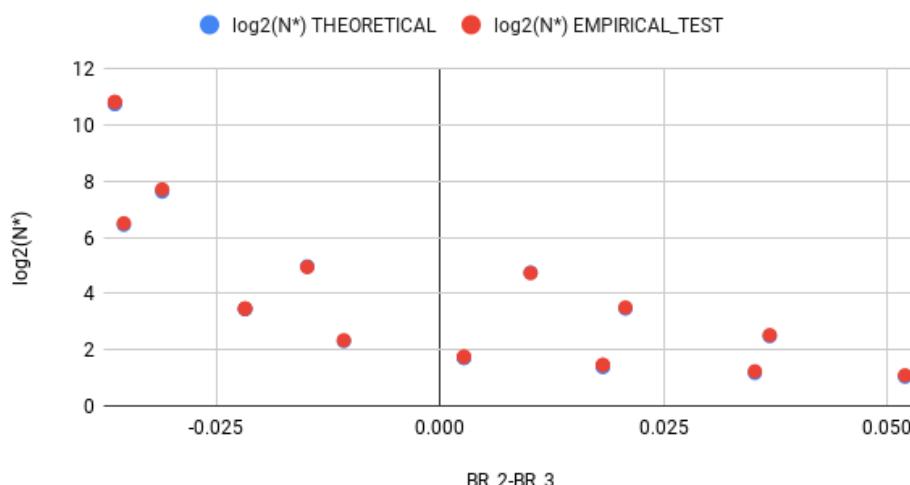
Figura 32 – Cenário 3 - $\hat{L}(\hat{h}^{(D)}, D')$ vs $\rho_{13} = \rho_{23}$

Loss | EMPIRICAL_TEST vs n

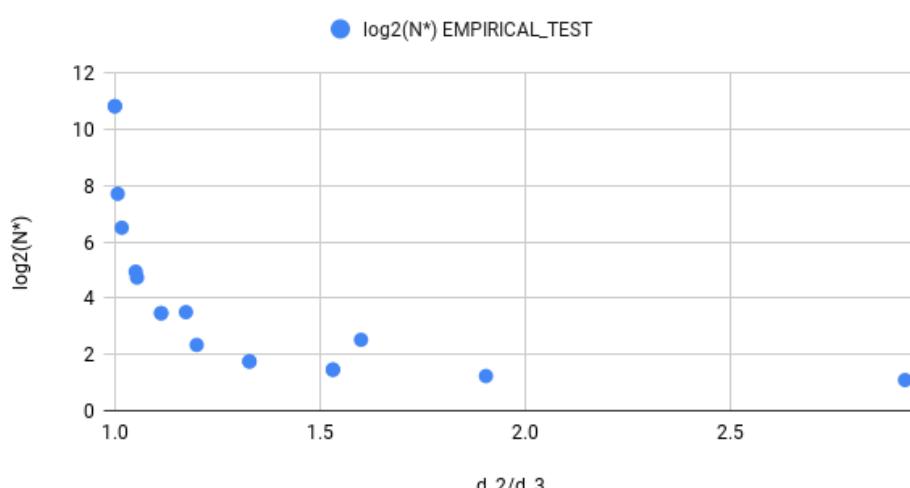
Figura 33 – Cenário 3 - $\hat{L}(\hat{h}^{(D)}, D')$ vs n

$\log_2(N^*)$ vs BR_2/BR_3

(a)

 $\log_2(N^*)$ vs BR_2-BR_3

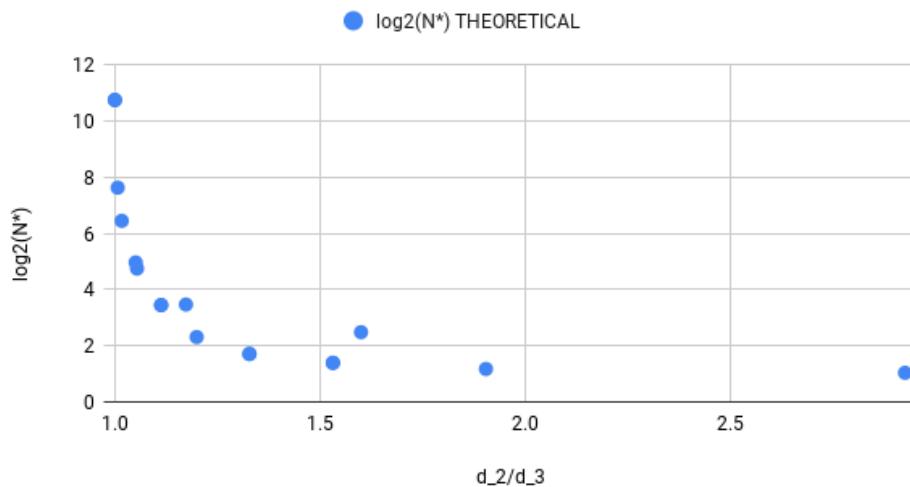
(b)

 $\log_2(N^*)$ vs d_2/d_3 | Empirical

(c)

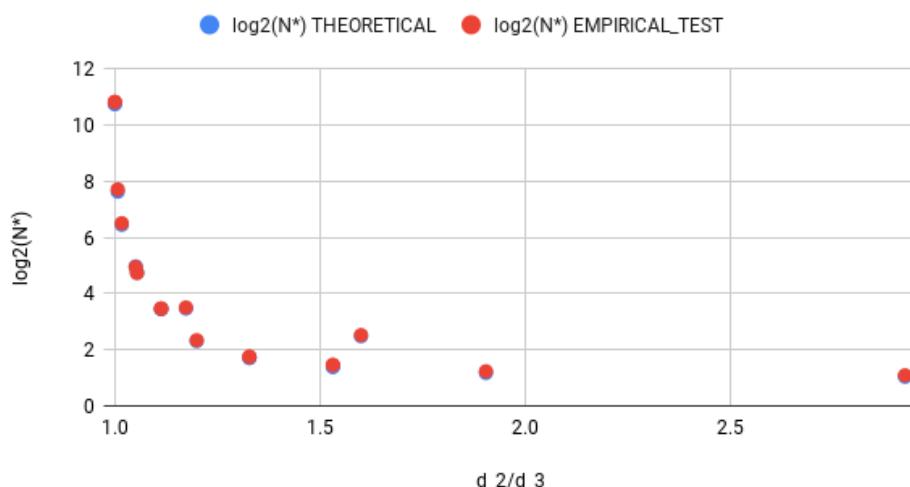
Figura 34 – Cenário 4 - indicadores 1/4

$\log_2(N^*)$ vs d_2/d_3 | Theoretical



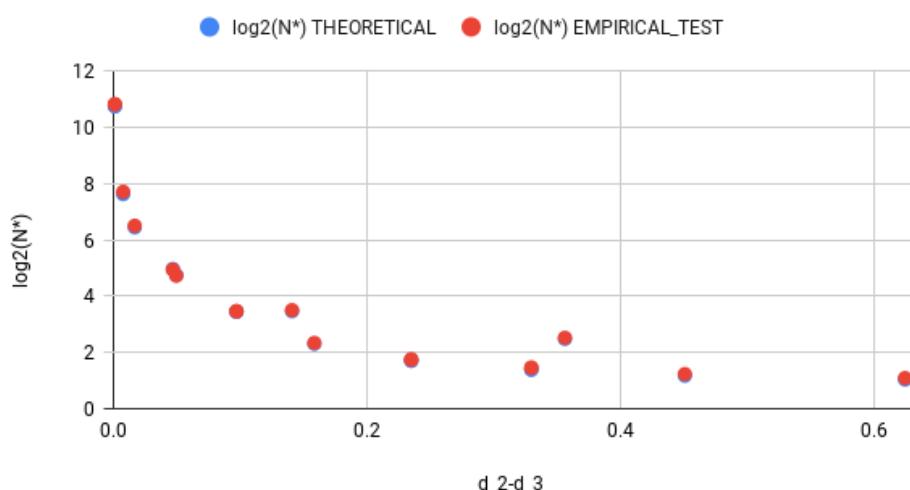
(a)

$\log_2(N^*)$ vs d_2/d_3



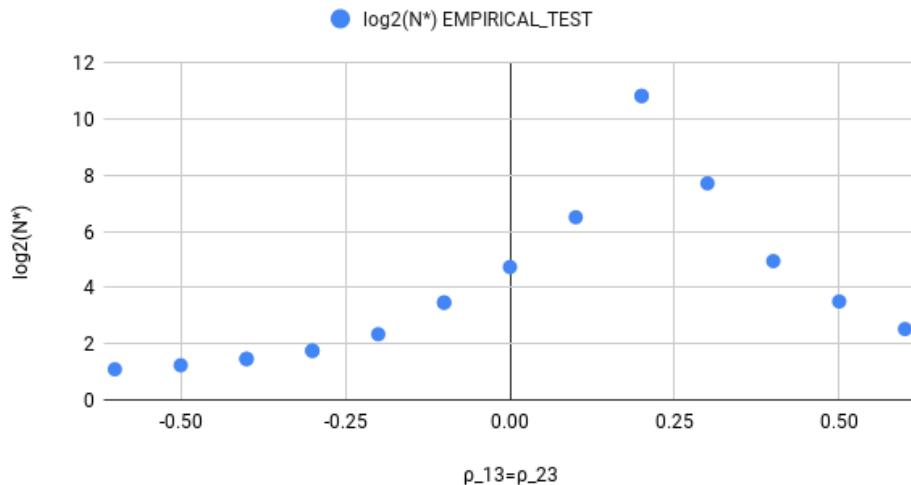
(b)

$\log_2(N^*)$ vs d_2-d_3

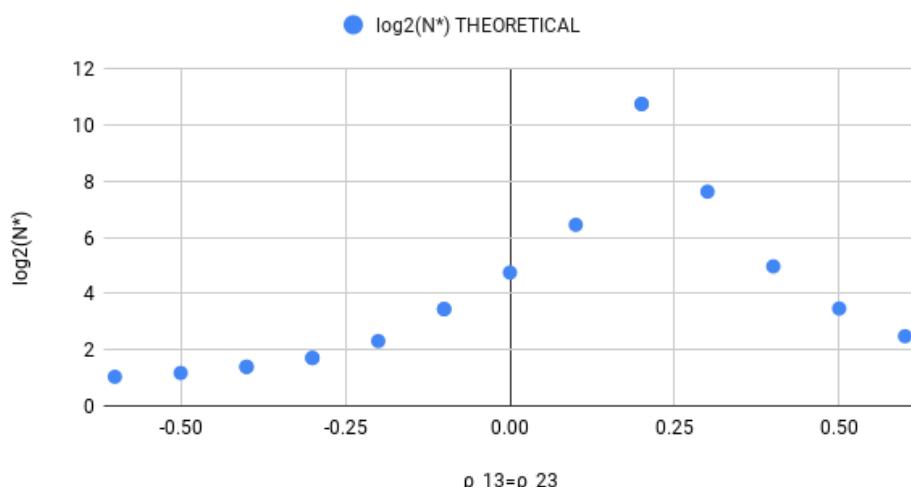


(c)

