ALGORITMOS II
TRABALHO PRÁTICO I

Thaís Ferreira da Silva - 2021092571
Filipe Araújo - 2021031920
Rodrigo Sales Nascimento - 2021067534

Departamento de Ciência da Computação - Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte - MG - Brasil

Introdução:

O objetivo deste trabalho prático é estudar sobre as aplicações da geometria computacional, mais especificamente a utilização de envoltórias convexas e verificação de interseção de segmentos para o problema de classificação em aprendizado de máquina supervisionado. Por isso, o código desenvolvido foi dividido em 6 etapas fundamentais que serão melhor detalhadas abaixo, sendo elas: cálculo da envoltória, varredura linear, modelagem, classificação e experimentos. Além disso, também será apresentada uma breve descrição dos tipos abstratos de dados que foram implementados para a inicialização e funcionamento das demais etapas citadas anteriormente.

Inicialização:

Ao todo, 3 classes foram implementadas para garantir a abstração de dados necessária na definição da envoltória convexa e varredura linear:

- A classe **Point** para armazenar os valores de x e y de cada ponto dos bancos de dados;
- A classe **Segment** armazenar um par de pontos que definem um segmento de reta, e também os valores dos coeficientes a e b que a descrevem;
- A classe **EndPoint** armazena um ponto, sua coordenada em relação ao segmento (endpoint da direita ou da esquerda), e um ponteiro para esse segmento. Ela é essencial para o processo de varredura linear, ao facilitar a inserção e remoção de um segmento na estrutura de dados desenvolvida para o algoritmo;

Determinação da envoltória convexa:

O algoritmo de Graham, dentre os que foram estudados ao longo da disciplina, foi escolhido para determinar as envoltórias convexas dos conjuntos de pontos dos experimentos. Ele é capaz de determinar o conjunto de pontos que compõem a envoltória em tempo **O(n log n)**, e é altamente eficiente quando se trabalha com pontos no plano 2D. Dessa forma, foram implementadas funções auxiliares responsáveis por encontrar o ponto âncora a ser utilizado no algoritmo e calcular orientações em mudanças de trajetória em caminhos de segmentos para ordenação e definição do conjunto solução.

Resumidamente, a ideia é que os vértices são analisados um a um a partir de um âncora no sentido anti-horário. Assim, partindo do ponto mais a esquerda e mais abaixo possível, organizam-se os demais pontos levando em consideração o seu ângulo polar em relação a âncora. Com isso, inicia-se o processo de extensão da envoltória, adicionando pontos em caso de rotação para a esquerda ou direita e removendo pontos em caso de rotações para a direita.

Varredura Linear:

Para o algoritmo de varredura linear, foi necessário construir uma **árvore AVL** que possui complexidade **O(log n)** na inserção, remoção e consulta de pontos dos segmentos em sua estrutura. Utilizando essa estrutura, podemos obter a varredura em um tempo consideravelmente menor em relação a outras implementações com árvore binária. A classe EndPoint foi implementada especialmente para essa parte do trabalho, pois foi facilitadora na criação de uma fila com os pontos a serem verificados na varredura e na manipulação desses na árvore sabendo qual é a posição e a qual segmento o ponto pertence. Além disso, foram utilizadas funções auxiliares responsáveis por verificar se dois segmentos se interceptam (baseadas nas primitivas estudadas em sala), obter a orientação de um ponto e verificar se um ponto está sobre um segmento ou não.

O restante das especificações do algoritmo foi baseado no modelo que foi visto em sala de aula. A varredura linear inicialmente realiza um pré-processamento onde ordena os endpoints dos segmentos obtidos das envoltórias determinadas no passo anterior da esquerda para a direita, e em caso de empate, de baixo para cima. Depois disso, iniciamos o processo da varredura onde para cada ponto do array de endpoints adicionamos um segmento na árvore AVL quando é o ponto inicial e o removemos quando é o final. Seguindo esse funcionamento, quando um ponto final é lindo, avalia-se a sua interseção com os outros segmentos mais próximos.

Determinação do modelo de classificação de pontos:

Conforme a verificação da separabilidade linear das envoltórias, foram implementadas duas funções responsáveis por encontrar os dois pontos mais próximos entre as envoltórias e definir a reta perpendicular que passa pelo ponto médio do segmento formado por esses pontos.

Para encontrar os dois pontos mais próximos, foram utilizados dois dois loops aninhados responsáveis por percorrer todos os pontos das envoltórias e calcular a distância entre eles, atualizando a distância mínima registrada.

Já para gerar a reta perpendicular que divide os pontos, foi criada uma função que recebe os dois pontos mais próximos e o ponto central entre eles. Desta forma, pode-se calcular os valores dos coeficientes a e b que descrevem a reta perpendicular que passa pelo ponto médio do segmento encontrado. Essa reta estabelece os limites de separação entre as envoltórias e será usada como o modelo de classificação de pontos nos experimentos.