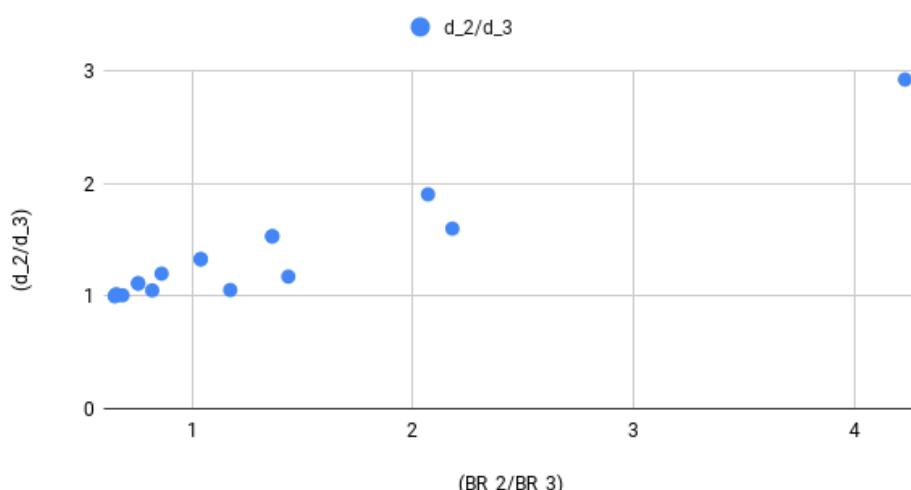
Figura 35 – Cenário 4 - indicadores 2/4

$\log_2(N^*)$ vs $\rho_{13}=\rho_{23}$ | Empirical

(a)

 $\log_2(N^*)$ vs $\rho_{13}=\rho_{23}$ | Theoretical

(b)

 (d_2/d_3) vs (BR_2/BR_3) 

(c)

Figura 36 – Cenário 4 - indicadores 3/4

Tabela 4 – Cenário 1 - $L(\hat{h}^{(D)})$: Erro Teórico

σ_3	$n = 2$	$n = 4$	$n = 8$	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$	$n = 512$	$n = 1024$	$\min(L)$
2 atributos	0.1496	0.1393	0.1196	0.1018	0.0915	0.0853	0.0823	0.0805	0.0796	0.0791	0.0786
1.3	0.1420	0.1260	0.1077	0.0870	0.0721	0.0635	0.0590	0.0565	0.0551	0.0544	0.0537
1.4	0.1506	0.1332	0.1131	0.0909	0.0755	0.0666	0.0620	0.0594	0.0579	0.0572	0.0566
1.5	0.1589	0.1399	0.1175	0.0945	0.0783	0.0691	0.0644	0.0618	0.0604	0.0597	0.0590
1.6	0.1674	0.1460	0.1226	0.0976	0.0807	0.0714	0.0665	0.0639	0.0624	0.0617	0.0610
1.7	0.1751	0.1520	0.1267	0.1002	0.0828	0.0734	0.0684	0.0657	0.0642	0.0635	0.0628
1.8	0.1823	0.1568	0.1303	0.1025	0.0846	0.0751	0.0698	0.0672	0.0657	0.0650	0.0643
1.9	0.1894	0.1623	0.1327	0.1045	0.0863	0.0765	0.0713	0.0686	0.0671	0.0663	0.0657
2	0.1955	0.1668	0.1363	0.1061	0.0875	0.0775	0.0724	0.0698	0.0682	0.0675	0.0668
3	0.2414	0.1967	0.1524	0.1155	0.0950	0.0842	0.0789	0.0760	0.0746	0.0739	0.0731
4	0.2651	0.2123	0.1592	0.1192	0.0975	0.0868	0.0813	0.0784	0.0769	0.0762	0.0755
5	0.2784	0.2210	0.1633	0.1208	0.0988	0.0879	0.0824	0.0795	0.0781	0.0774	0.0766
6	0.2859	0.2261	0.1643	0.1216	0.0996	0.0885	0.0831	0.0801	0.0787	0.0780	0.0772
7	0.2902	0.2292	0.1660	0.1220	0.0999	0.0890	0.0834	0.0805	0.0791	0.0783	0.0776
8	0.2930	0.2314	0.1671	0.1223	0.1000	0.0890	0.0837	0.0808	0.0793	0.0786	0.0778
9	0.2953	0.2335	0.1674	0.1229	0.1002	0.0893	0.0839	0.0809	0.0795	0.0788	0.0780
10	0.2970	0.2347	0.1676	0.1227	0.1002	0.0894	0.0840	0.0811	0.0796	0.0789	0.0781

Tabela 5 – Cenário 1 - $\hat{L}(\hat{h}^{(D)}, D')$: Erro Empírico de Teste

σ_3	$n = 2$	$n = 4$	$n = 8$	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$	$n = 512$	$n = 1024$	$\min(L)$
2 atributos	0.1587	0.1307	0.1081	0.0866	0.0722	0.0636	0.0589	0.0564	0.0550	0.0543	0.0537
1.3	0.1704	0.1382	0.1138	0.0912	0.0756	0.0666	0.0618	0.0594	0.0578	0.0572	0.0566
1.4	0.1826	0.1463	0.1188	0.0945	0.0784	0.0690	0.0643	0.0617	0.0603	0.0596	0.0590
1.5	0.1943	0.1536	0.1233	0.0976	0.0808	0.0714	0.0663	0.0639	0.0625	0.0617	0.0610
1.6	0.2056	0.1602	0.1276	0.1002	0.0829	0.0732	0.0681	0.0656	0.0642	0.0634	0.0628
1.7	0.2166	0.1660	0.1305	0.1025	0.0847	0.0750	0.0696	0.0671	0.0657	0.0649	0.0643
1.8	0.2261	0.1724	0.1343	0.1047	0.0863	0.0763	0.0711	0.0686	0.0670	0.0662	0.0657
1.9	0.2371	0.1779	0.1371	0.1064	0.0876	0.0775	0.0724	0.0697	0.0682	0.0674	0.0668
2	0.3195	0.2174	0.1547	0.1158	0.0951	0.0841	0.0789	0.0758	0.0744	0.0738	0.0731
3	0.3706	0.2393	0.1624	0.1190	0.0977	0.0869	0.0812	0.0782	0.0768	0.0761	0.0755
4	0.4034	0.2551	0.1661	0.1210	0.0991	0.0879	0.0823	0.0795	0.0780	0.0773	0.0766
5	0.4248	0.2652	0.1687	0.1217	0.0999	0.0885	0.0830	0.0800	0.0787	0.0778	0.0772
6	0.4393	0.2710	0.1704	0.1223	0.1001	0.0890	0.0834	0.0804	0.0791	0.0784	0.0776
7	0.4498	0.2761	0.1716	0.1227	0.1003	0.0891	0.0837	0.0807	0.0793	0.0786	0.0778
8	0.4571	0.2802	0.1730	0.1229	0.1004	0.0893	0.0840	0.0808	0.0795	0.0788	0.0780
9	0.4641	0.2848	0.1731	0.1231	0.1005	0.0894	0.0840	0.0810	0.0795	0.0788	0.0781
10	0.2970	0.2347	0.1676	0.1227	0.1002	0.0894	0.0840	0.0811	0.0796	0.0789	0.0781

Tabela 6 – Cenário 2 - indicadores

σ_1	σ_2	σ_3	ρ_{12}	ρ_{13}	ρ_{23}	$BR_2 - BR_3$	$\frac{BR_2}{BR_3}$	$\frac{d_2}{d_3}$	$d_2 - d_3$	theoretical	$\log_2(n*)$ empirical, D'
1	1	2	-0.8	0	0	0.0001	1.1453	0.0055	1.0125	9.1921	9.5994
1	1	2	-0.7	0	0	0.0006	1.1503	0.0100	1.0185	7.5813	7.5837
1	1	2	-0.6	0	0	0.0017	1.1549	0.0153	1.0247	6.6653	6.6052
1	1	2	-0.5	0	0	0.0031	1.1592	0.0211	1.0308	6.1293	6.0872
1	1	2	-0.4	0	0	0.0048	1.1632	0.0274	1.0367	5.6412	5.6305
1	1	2	-0.3	0	0	0.0065	1.1670	0.0344	1.0428	5.2560	5.2794
1	1	2	-0.2	0	0	0.0083	1.1706	0.0416	1.0488	4.9574	4.9592
1	1	2	-0.1	0	0	0.0101	1.1740	0.0492	1.0547	4.7477	4.7310
1	1	2	0	0	0	0.0118	1.1773	0.0572	1.0607	4.5153	4.5309
1	1	2	0.1	0	0	0.0136	1.1804	0.0655	1.0666	4.3233	4.3525
1	1	2	0.2	0	0	0.0152	1.1834	0.0740	1.0724	4.1769	4.1362
1	1	2	0.3	0	0	0.0169	1.1863	0.0827	1.0782	3.9417	3.9686
1	1	2	0.4	0	0	0.0184	1.1890	0.0916	1.0840	3.8182	3.8693
1	1	2	0.5	0	0	0.0200	1.1917	0.1008	1.0897	3.7441	3.7699
1	1	2	0.6	0	0	0.0214	1.1943	0.1102	1.0955	3.6603	3.6889
1	1	2	0.7	0	0	0.0229	1.1968	0.1198	1.1012	3.6103	3.6415
1	1	2	0.8	0	0	0.0242	1.1993	0.1295	1.1068	3.5899	3.6329

Tabela 7 – Cenário 2 - $\hat{L}(\hat{h}^{(D)})$, $\sigma_3 = 0$: Erro Empírico de Treino

ρ_{12}	$n = 2$	$n = 4$	$n = 8$	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$	$n = 512$	$n = 1024$	$\min(L)$
-0.8	0.0000	0.0002	0.0004	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0008
-0.7	0.0000	0.0013	0.0024	0.0029	0.0034	0.0038	0.0041	0.0043	0.0044	0.0046	0.0049
-0.6	0.0000	0.0036	0.0063	0.0079	0.0094	0.0102	0.0112	0.0116	0.0119	0.0123	0.0127
-0.5	0.0000	0.0070	0.0119	0.0151	0.0175	0.0197	0.0205	0.0213	0.0218	0.0223	0.0228
-0.4	0.0000	0.0112	0.0183	0.0234	0.0269	0.0301	0.0316	0.0325	0.0330	0.0335	0.0339
-0.3	0.0000	0.0158	0.0253	0.0319	0.0372	0.0412	0.0428	0.0438	0.0442	0.0450	0.0455
-0.2	0.0000	0.0204	0.0327	0.0411	0.0480	0.0521	0.0542	0.0552	0.0556	0.0564	0.0569
-0.1	0.0000	0.0248	0.0401	0.0504	0.0579	0.0626	0.0649	0.0662	0.0666	0.0676	0.0680
0	0.0000	0.0295	0.0478	0.0593	0.0669	0.0723	0.0751	0.0769	0.0774	0.0780	0.0786
0.1	0.0000	0.0335	0.0539	0.0679	0.0760	0.0826	0.0848	0.0866	0.0873	0.0883	0.0888
0.2	0.0000	0.0387	0.0607	0.0766	0.0852	0.0919	0.0945	0.0963	0.0971	0.0980	0.0984
0.3	0.0000	0.0435	0.0669	0.0841	0.0940	0.1005	0.1033	0.1054	0.1062	0.1071	0.1074
0.4	0.0000	0.0486	0.0738	0.0922	0.1021	0.1090	0.1119	0.1140	0.1148	0.1157	0.1160
0.5	0.0000	0.0541	0.0806	0.0994	0.1097	0.1168	0.1201	0.1220	0.1229	0.1238	0.1241
0.6	0.0000	0.0594	0.0871	0.1069	0.1172	0.1244	0.1275	0.1297	0.1307	0.1315	0.1318
0.7	0.0000	0.0673	0.0950	0.1145	0.1244	0.1311	0.1348	0.1372	0.1379	0.1389	0.1390
0.8	0.0000	0.0763	0.1040	0.1221	0.1315	0.1382	0.1415	0.1442	0.1448	0.1456	0.1459

Tabela 8 – Cenário 2 - $\hat{L}(\hat{h}^{(D)})$, $\sigma_3 > 0$: Erro Empírico de Treino

ρ_{12}	$n = 2$	$n = 4$	$n = 8$	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$	$n = 512$	$n = 1024$	$\min(L)$
-0.8	0.0000	0.0000	0.0001	0.0002	0.0003	0.0004	0.0003	0.0005	0.0005	0.0006	0.0007
-0.7	0.0000	0.0003	0.0007	0.0016	0.0022	0.0027	0.0032	0.0035	0.0036	0.0039	0.0043
-0.6	0.0000	0.0008	0.0022	0.0044	0.0063	0.0079	0.0089	0.0096	0.0100	0.0104	0.0110
-0.5	0.0000	0.0016	0.0046	0.0084	0.0119	0.0149	0.0166	0.0177	0.0183	0.0190	0.0196
-0.4	0.0000	0.0030	0.0076	0.0137	0.0191	0.0231	0.0253	0.0268	0.0276	0.0286	0.0292
-0.3	0.0000	0.0043	0.0112	0.0195	0.0264	0.0316	0.0347	0.0363	0.0373	0.0383	0.0390
-0.2	0.0000	0.0056	0.0147	0.0253	0.0339	0.0402	0.0439	0.0456	0.0467	0.0480	0.0486
-0.1	0.0000	0.0071	0.0181	0.0317	0.0411	0.0488	0.0528	0.0548	0.0561	0.0571	0.0579
0	0.0000	0.0094	0.0222	0.0366	0.0482	0.0568	0.0613	0.0640	0.0652	0.0661	0.0668
0.1	0.0000	0.0110	0.0262	0.0436	0.0564	0.0651	0.0696	0.0721	0.0734	0.0746	0.0752
0.2	0.0000	0.0125	0.0300	0.0498	0.0632	0.0725	0.0773	0.0799	0.0813	0.0824	0.0831
0.3	0.0000	0.0146	0.0335	0.0551	0.0697	0.0794	0.0847	0.0875	0.0889	0.0899	0.0906
0.4	0.0000	0.0165	0.0376	0.0605	0.0758	0.0863	0.0918	0.0946	0.0958	0.0970	0.0976
0.5	0.0000	0.0187	0.0416	0.0656	0.0814	0.0927	0.0982	0.1015	0.1025	0.1034	0.1041
0.6	0.0000	0.0213	0.0460	0.0708	0.0873	0.0985	0.1045	0.1076	0.1089	0.1098	0.1103
0.7	0.0000	0.0244	0.0513	0.0762	0.0927	0.1042	0.1104	0.1132	0.1147	0.1157	0.1162
0.8	0.0000	0.0278	0.0567	0.0825	0.0989	0.1097	0.1161	0.1187	0.1202	0.1213	0.1217

Tabela 9 – Cenário 2 - $L(\hat{h}^{(D)})$, $\sigma_3 = 0$: Erro Teórico

ρ_{12}	$n = 2$	$n = 4$	$n = 8$	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$	$n = 512$	$n = 1024$	$\min(L)$
-0.8	0.0888	0.0294	0.0106	0.0048	0.0026	0.0017	0.0013	0.0011	0.0010	0.0009	0.0008
-0.7	0.0962	0.0458	0.0241	0.0140	0.0095	0.0074	0.0064	0.0057	0.0054	0.0052	0.0049
-0.6	0.1050	0.0630	0.0395	0.0265	0.0198	0.0168	0.0149	0.0139	0.0134	0.0130	0.0127
-0.5	0.1144	0.0786	0.0550	0.0398	0.0321	0.0277	0.0255	0.0242	0.0235	0.0231	0.0228
-0.4	0.1221	0.0932	0.0698	0.0534	0.0446	0.0397	0.0369	0.0356	0.0348	0.0344	0.0339
-0.3	0.1300	0.1064	0.0840	0.0663	0.0571	0.0517	0.0487	0.0471	0.0464	0.0459	0.0455
-0.2	0.1371	0.1186	0.0969	0.0789	0.0692	0.0634	0.0604	0.0586	0.0578	0.0574	0.0569
-0.1	0.1437	0.1291	0.1087	0.0907	0.0806	0.0747	0.0715	0.0698	0.0689	0.0685	0.0680
0	0.1496	0.1393	0.1196	0.1018	0.0915	0.0853	0.0823	0.0805	0.0796	0.0791	0.0786
0.1	0.1556	0.1477	0.1297	0.1126	0.1014	0.0954	0.0923	0.0906	0.0896	0.0892	0.0888
0.2	0.1608	0.1550	0.1389	0.1219	0.1110	0.1051	0.1020	0.1002	0.0992	0.0988	0.0984
0.3	0.1650	0.1614	0.1466	0.1309	0.1201	0.1142	0.1111	0.1093	0.1083	0.1079	0.1074
0.4	0.1696	0.1668	0.1539	0.1389	0.1287	0.1228	0.1197	0.1179	0.1169	0.1165	0.1160
0.5	0.1735	0.1714	0.1599	0.1459	0.1365	0.1308	0.1278	0.1260	0.1250	0.1246	0.1241
0.6	0.1770	0.1754	0.1653	0.1524	0.1437	0.1384	0.1355	0.1337	0.1327	0.1323	0.1318
0.7	0.1802	0.1785	0.1697	0.1581	0.1505	0.1455	0.1426	0.1409	0.1399	0.1395	0.1390
0.8	0.1834	0.1806	0.1725	0.1628	0.1562	0.1519	0.1494	0.1478	0.1468	0.1464	0.1459

Tabela 10 – Cenário 2 - $L(\hat{h}^{(D)})$, $\sigma_3 > 0$: Erro Teórico

ρ_{12}	$n = 2$	$n = 4$	$n = 8$	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$	$n = 512$	$n = 1024$	$\min(L)$
-0.8	0.1756	0.0886	0.0314	0.0108	0.0045	0.0024	0.0016	0.0012	0.0010	0.0009	0.0007
-0.7	0.1783	0.1019	0.0479	0.0230	0.0129	0.0087	0.0066	0.0055	0.0051	0.0047	0.0043
-0.6	0.1809	0.1145	0.0641	0.0367	0.0237	0.0176	0.0145	0.0128	0.0120	0.0115	0.0110
-0.5	0.1835	0.1261	0.0792	0.0502	0.0352	0.0279	0.0239	0.0218	0.0208	0.0203	0.0196
-0.4	0.1861	0.1362	0.0929	0.0633	0.0469	0.0384	0.0340	0.0317	0.0305	0.0298	0.0292
-0.3	0.1886	0.1452	0.1051	0.0749	0.0581	0.0487	0.0442	0.0416	0.0404	0.0397	0.0390
-0.2	0.1910	0.1528	0.1162	0.0862	0.0688	0.0590	0.0542	0.0514	0.0501	0.0493	0.0486
-0.1	0.1936	0.1600	0.1261	0.0965	0.0786	0.0687	0.0637	0.0607	0.0594	0.0587	0.0579
0	0.1955	0.1668	0.1363	0.1061	0.0875	0.0775	0.0724	0.0698	0.0682	0.0675	0.0668
0.1	0.1986	0.1738	0.1443	0.1151	0.0963	0.0859	0.0810	0.0780	0.0765	0.0759	0.0752
0.2	0.2007	0.1782	0.1515	0.1233	0.1043	0.0940	0.0889	0.0860	0.0845	0.0838	0.0831
0.3	0.2031	0.1832	0.1580	0.1302	0.1118	0.1015	0.0965	0.0934	0.0920	0.0913	0.0906
0.4	0.2051	0.1876	0.1640	0.1366	0.1189	0.1085	0.1035	0.1004	0.0990	0.0983	0.0976
0.5	0.2071	0.1912	0.1691	0.1427	0.1253	0.1152	0.1100	0.1070	0.1056	0.1049	0.1041
0.6	0.2090	0.1952	0.1736	0.1481	0.1310	0.1213	0.1162	0.1132	0.1118	0.1111	0.1103
0.7	0.2108	0.1985	0.1777	0.1530	0.1363	0.1270	0.1220	0.1191	0.1176	0.1169	0.1162
0.8	0.2126	0.2016	0.1812	0.1568	0.1410	0.1322	0.1273	0.1246	0.1231	0.1224	0.1217