Etapas de classificação de pontos:

A técnica de separabilidade com envoltórias convexas é um método de classificação que se baseia na ideia de separar pontos em classes com base em suas envoltórias convexas. O processo geral de classificação de pontos adotado nos experimentos segue os seguintes passos:

- **Coleta de dados:** Encontrar o dataset que apresente um conjunto de dados que podem ser convertidos em pontos e que apresente suas respectivas classes (rótulos).

- **Divisão dos dados:** Separar parte dos pontos em duas classes com base em seus rótulos. Por exemplo, uma classe de "tem característica x" e outra de "não tem característica x".

- **Cálculo das envoltórias convexas:** Para cada classe, é calculada a envoltória convexa dos pontos pertencentes a essa classe utilizando o algoritmo de Graham.

- **Verificação de interseção:** É verificado se as envoltórias convexas das duas classes se interceptam. Se houver interseção, isso indica que as classes não podem ser separadas por uma linha reta (ou um hiperplano) e, portanto, não são linearmente separáveis.

- **Classificação baseada nas envoltórias convexas:** Se as envoltórias convexas das duas classes não se interceptam, a parte restante dos pontos que não foi utilizada para construir a reta de separabilidade podem ser classificados com base em qual envoltória convexa eles pertencem. Um ponto estará em na classe de uma envoltória se ele se encontrar do lado do plano que contém a envoltória.

- **Avaliação do Desempenho:** O desempenho do modelo de classificação determinado é avaliado com base em métricas apropriadas, como precisão, revocação, F1-score.

É importante destacar que a técnica de envoltórias convexas é mais apropriada para problemas de classificação binária, onde você está tentando separar dois grupos distintos de pontos. Para problemas de classificação multiclasse, o método foi estendido para tratar pares de classe “pertence a classe x” e “não pertence a classe x”.

Experimentos:

Ao todo, 10 bancos de dados diferentes foram utilizados para os experimentos. Todos os bancos são de dados reais para estudos de classificação e foram encontrados na plataforma KEEL. Os links contendo as características de cada banco de dados se encontram no final do relatório.

Para cada experimento, é necessário realizar um pré-processamento responsável pela redução de dimensionalidade dos dados para 2D. Os dados em duas dimensões são formatados para que possam ser devidamente passados como parâmetros nas funções dos algoritmos apresentados nas etapas apresentadas anteriormente.

Para reduzir os dados de K dimensões para 2 dimensões, foi utilizado o algoritmo PCA (Principal Component Analysis). Ele fornece um mapeamento de um espaço com N dimensões (tamanho original) para um espaço com M dimensões (que nesse caso é 2), através de um procedimento algébrico que converte as variáveis originais num conjunto de variáveis não correlacionadas, através da decomposição dos dados originais (uma matriz X) em duas matrizes. Assim, foi utilizada a função sklearn.decomposition.PCA já existente nas bibliotecas incluídas em Python.

Cerca de 70% dos pontos de cada experimento foram utilizados para a construção do modelo de classificação. Os outros 30% restantes foram classificados com base no modelo determinado e a eficiência da classificação foi verificada. A escolha dos pontos que fazem parte do treinamento do modelo e da classificação foi feita de forma aleatória. Assim, cada execução de um experimento pode resultar em configurações de separabilidade diferentes para certos casos de distribuição de pontos. Para alguns experimentos, foram necessárias diversas tentativas até encontrar uma separabilidade adequada das envoltórias.

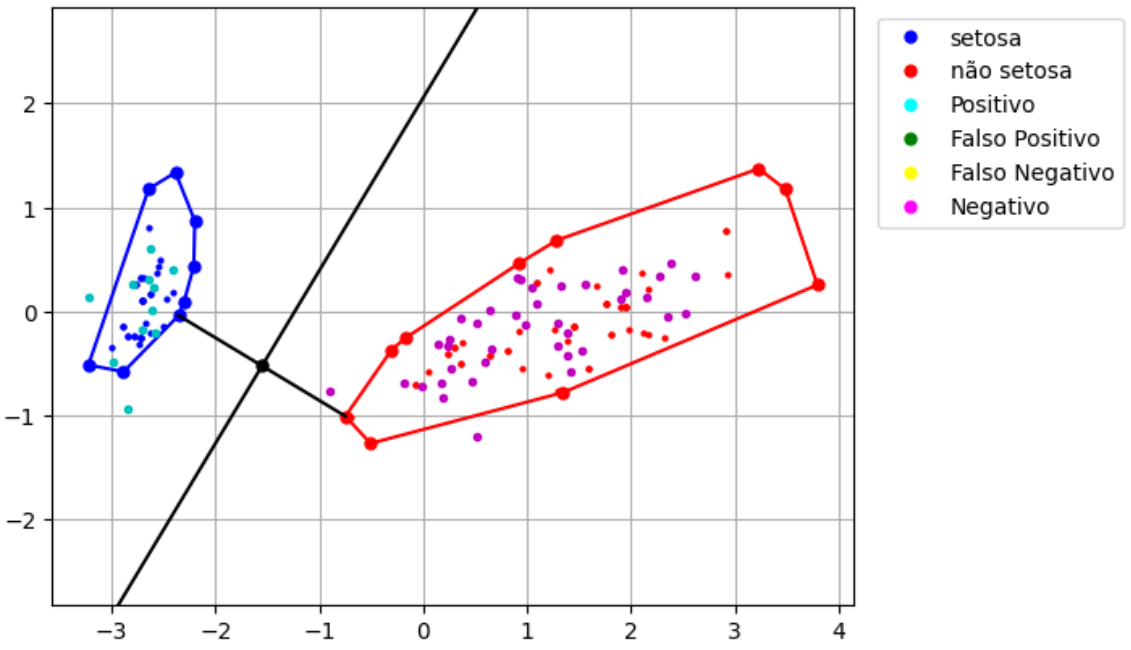
Todos os modelos de classificação foram definidos com base na equação da reta perpendicular que separa as envoltórias apresentada nos gráficos na seção a seguir.