Tabela 11 – Cenário 2 - $\hat{L}(\hat{h}^{(D)}, D')$, $\sigma_3 = 0$

ρ_{12}	$n = 2$	$n = 4$	$n = 8$	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$	$n = 512$	$n = 1024$	$\min(L)$
-0.8	0.0970	0.0297	0.0106	0.0048	0.0026	0.0016	0.0013	0.0011	0.0009	0.0009	0.0008
-0.7	0.1055	0.0470	0.0241	0.0140	0.0095	0.0074	0.0063	0.0057	0.0053	0.0051	0.0049
-0.6	0.1153	0.0642	0.0396	0.0262	0.0199	0.0167	0.0149	0.0139	0.0133	0.0130	0.0127
-0.5	0.1244	0.0808	0.0551	0.0398	0.0319	0.0277	0.0254	0.0243	0.0236	0.0232	0.0228
-0.4	0.1336	0.0959	0.0700	0.0533	0.0446	0.0396	0.0369	0.0356	0.0348	0.0345	0.0339
-0.3	0.1430	0.1099	0.0841	0.0664	0.0570	0.0517	0.0487	0.0472	0.0463	0.0460	0.0455
-0.2	0.1528	0.1228	0.0971	0.0789	0.0692	0.0634	0.0602	0.0586	0.0579	0.0573	0.0569
-0.1	0.1629	0.1346	0.1092	0.0908	0.0806	0.0745	0.0715	0.0699	0.0689	0.0686	0.0680
0	0.1721	0.1455	0.1204	0.1022	0.0914	0.0854	0.0822	0.0806	0.0796	0.0790	0.0786
0.1	0.1806	0.1548	0.1311	0.1126	0.1018	0.0957	0.0923	0.0906	0.0896	0.0894	0.0888
0.2	0.1892	0.1636	0.1395	0.1222	0.1112	0.1053	0.1021	0.1002	0.0992	0.0988	0.0984
0.3	0.1975	0.1715	0.1480	0.1310	0.1205	0.1145	0.1112	0.1094	0.1084	0.1082	0.1074
0.4	0.2058	0.1784	0.1548	0.1388	0.1291	0.1229	0.1199	0.1179	0.1170	0.1168	0.1160
0.5	0.2134	0.1844	0.1613	0.1464	0.1369	0.1309	0.1279	0.1261	0.1252	0.1250	0.1241
0.6	0.2216	0.1895	0.1669	0.1527	0.1440	0.1386	0.1355	0.1336	0.1329	0.1327	0.1318
0.7	0.2296	0.1936	0.1717	0.1584	0.1507	0.1456	0.1427	0.1410	0.1403	0.1399	0.1390
0.8	0.2372	0.1969	0.1747	0.1631	0.1565	0.1521	0.1495	0.1478	0.1472	0.1468	0.1459

Tabela 12 – Cenário 2 - $\hat{L}(\hat{h}^{(D)}, D')$, $\sigma_3 > 0$

ρ_{12}	$n = 2$	$n = 4$	$n = 8$	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$	$n = 512$	$n = 1024$	$\min(L)$
-0.8	0.2074	0.0906	0.0313	0.0107	0.0045	0.0024	0.0016	0.0011	0.0009	0.0008	0.0007
-0.7	0.2110	0.1059	0.0481	0.0229	0.0128	0.0086	0.0066	0.0055	0.0050	0.0046	0.0043
-0.6	0.2148	0.1190	0.0642	0.0366	0.0236	0.0175	0.0143	0.0127	0.0120	0.0115	0.0110
-0.5	0.2185	0.1315	0.0794	0.0502	0.0351	0.0279	0.0238	0.0218	0.0209	0.0204	0.0196
-0.4	0.2223	0.1425	0.0933	0.0630	0.0468	0.0383	0.0339	0.0316	0.0304	0.0298	0.0292
-0.3	0.2260	0.1526	0.1056	0.0749	0.0582	0.0487	0.0441	0.0416	0.0403	0.0397	0.0390
-0.2	0.2298	0.1617	0.1168	0.0862	0.0689	0.0590	0.0540	0.0514	0.0500	0.0494	0.0486
-0.1	0.2333	0.1700	0.1271	0.0965	0.0785	0.0686	0.0634	0.0608	0.0594	0.0587	0.0579
0	0.2371	0.1779	0.1371	0.1064	0.0876	0.0775	0.0724	0.0697	0.0682	0.0674	0.0668
0.1	0.2416	0.1852	0.1460	0.1154	0.0966	0.0861	0.0811	0.0780	0.0765	0.0758	0.0752
0.2	0.2454	0.1918	0.1536	0.1232	0.1046	0.0942	0.0891	0.0861	0.0845	0.0839	0.0831
0.3	0.2480	0.1980	0.1604	0.1306	0.1120	0.1017	0.0966	0.0935	0.0921	0.0914	0.0906
0.4	0.2516	0.2043	0.1664	0.1371	0.1189	0.1087	0.1036	0.1005	0.0990	0.0984	0.0976
0.5	0.2561	0.2091	0.1719	0.1432	0.1256	0.1153	0.1102	0.1071	0.1058	0.1050	0.1041
0.6	0.2601	0.2142	0.1767	0.1483	0.1314	0.1215	0.1163	0.1134	0.1120	0.1112	0.1103
0.7	0.2639	0.2189	0.1810	0.1533	0.1368	0.1272	0.1221	0.1192	0.1178	0.1171	0.1162
0.8	0.2676	0.2234	0.1845	0.1574	0.1413	0.1324	0.1275	0.1246	0.1232	0.1226	0.1217

Tabela 13 – Cenário 3 - indicadores

σ_1	σ_2	σ_3	ρ_{12}	ρ_{13}	ρ_{23}	$BR_2 - BR_3$	$\frac{BR_2}{BR_3}$	$\frac{d_2}{d_3}$	$d_2 - d_3$	$\log_2(n*)$	
										theoretical	empirical, D'
1	1	2	0	-0.7	-0.7	0.07865		9.54895	0.89527	0	0
1	1	2	0	-0.6	-0.6	0.07843	351.28064	2.48236	0.59715	1.001657869	1.033789344
1	1	2	0	-0.5	-0.5	0.07326	14.58169	1.80268	0.44527	1.141669611	1.17710099
1	1	2	0	-0.4	-0.4	0.06156	4.60105	1.49762	0.33227	1.355877509	1.408962482
1	1	2	0	-0.3	-0.3	0.04751	2.52569	1.31823	0.24141	1.667955295	1.706679823
1	1	2	0	-0.2	-0.2	0.03382	1.75436	1.20010	0.16674	2.224947163	2.228173176
1	1	2	0	-0.1	-0.1	0.02173	1.38168	1.11803	0.10557	3.298873181	3.293492842
1	1	2	0	0	0	0.01184	1.17726	1.06068	0.05720	4.515332922	4.530928467
1	1	2	0	0.1	0.1	0.00461	1.06220	1.02271	0.02220	5.911538892	5.972096318
1	1	2	0	0.2	0.2	0.00056	1.00719	1.00269	0.00269	8.931092511	9.024048168
1	1	2	0	0.3	0.3	0.00063	1.00807	1.00305	0.00304	8.864246742	8.813441414
1	1	2	0	0.4	0.4	0.00654	1.09070	1.03256	0.03154	5.586478316	5.582435019
1	1	2	0	0.5	0.5	0.02173	1.38168	1.11803	0.10557	3.815312433	3.848268197
1	1	2	0	0.6	0.6	0.05225	2.97873	1.36929	0.26969	2.703825906	2.747037919
1	1	2	0	0.7	0.7	0.07865	2225445301	4.60952	0.78305	1.924144439	1.967040628

Tabela 14 – Cenário 3 - $\hat{L}(\hat{h}^{(D)})$: Erro Empírico de Treino

$\rho_{13} = \rho_{23}$	$n = 2$	$n = 4$	$n = 8$	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$	$n = 512$	$n = 1024$	$\min(L)$
2 atributos	0	0.0295	0.0478	0.0593	0.0669	0.0723	0.0751	0.0769	0.0774	0.0780	0.0786
-0.7	0	0	0	0	0	0	0	0	0	0	0
-0.6	0	0	0	0.0000	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002	0.0002
-0.5	0	0.0003	0.0007	0.0017	0.0025	0.0033	0.0038	0.0043	0.0046	0.0049	0.0054
-0.4	0	0.0013	0.0031	0.0063	0.0097	0.0123	0.0140	0.0152	0.0158	0.0165	0.0171
-0.3	0	0.0028	0.0075	0.0139	0.0198	0.0242	0.0269	0.0287	0.0295	0.0304	0.0311
-0.2	0	0.0049	0.0134	0.0225	0.0306	0.0365	0.0403	0.0419	0.0429	0.0442	0.0448
-0.1	0	0.0067	0.0179	0.0305	0.0401	0.0480	0.0518	0.0539	0.0551	0.0561	0.0569
0	0	0.0094	0.0222	0.0366	0.0482	0.0568	0.0613	0.0640	0.0652	0.0661	0.0668
0.1	0	0.0108	0.0261	0.0423	0.0544	0.0636	0.0682	0.0709	0.0726	0.0735	0.0740
0.2	0	0.0120	0.0280	0.0462	0.0584	0.0679	0.0723	0.0749	0.0765	0.0774	0.0781
0.3	0	0.0125	0.0296	0.0472	0.0590	0.0679	0.0722	0.0749	0.0765	0.0775	0.0780
0.4	0	0.0123	0.0284	0.0439	0.0545	0.0628	0.0668	0.0692	0.0708	0.0716	0.0721
0.5	0	0.0106	0.0238	0.0349	0.0430	0.0491	0.0523	0.0541	0.0554	0.0563	0.0569
0.6	0	0.0060	0.0129	0.0171	0.0204	0.0226	0.0237	0.0244	0.0252	0.0258	0.0264
0.7	0	0	0	0	0	0	0	0	0	0	0