Experimento 1: IRIS

Classes: setosa, não setosa

Separabilidade: Separável

Gráfico:



Métricas:

Precisão	Revocação	F1-Score
1	1	1

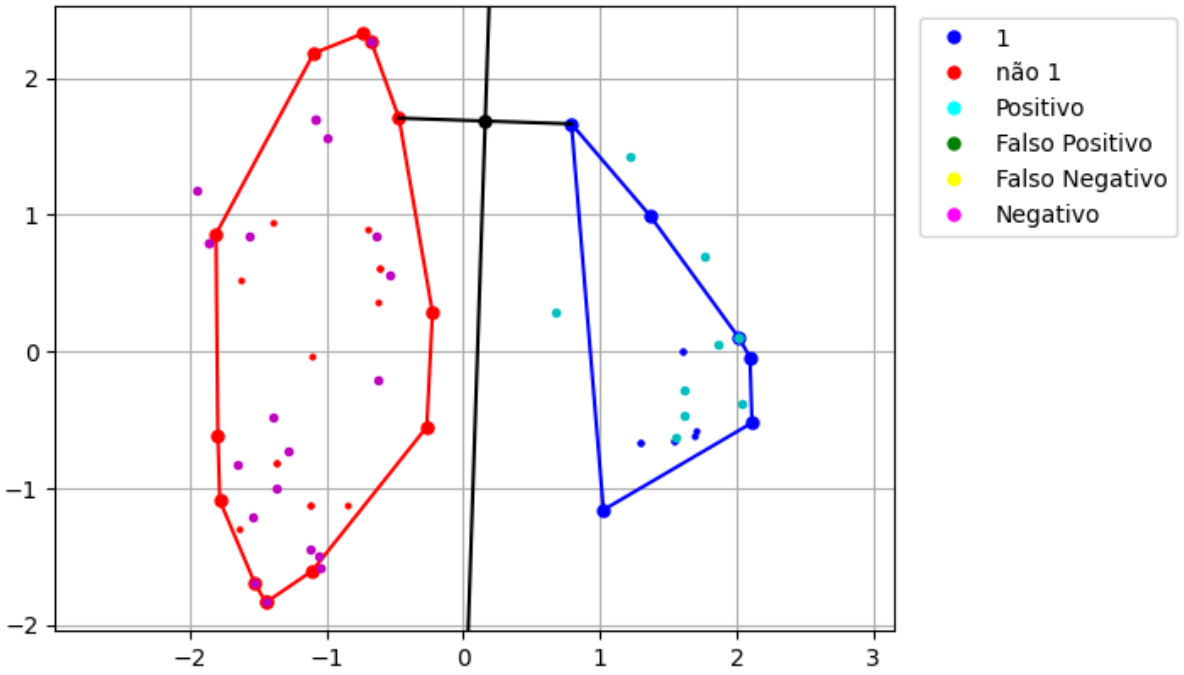
Análise dos resultados: Nesse cenário, as envoltórias convexas apresentam uma separabilidade muito bem definida, o que facilita a classificação dos pontos. Dessa forma, os pontos de teste foram inteiramente classificados de forma correta, resultando em métricas de avaliação do modelo com alta eficiência.

Experimento 2: ZOO

Classes: animal tipo 1, animal tipo não 1

Separabilidade: Separável

Gráfico:



Métricas:

Precisão	Revocação	F1-Score
1	1	1

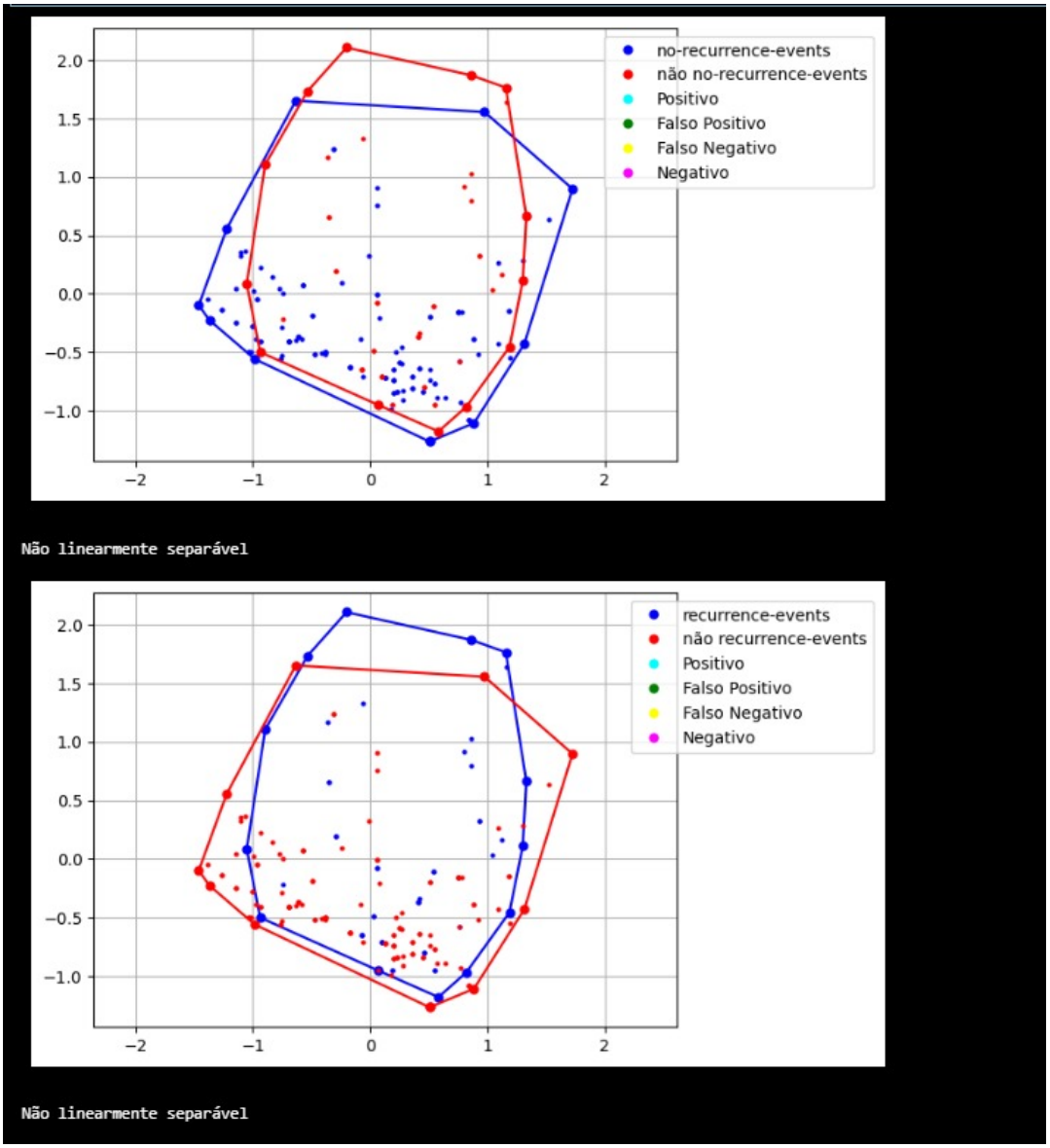
Análise dos resultados: Da mesma maneira que ocorreu no experimento anterior, as envoltórias convexas apresentam uma separabilidade muito bem definida, tornando a classificação dos pontos muito eficiente. Podemos chegar a uma conclusão de que o tipo animal 1 tem características bem distintas do restante das classes.

Experimento 3: BREAST CANCER

Classes: no-recurrence-events, recurrence-events

Separabilidade: Não separável

Gráfico:



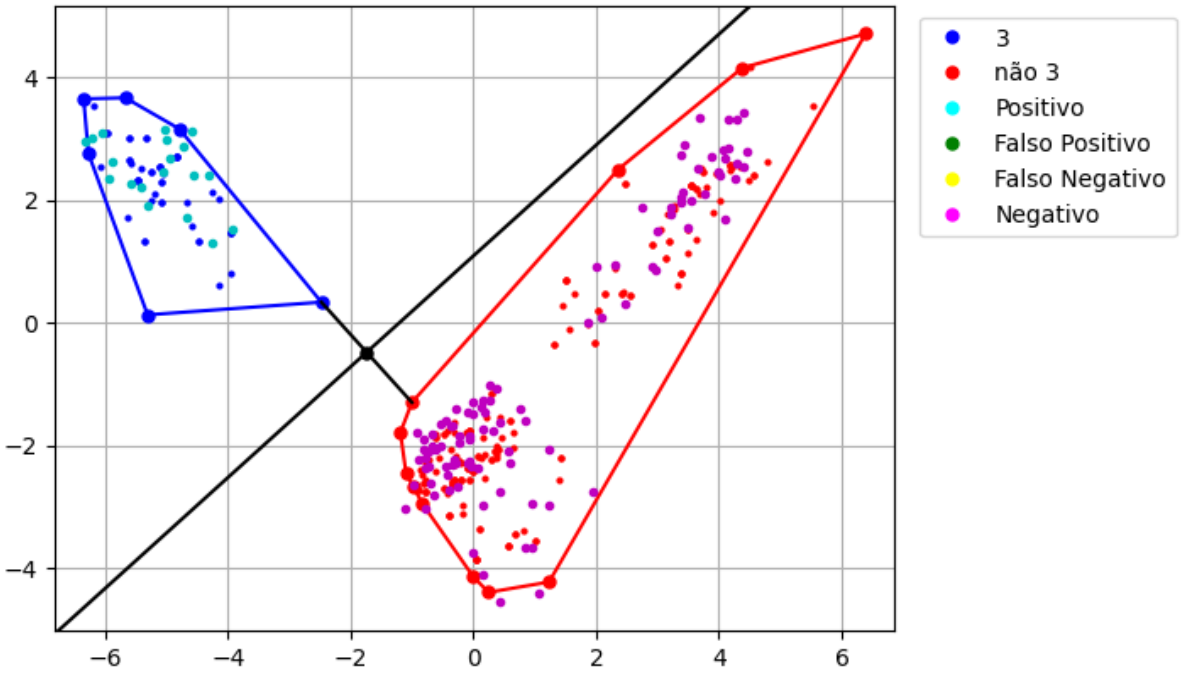
Análise dos resultados: Como pode ser observado nos gráficos apresentados, as envoltórias encontradas se interceptam em vários pontos. As classes de pontos não se distinguem o suficiente em suas características, tornando o método de classificação por separabilidade de envoltórias convexas inadequado.

Experimento 4: DERMATOLOGY

Classes: classe 3, classe não 3

Separabilidade: Separável

Gráfico:



Métricas:

Precisão	Revocação	F1-Score
1	1	1

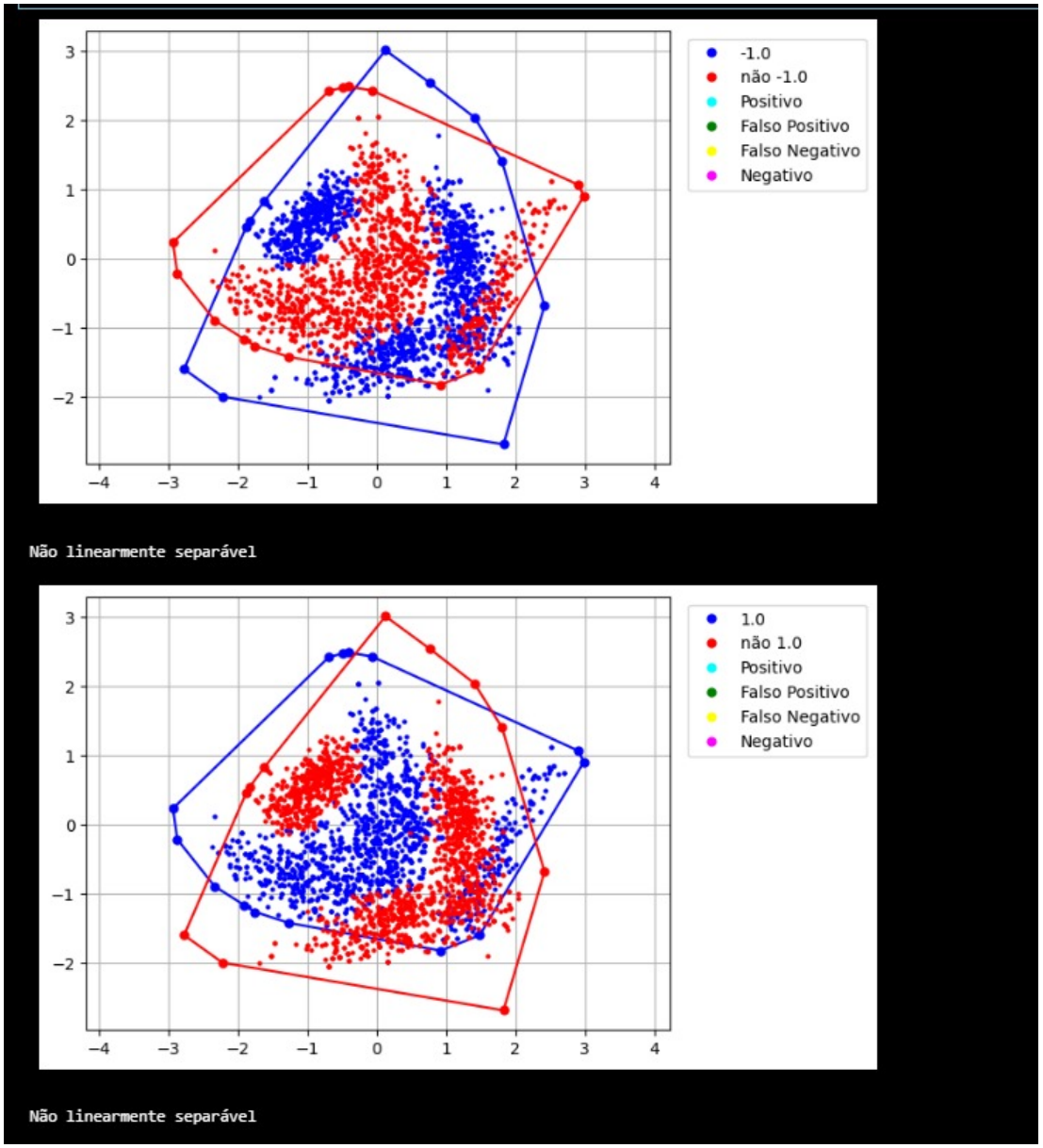
Análise dos resultados: Temos um outro caso onde as envoltórias convexas apresentam uma separabilidade bem definida e a classificação dos pontos foi totalmente correta. Assim, a classe 3 para o diagnóstico da doença tem características bem distintas do restante das classes e falsos positivos seriam improváveis.