Tabela 15 – Cenário 3 - $L(\hat{h}^{(D)})$: Erro Teórico

$\rho_{13} = \rho_{23}$	$n = 2$	$n = 4$	$n = 8$	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$	$n = 512$	$n = 1024$	$\min(L)$
2 atributos	0.1496	0.1393	0.1196	0.1018	0.0915	0.0853	0.0823	0.0805	0.0796	0.0791	0.0786
-0.7	0.1419	0.0270	0.0011	0	0	0	0	0	0	0	0
-0.6	0.1504	0.0520	0.0136	0.0045	0.0019	0.0011	0.0007	0.0005	0.0004	0.0003	0.00022
-0.5	0.1602	0.0801	0.0391	0.0225	0.0139	0.0101	0.0080	0.0068	0.0063	0.0059	0.00539
-0.4	0.1694	0.1049	0.0659	0.0438	0.0312	0.0247	0.0212	0.0192	0.0183	0.0177	0.01709
-0.3	0.1774	0.1259	0.0886	0.0635	0.0486	0.0403	0.0359	0.0337	0.0325	0.0318	0.03114
-0.2	0.1844	0.1429	0.1072	0.0802	0.0642	0.0549	0.0502	0.0475	0.0463	0.0455	0.04483
-0.1	0.1905	0.1563	0.1228	0.0948	0.0774	0.0676	0.0626	0.0598	0.0584	0.0576	0.05692
0	0.1955	0.1668	0.1363	0.1061	0.0875	0.0775	0.0724	0.0698	0.0682	0.0675	0.06681
0.1	0.2009	0.1754	0.1444	0.1142	0.0961	0.0852	0.0801	0.0769	0.0755	0.0748	0.07404
0.2	0.2050	0.1812	0.1497	0.1187	0.1005	0.0893	0.0841	0.0810	0.0795	0.0788	0.07809
0.3	0.2084	0.1849	0.1515	0.1195	0.1005	0.0893	0.0840	0.0809	0.0795	0.0788	0.07802
0.4	0.2112	0.1855	0.1480	0.1140	0.0947	0.0833	0.0780	0.0750	0.0735	0.0729	0.07211
0.5	0.2133	0.1817	0.1362	0.0984	0.0787	0.0677	0.0627	0.0596	0.0583	0.0576	0.05692
0.6	0.2150	0.1691	0.1076	0.0640	0.0446	0.0351	0.0310	0.0287	0.0276	0.0270	0.02640
0.7	0.2159	0.1337	0.0449	0.0050	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0

Tabela 16 – Cenário 3 - $\hat{L}(\hat{h}^{(D)}, D')$: Erro Empírico de Teste

$\rho_{13} = \rho_{23}$	$n = 2$	$n = 4$	$n = 8$	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$	$n = 512$	$n = 1024$	$\min(L)$
2 atributos	0.1721	0.1455	0.1204	0.1022	0.0914	0.0854	0.0822	0.0806	0.0796	0.0790	0.0786
-0.7	0.1665	0.0279	0.0011	0	0	0	0	0	0	0	0
-0.6	0.1761	0.0534	0.0136	0.0045	0.0019	0.0011	0.0007	0.0005	0.0003	0.0003	0.0002
-0.5	0.1870	0.0818	0.0391	0.0224	0.0139	0.0101	0.0079	0.0068	0.0062	0.0057	0.0054
-0.4	0.1991	0.1089	0.0658	0.0438	0.0309	0.0246	0.0211	0.0193	0.0184	0.0177	0.0171
-0.3	0.2098	0.1313	0.0887	0.0636	0.0487	0.0403	0.0359	0.0337	0.0324	0.0318	0.0311
-0.2	0.2196	0.1500	0.1080	0.0801	0.0642	0.0550	0.0500	0.0475	0.0464	0.0455	0.0448
-0.1	0.2273	0.1646	0.1237	0.0944	0.0775	0.0675	0.0625	0.0597	0.0583	0.0577	0.0569
0	0.2371	0.1779	0.1371	0.1064	0.0876	0.0775	0.0724	0.0697	0.0682	0.0674	0.0668
0.1	0.2440	0.1885	0.1459	0.1141	0.0959	0.0852	0.0800	0.0770	0.0754	0.0748	0.0740
0.2	0.2527	0.1956	0.1518	0.1188	0.1004	0.0894	0.0840	0.0810	0.0796	0.0788	0.0781
0.3	0.2584	0.1993	0.1533	0.1194	0.1005	0.0891	0.0838	0.0809	0.0793	0.0788	0.0780
0.4	0.2653	0.2003	0.1496	0.1142	0.0945	0.0831	0.0779	0.0751	0.0735	0.0729	0.0721
0.5	0.2701	0.1956	0.1374	0.0988	0.0787	0.0675	0.0626	0.0596	0.0583	0.0576	0.0569
0.6	0.2749	0.1813	0.1084	0.0642	0.0445	0.0351	0.0308	0.0287	0.0275	0.0270	0.0264
0.7	0.2796	0.1429	0.0451	0.0051	0.0002	0.0000	0.0000	0	0	0	0

Tabela 17 – Cenário 4 - indicadores

σ_1	σ_2	σ_3	ρ_{12}	ρ_{13}	ρ_{23}	$BR_2 - BR_3$	$\frac{BR_2}{BR_3}$	$\frac{d_2}{d_3}$	$d_2 - d_3$	$\log_2(n*)$	
										theoretical	empirical, D'
1	1	2	0	-0.7	-0.7	0.07865		9.54895	0.89527	0	0
1	1	2	0	-0.6	-0.6	0.07843	351.28064	2.48236	0.59715	1.001657869	1.033789344
1	1	2	0	-0.5	-0.5	0.07326	14.58169	1.80268	0.44527	1.141669611	1.17710099
1	1	2	0	-0.4	-0.4	0.06156	4.60105	1.49762	0.33227	1.355877509	1.408962482
1	1	2	0	-0.3	-0.3	0.04751	2.52569	1.31823	0.24141	1.667955295	1.706679823
1	1	2	0	-0.2	-0.2	0.03382	1.75436	1.20010	0.16674	2.224947163	2.228173176
1	1	2	0	-0.1	-0.1	0.02173	1.38168	1.11803	0.10557	3.298873181	3.293492842
1	1	2	0	0	0	0.01184	1.17726	1.06068	0.05720	4.515332922	4.530928467
1	1	2	0	0.1	0.1	0.00461	1.06220	1.02271	0.02220	5.911538892	5.972096318
1	1	2	0	0.2	0.2	0.00056	1.00719	1.00269	0.00269	8.931092511	9.024048168
1	1	2	0	0.3	0.3	0.00063	1.00807	1.00305	0.00304	8.864246742	8.813441414
1	1	2	0	0.4	0.4	0.00654	1.09070	1.03256	0.03154	5.586478316	5.582435019
1	1	2	0	0.5	0.5	0.02173	1.38168	1.11803	0.10557	3.815312433	3.848268197
1	1	2	0	0.6	0.6	0.05225	2.97873	1.36929	0.26969	2.703825906	2.747037919
1	1	2	0	0.7	0.7	0.07865	2225445301	4.60952	0.78305	1.924144439	1.967040628

Tabela 18 – Cenário 4 - $\hat{L}(\hat{h}^{(D)})$: Erro Empírico de Treino

Tabela 19 – Cenário 4 - $L(\hat{h}^{(D)})$: Erro Teórico

$\rho_{13} = \rho_{23}$	$n = 2$	$n = 4$	$n = 8$	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$	$n = 512$	$n = 1024$	$\min(L)$
2 atributos	0.1496	0.1393	0.1196	0.1018	0.0915	0.0853	0.0823	0.0805	0.0796	0.0791	0.0786
-0.7	0.1419	0.0270	0.0011	0	0	0	0	0	0	0	0
-0.6	0.1504	0.0520	0.0136	0.0045	0.0019	0.0011	0.0007	0.0005	0.0004	0.0003	0.00022
-0.5	0.1602	0.0801	0.0391	0.0225	0.0139	0.0101	0.0080	0.0068	0.0063	0.0059	0.00539
-0.4	0.1694	0.1049	0.0659	0.0438	0.0312	0.0247	0.0212	0.0192	0.0183	0.0177	0.01709
-0.3	0.1774	0.1259	0.0886	0.0635	0.0486	0.0403	0.0359	0.0337	0.0325	0.0318	0.03114
-0.2	0.1844	0.1429	0.1072	0.0802	0.0642	0.0549	0.0502	0.0475	0.0463	0.0455	0.04483
-0.1	0.1905	0.1563	0.1228	0.0948	0.0774	0.0676	0.0626	0.0598	0.0584	0.0576	0.05692
0	0.1955	0.1668	0.1363	0.1061	0.0875	0.0775	0.0724	0.0698	0.0682	0.0675	0.06681
0.1	0.2009	0.1754	0.1444	0.1142	0.0961	0.0852	0.0801	0.0769	0.0755	0.0748	0.07404
0.2	0.2050	0.1812	0.1497	0.1187	0.1005	0.0893	0.0841	0.0810	0.0795	0.0788	0.07809
0.3	0.2084	0.1849	0.1515	0.1195	0.1005	0.0893	0.0840	0.0809	0.0795	0.0788	0.07802
0.4	0.2112	0.1855	0.1480	0.1140	0.0947	0.0833	0.0780	0.0750	0.0735	0.0729	0.07211
0.5	0.2133	0.1817	0.1362	0.0984	0.0787	0.0677	0.0627	0.0596	0.0583	0.0576	0.05692
0.6	0.2150	0.1691	0.1076	0.0640	0.0446	0.0351	0.0310	0.0287	0.0276	0.0270	0.02640
0.7	0.2159	0.1337	0.0449	0.0050	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0

Tabela 20 – Cenário 4 - $\hat{L}(\hat{h}^{(D)}, D')$: Erro Empírico de Teste

$\rho_{13} = \rho_{23}$	$n = 2$	$n = 4$	$n = 8$	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$	$n = 512$	$n = 1024$	$\min(L)$
2 atributos	0.1721	0.1455	0.1204	0.1022	0.0914	0.0854	0.0822	0.0806	0.0796	0.0790	0.0786
-0.7	0.1665	0.0279	0.0011	0	0	0	0	0	0	0	0
-0.6	0.1761	0.0534	0.0136	0.0045	0.0019	0.0011	0.0007	0.0005	0.0003	0.0003	0.0002
-0.5	0.1870	0.0818	0.0391	0.0224	0.0139	0.0101	0.0079	0.0068	0.0062	0.0057	0.0054
-0.4	0.1991	0.1089	0.0658	0.0438	0.0309	0.0246	0.0211	0.0193	0.0184	0.0177	0.0171
-0.3	0.2098	0.1313	0.0887	0.0636	0.0487	0.0403	0.0359	0.0337	0.0324	0.0318	0.0311
-0.2	0.2196	0.1500	0.1080	0.0801	0.0642	0.0550	0.0500	0.0475	0.0464	0.0455	0.0448
-0.1	0.2273	0.1646	0.1237	0.0944	0.0775	0.0675	0.0625	0.0597	0.0583	0.0577	0.0569
0	0.2371	0.1779	0.1371	0.1064	0.0876	0.0775	0.0724	0.0697	0.0682	0.0674	0.0668
0.1	0.2440	0.1885	0.1459	0.1141	0.0959	0.0852	0.0800	0.0770	0.0754	0.0748	0.0740
0.2	0.2527	0.1956	0.1518	0.1188	0.1004	0.0894	0.0840	0.0810	0.0796	0.0788	0.0781
0.3	0.2584	0.1993	0.1533	0.1194	0.1005	0.0891	0.0838	0.0809	0.0793	0.0788	0.0780
0.4	0.2653	0.2003	0.1496	0.1142	0.0945	0.0831	0.0779	0.0751	0.0735	0.0729	0.0721
0.5	0.2701	0.1956	0.1374	0.0988	0.0787	0.0675	0.0626	0.0596	0.0583	0.0576	0.0569
0.6	0.2749	0.1813	0.1084	0.0642	0.0445	0.0351	0.0308	0.0287	0.0275	0.0270	0.0264
0.7	0.2796	0.1429	0.0451	0.0051	0.0002	0.0000	0.0000	0	0	0	0