Experimento 5: BANANA

Classes: formato 1.0, formato -1.0

Separabilidade: Não separável

Gráfico:



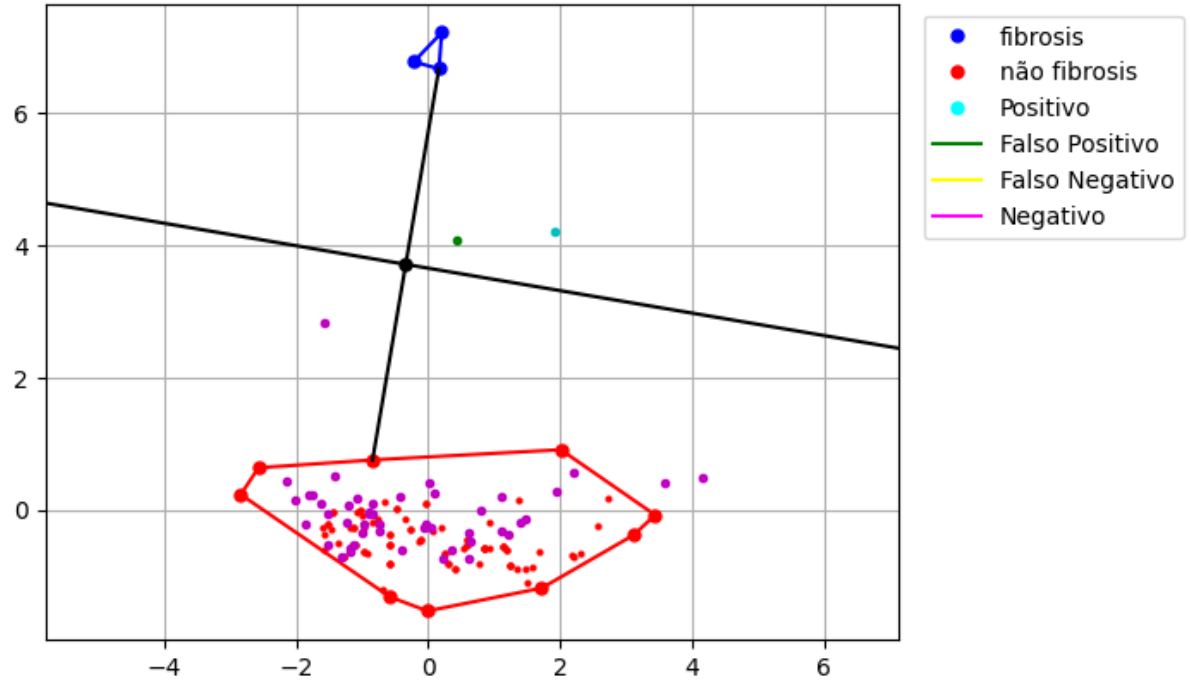
Análise dos resultados: Como pode ser observado nos gráficos apresentados, as envoltórias encontradas se interceptam em vários pontos. O banco de dados apresenta muitas instâncias de pontos e as classes não são se distinguem o suficiente nas suas características, dificultando as chances de aplicação do método por separabilidade por envoltórias convexas.