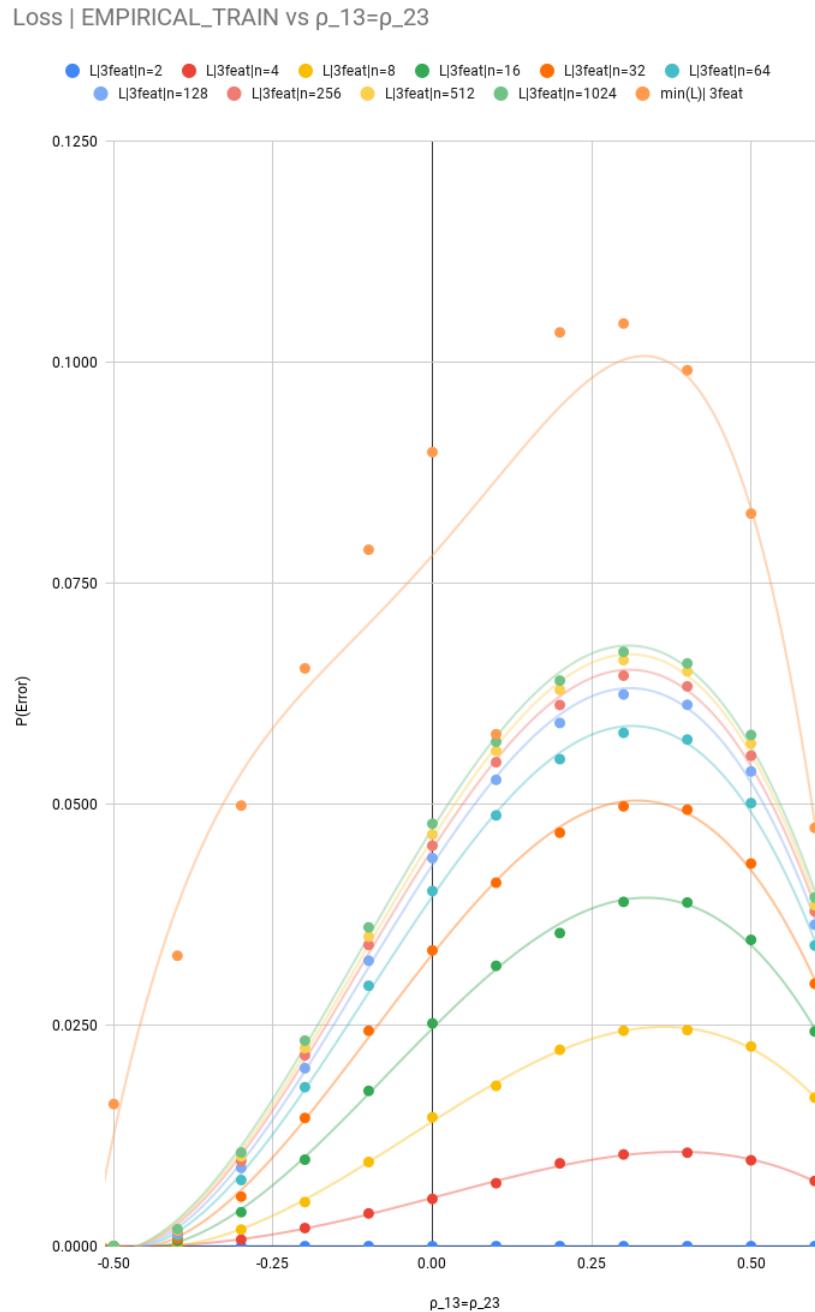


Figura 38 – Cenário 4 - $\hat{L}(\hat{h}^{(D)})$ vs $\rho_{13} = \rho_{23}$

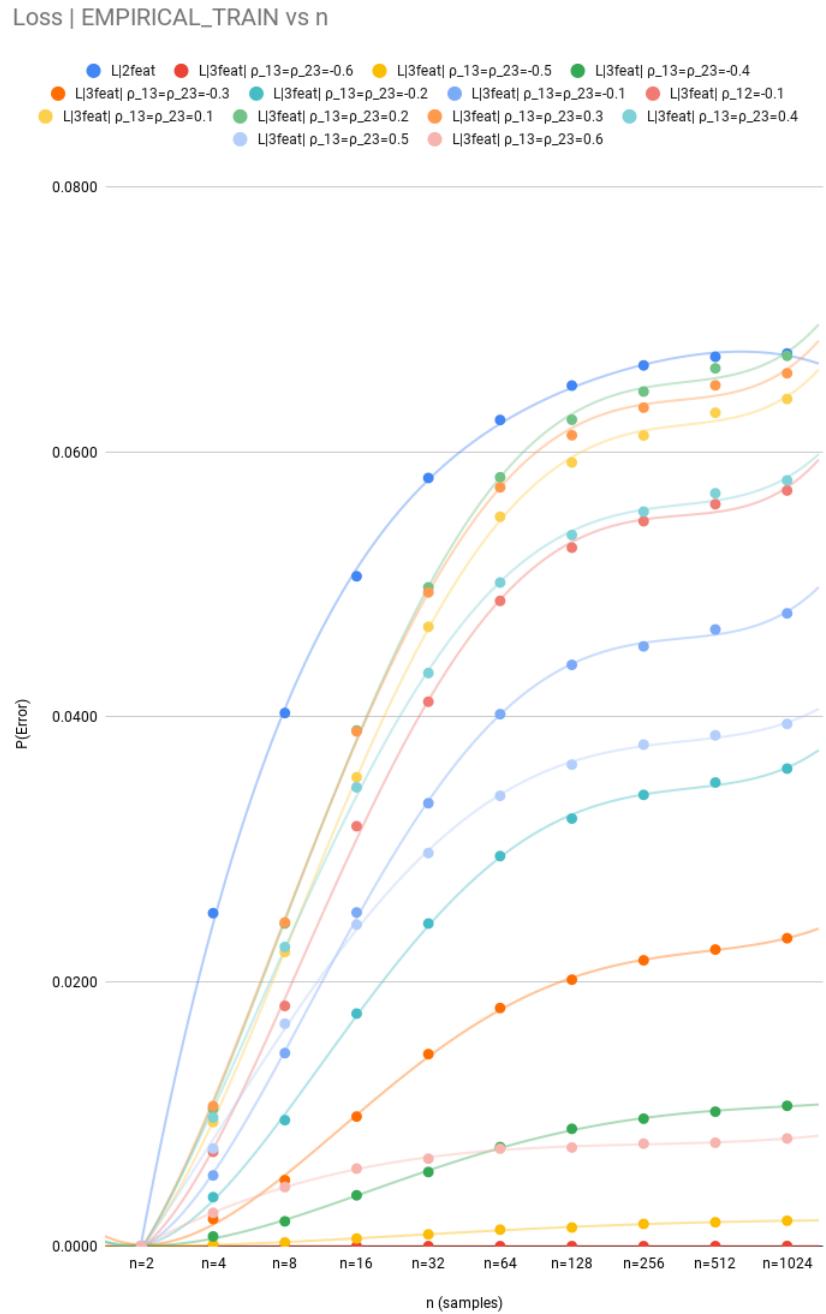
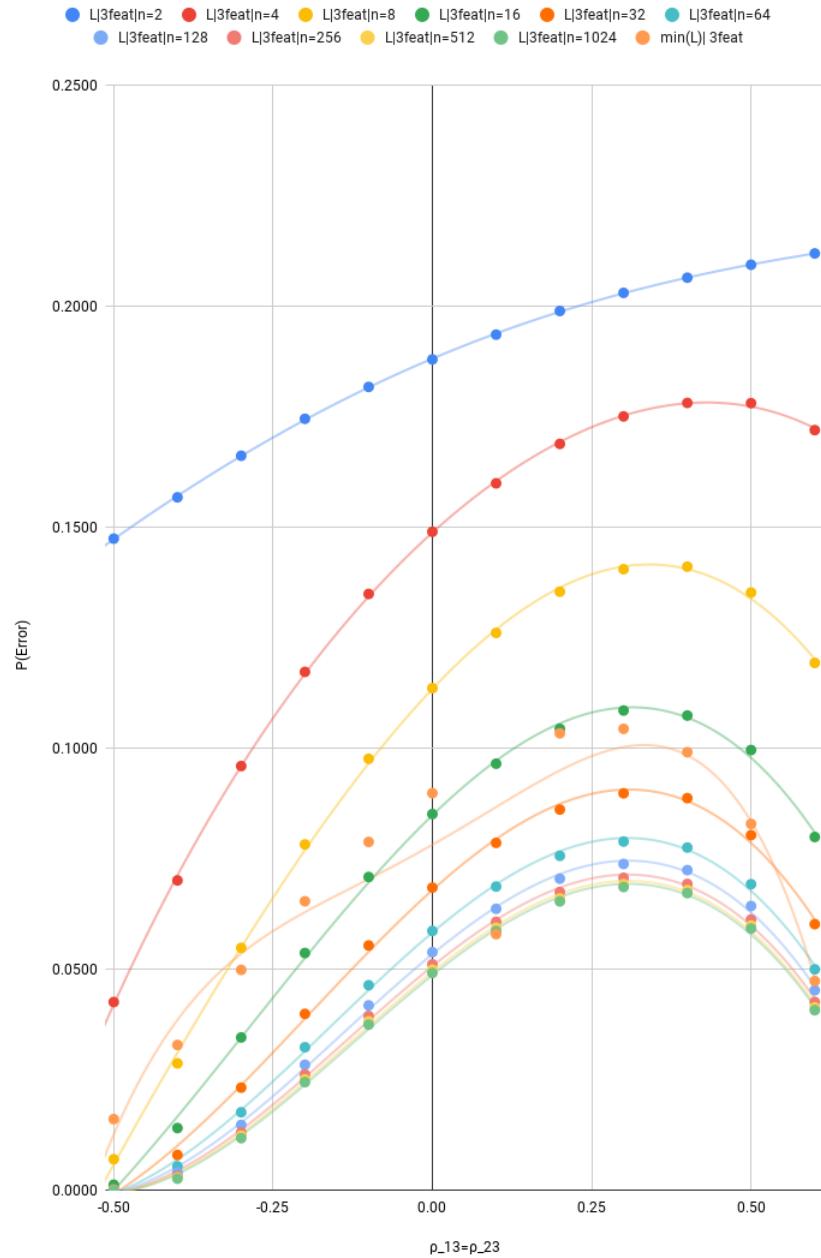