Experimento 6: LYMPHOGRAPHY

Classes: linfoma fibrosis, linfoma não fibrosis

Separabilidade: Separável

Gráfico:



Métricas:

Precisão	Revocação	F1-Score
0.5	1	0.66

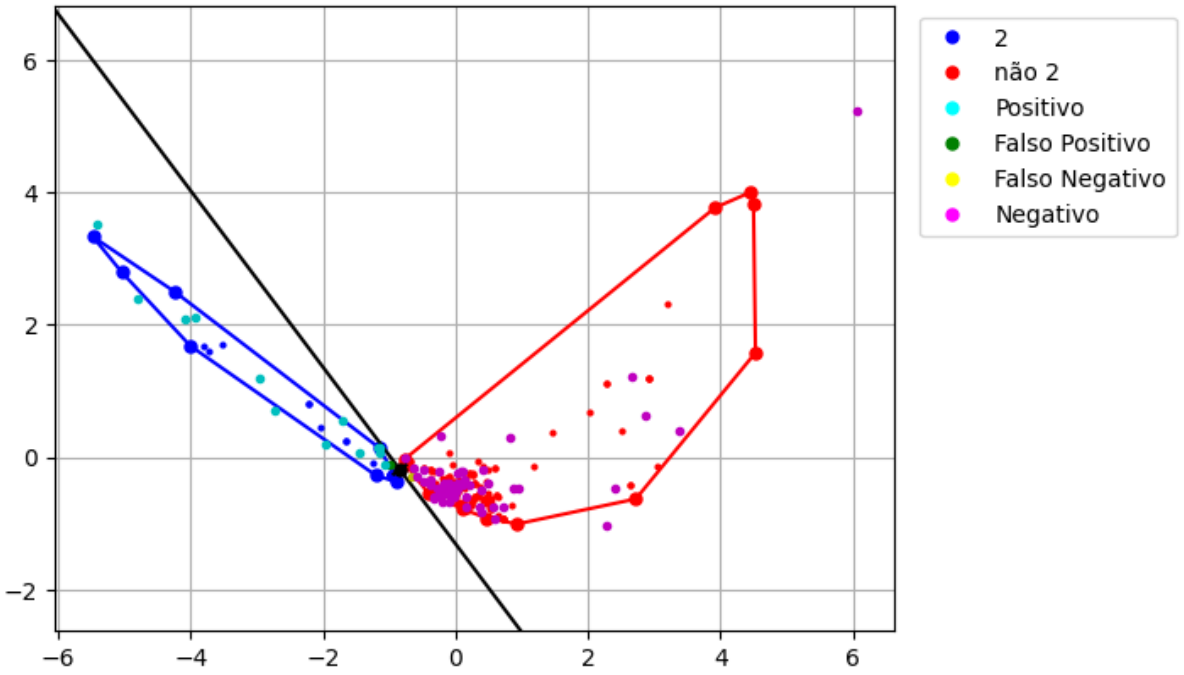
Análise dos resultados: O grupo fibrose, que foi separado, possui somente alguns pontos, o que faz com que seja difícil achar, com a escolha aleatória de pontos, um modelo linearmente separável, e quando um modelo é achado, não existem muitos pontos que seriam classificados como fibrose. Assim, com 1 de 2 pontos tendo sido um falso positivo, a precisão ficou no menor valor encontrado.

Experimento 7: THYROID

Classes: classe 2, classe não 2

Separabilidade: Separável

Gráfico:



Métricas:

Precisão	Revocação	F1-Score
0.93	0.93	0.93

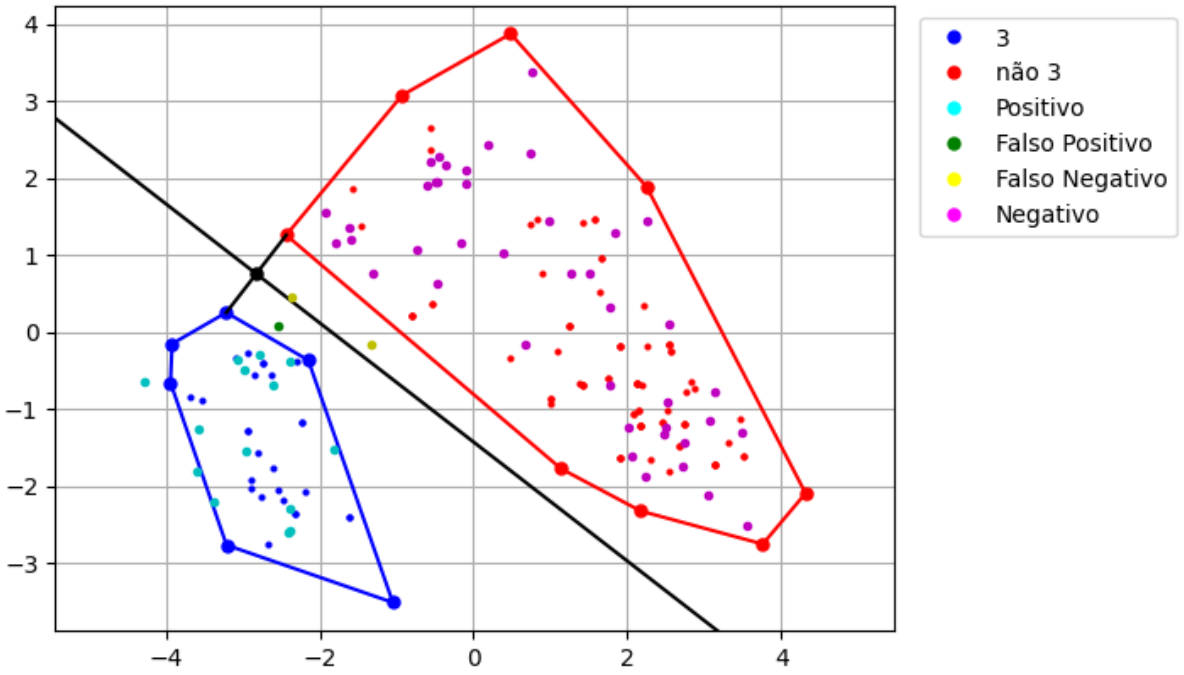
Análise dos resultados: A separação entre as classificações não é muito bem definida, então, assim como o anterior, encontrar um modelo com seleção aleatória de pontos dificilmente acontece na primeira tentativa e o modelo não é perfeito, tendo um falso positivo e um falso negativo bem perto da reta, a precisão e revocação não ficam perfeitas, com o f1-score sofrendo por ser calculado a partir dos dois.