Figura 39 – Cenário 4 - $\hat{L}(\hat{h}^{(D)})$ vs n

Loss | THEORETICAL vs $\rho_{13}=\rho_{23}$ Figura 40 – Cenário 4 - $L(\hat{h}^{(D)})$ vs $\rho_{13} = \rho_{23}$

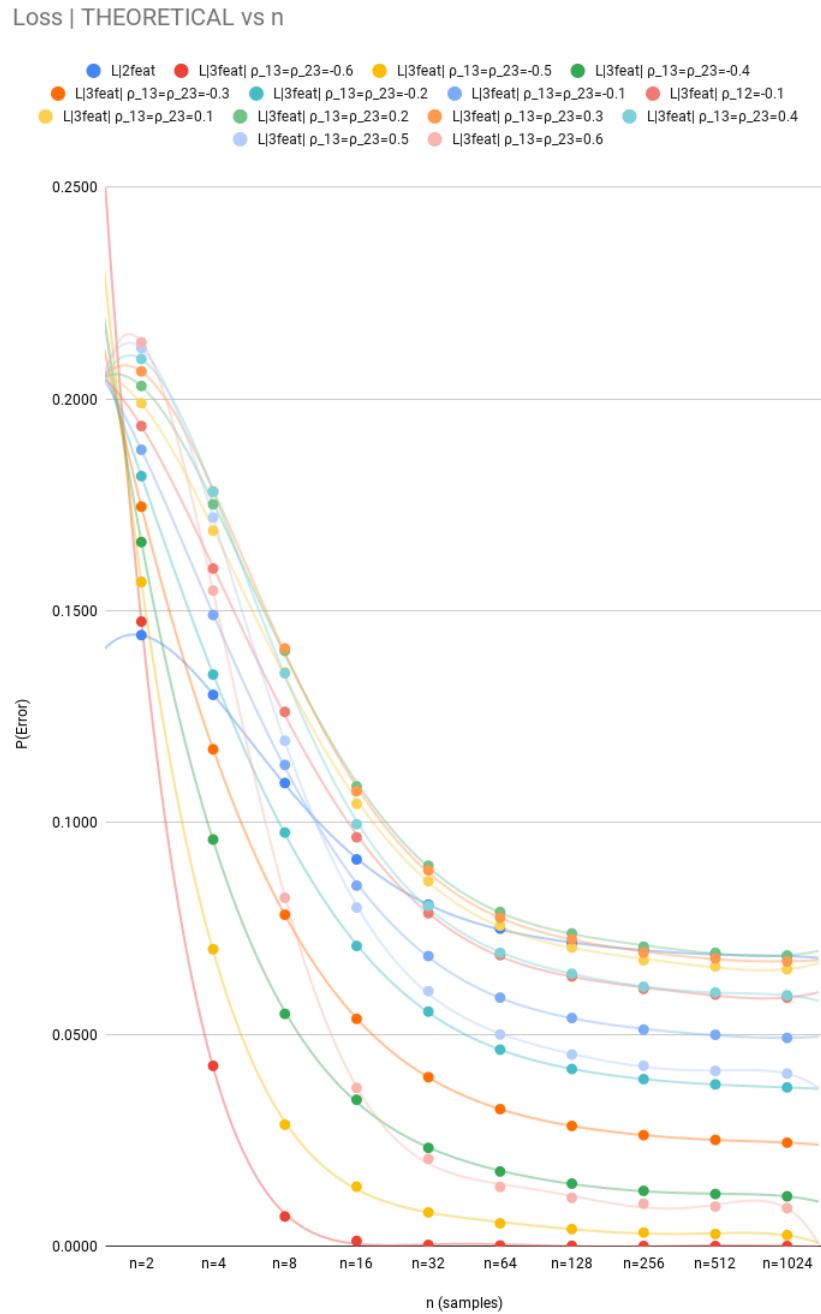


Figura 41 – Cenário 4 - $L(\hat{h}^{(D)})$ vs n

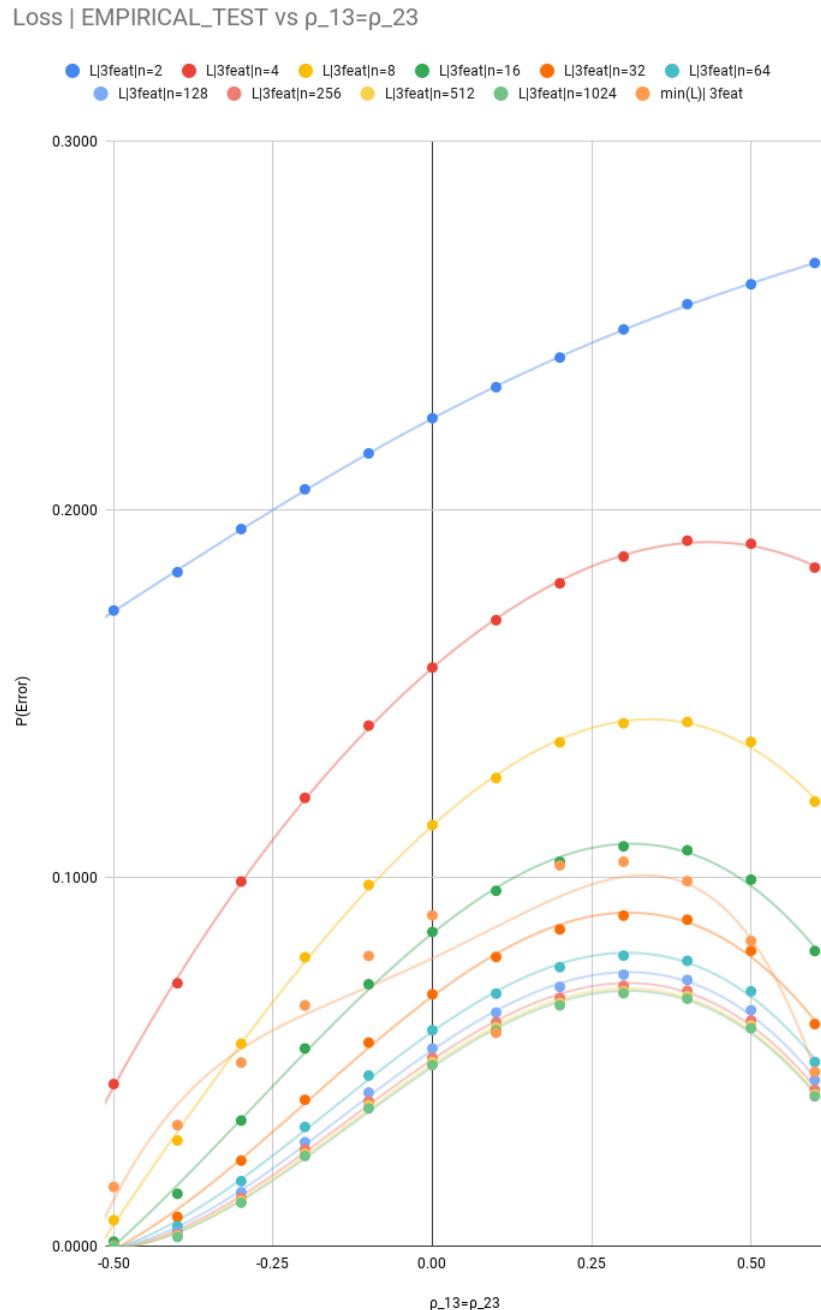


Figura 42 – Cenário 4 - $\hat{L}(\hat{h}^{(D)}, D')$ vs $\rho_{13} = \rho_{23}$

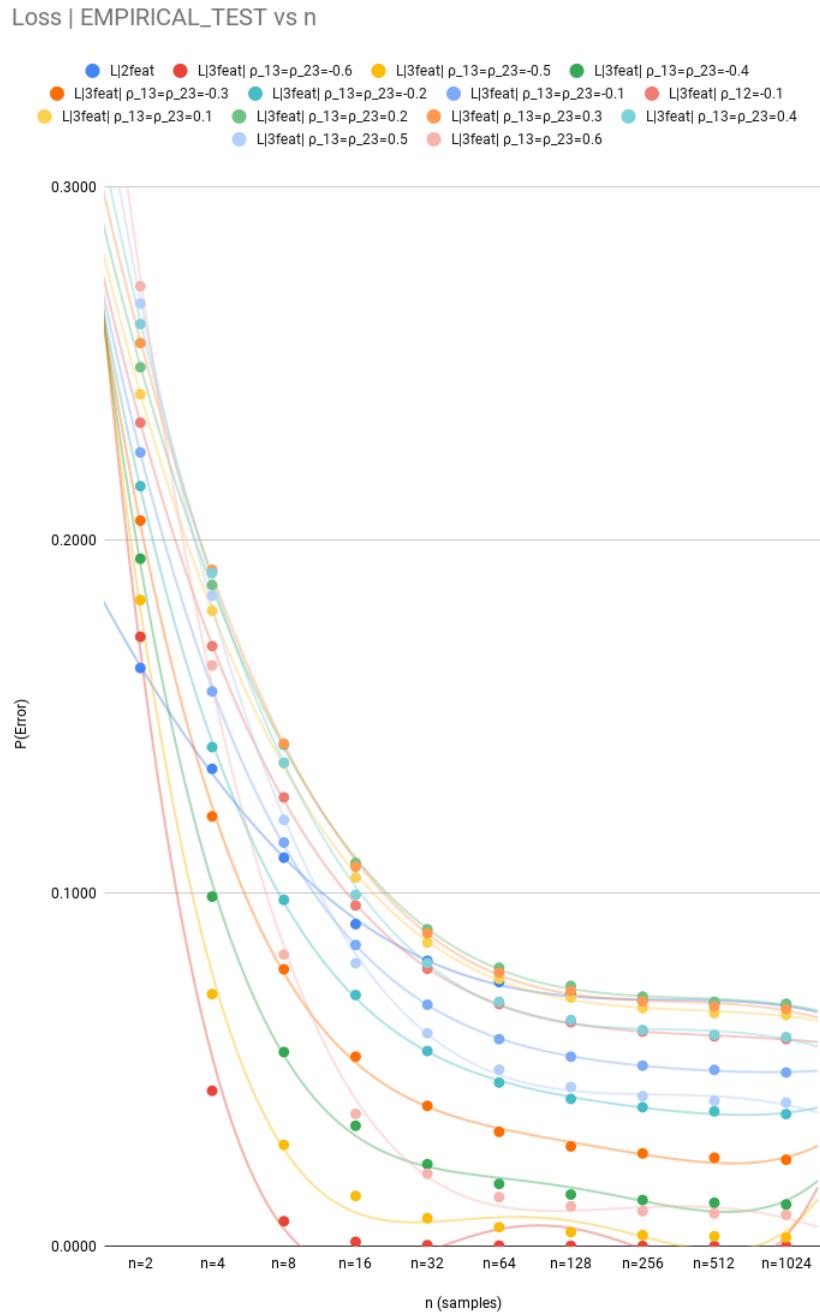


Figura 43 – Cenário 4 - $\hat{L}(\hat{h}^{(D)}, D')$ vs n

5 TRABALHOS RELACIONADOS

Os modelos analíticos e de simulação deste trabalho foram baseados principalmente nos estudos do professor João I. Pinheiro como "PINHEIRO, J.I.; Aprendendo a classificar com poucos atributos e observações – Tese de doutorado no PPGI/UFRJ,2021" entre outros materiais disponibilizados. (??)

Treinar classificadores com poucas amostras é um tema que recentemente tem recebido a atenção da comunidade de sensoriamento remoto (ZHANG et al., 2021), (HE et al., 2021) entre outros (HANCZAR; DOUGHERTY, 2013; MILLER; MATSAKIS; VIOLA, 2000; WANG et al., 2020). Considere um sistema que implementa um algoritmo para a detecção de ponto de mudança visando rastrear abruptas novidades no meio ambiente. Uma vez que ocorre um ponto de mudança, podemos ter algumas amostras de cada uma das classes que representam nosso alvo de interesse. O treinamento de um classificador, imediatamente após um tal ponto de mudança, fica então restrito a pequenos tamanhos amostrais e a um número possivelmente limitado de atributos (NOH; RAJAGOPAL; KIREMIDJIAN, 2013). Embora já tenha sido feito um esforço significativo para se investigar como o número de amostras e o número de atributos afetam a precisão do classificador (HANCZAR; DOUGHERTY, 2013), (NOH; RAJAGOPAL; KIREMIDJIAN, 2013), a relação entre esses dois atributos ainda apresenta problemas analíticos e práticos em aberto.

Semelhante ao que fazemos aqui, em (HUA et al., 2005) os autores também tratam de Classificação Binária por meio de uma abordagem baseada em simulação.

Nossas simulações abordam exemplos específicos, e consideramos o que acontece com a probabilidade de erro à medida que a cardinalidade aumenta, mantendo constante o número de atributos.

Por outro lado, em (HUA et al., 2005) os autores consideram uma combinação muito mais ampla de modelos probabilísticos e procedimentos estatísticos. Eles mostram que, em condições adequadas, para cada cardinalidade do conjunto de dados n , existe um número ótimo de atributos a serem usadas no classificador, a fim de minimizar sua probabilidade de classificação incorreta.

A propósito, em (HUGHES, 1968), ao abordar o caso específico de atributos discretos, o autor conclui algo muito semelhante ao resultado de (HUA et al., 2005). Em vez de considerar a dimensionalidade d , em (HUGHES, 1968) se introduz o conceito de “complexidade do padrão de medição” c , a saber, o número total de valores possíveis do vetor de atributos. O autor afirma que, para cada cardinalidade n do conjunto de dados, existe um c ideal, para a qual a probabilidade de erro do classificador é mínima. Assim, trocando d por c , as conclusões de (HUGHES, 1968) e (HUA et al., 2005) são análogas entre si.

Aliás, tanto nosso trabalho como quanto (HUA et al., 2005) concordam que, devido

à ocorrência de *overfitting*, se a cardinalidade do conjunto de dados for relativamente pequena, uma dimensionalidade alta pode prejudicar substancialmente o desempenho esperado do classificador.

Nosso trabalho sugere que, em muitas situações, pode haver um *trade-off* entre amostras e atributos, no que diz respeito à busca de minimização da probabilidade de erro.

Em outras palavras, uma eventual escassez de atributos poderia ser compensada apenas pelo aumento do tamanho amostral. E isso geralmente tende a acontecer quando os novos atributos que poderiam eventualmente ser adicionados, não possuem um poder de discriminação tão alto quanto as já presentes no modelo.

Por outro lado, uma eventual escassez de amostras também poderia ser compensada pela adição de novos atributos ao conjunto de dados. Já isso exigiria a disponibilidade de novos atributos cuja inclusão melhorasse expressivamente o poder de discriminação global do modelo.

6 CONCLUSÃO E TRABALHOS FUTUROS

Neste trabalho, consideramos um problema de classificação binária, envolvendo normais multivariadas. Inicialmente, supondo que $\rho = 0$ (i.e., atributos condicionalmente independentes), analisamos como a cardinalidade n do conjunto de dados, aliada ao valor do parâmetro δ , influenciam simultaneamente a decisão de se usar um ou dois atributos na definição do classificador, tendo em vista que se deseja minimizar a sua probabilidade esperada de erro. Verificamos a existência de um limiar para a cardinalidade n , abaixo da qual é preferível usar apenas o atributo mais discriminativo.

Em seguida, propusemos uma modelagem puramente analítica que representasse com boa precisão o comportamento do fenômeno acima descrito, para uma ampla região de valores possíveis da cardinalidade n .

Exploramos espaços de 1, 2 e 3 dimensões, mas podemos via simulação, com o simulador desenvolvidos, considerar mais dimensões. Pretendemos em trabalhos futuros avaliar se nossas conclusões poderiam também serem estendidas a espaços de dimensão mais elevada. Pretendemos também experimentar o *tradeoff* entre atributos e amostras em um ambiente real, por exemplo, envolvendo uma rede de sensores sem fio em um ambiente de teste.

REFERÊNCIAS

- BLANCHARD, G.; BOUSQUET, O.; MASSART, P. Statistical performance of support vector machines. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 36, n. 2, p. 489–531, 2008.
- BOTTOU, L.; LIN, C.-J. Support vector machine solvers. **Large scale kernel machines**, Citeseer, v. 3, n. 1, p. 301–320, 2007.
- D'ASCOLI, S.; SAGUN, L.; BIROLI, G. Triple descent and the two kinds of overfitting: where & why do they appear? In: **NeurIPS**. [S.l.: s.n.], 2020.
- DEVROYE, L.; GYÖRFI, L.; LUGOSI, G. **A probabilistic theory of pattern recognition**. [S.l.]: Springer Science & Business Media, 2013. v. 31.
- FARAGÓ, A.; LUGOSI, G. Strong universal consistency of neural network classifiers. **IEEE Transactions on Information Theory**, IEEE, v. 39, n. 4, p. 1146–1151, 1993.
- GLASMACHERS, T. Universal consistency of multi-class support vector classification. **Advances in Neural Information Processing Systems**, v. 23, p. 739–747, 2010.
- HANCZAR, B.; DOUGHERTY, E. R. The reliability of estimated confidence intervals for classification error rates when only a single sample is available. **Pattern Recognition**, Elsevier, v. 46, n. 3, p. 1067–1077, 2013.
- HE, F. et al. One-Shot Distributed Algorithm for PCA With RBF Kernels. **IEEE Signal Processing Letters**, IEEE, v. 28, p. 1465–1469, 2021.
- HSIEH, C.-J. et al. A dual coordinate descent method for large-scale linear SVM. In: **Proceedings of the 25th international conference on Machine learning**. [S.l.: s.n.], 2008. p. 408–415.
- HUA, J. et al. Optimal number of features as a function of sample size for various classification rules. **Bioinformatics**, Oxford University Press, v. 21, n. 8, p. 1509–1515, 2005.
- HUGHES, G. On the mean accuracy of statistical pattern recognizers. **IEEE Transactions on Information Theory**, IEEE, v. 14, n. 1, p. 55–63, 1968.
- LIST, N.; SIMON, H. U. SVM-optimization and steepest-descent line search. In: CITESEER. **Proceedings of the 22nd Annual Conference on Computational Learning Theory**. [S.l.], 2009.
- MILLER, E. G.; MATSAKIS, N. E.; VIOLA, P. A. Learning from one example through shared densities on transforms. In: **IEEE. Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)**. [S.l.], 2000. v. 1, p. 464–471.
- NAKKIRAN, P. et al. Optimal regularization can mitigate double descent. In: **International Conference on Learning Representations**. [S.l.: s.n.], 2020.

NOH, H.; RAJAGOPAL, R.; KIREMIDJIAN, A. Sequential structural damage diagnosis algorithm using a change point detection method. **Journal of Sound and Vibration**, Elsevier, v. 332, n. 24, p. 6419–6433, 2013.

R Documentation. 2021. <<https://www.rdocumentation.org/packages/e1071>>.

VERT, R.; VERT, J.-P.; SCHÖLKOPF, B. Consistency and convergence rates of one-class SVMs and related algorithms. **Journal of Machine Learning Research**, v. 7, n. 5, 2006.

WANG, Y. et al. Generalizing from a few examples: A survey on few-shot learning. **ACM Computing Surveys (CSUR)**, ACM New York, NY, USA, v. 53, n. 3, p. 1–34, 2020.

WILLETT, P.; SWASZEK, P. F.; BLUM, R. S. The good, bad and ugly: distributed detection of a known signal in dependent gaussian noise. **IEEE Transactions on signal processing**, IEEE, v. 48, n. 12, p. 3266–3279, 2000.

ZHANG, S. et al. Polygon structure-guided hyperspectral image classification with single sample for strong geometric characteristics scenes. **IEEE Transactions on Geoscience and Remote Sensing**, IEEE, 2021.