Experimento 8: WINE

Classes: classe 3, classe não 3

Separabilidade: Separável

Gráfico:



Métricas:

Precisão	Revocação	F1-Score
0.93	0.87	0.90

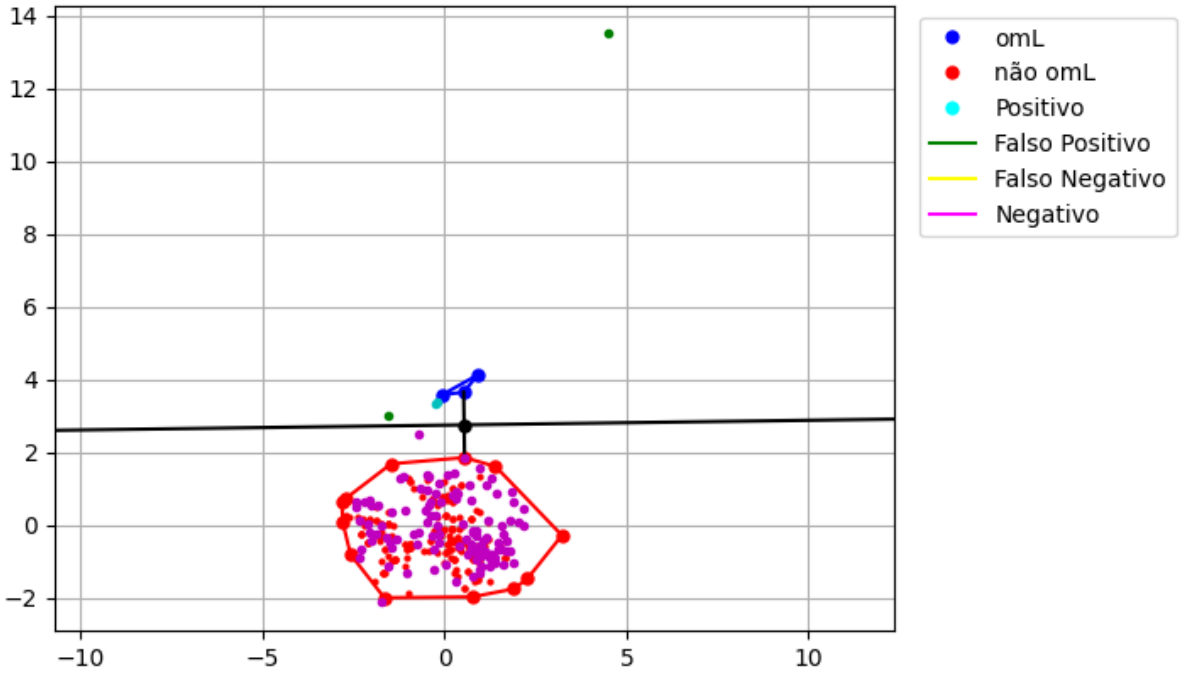
Análise dos resultados: A região da reta do modelo possui pontos de classificações diversas, então além de não ser instantâneo achar o modelo, ficam pontos mal classificados, igual no gráfico acima, onde ocorre um falso positivo e dois falso negativos. Assim, a precisão fica pouco acima da revocação e o f1-score entre os dois.

Experimento 9: ECOLI

Classes: classe omL, classe não omL

Separabilidade: Separável

Gráfico:



Métricas:

Precisão	Revocação	F1-Score
0.5	1	0.66

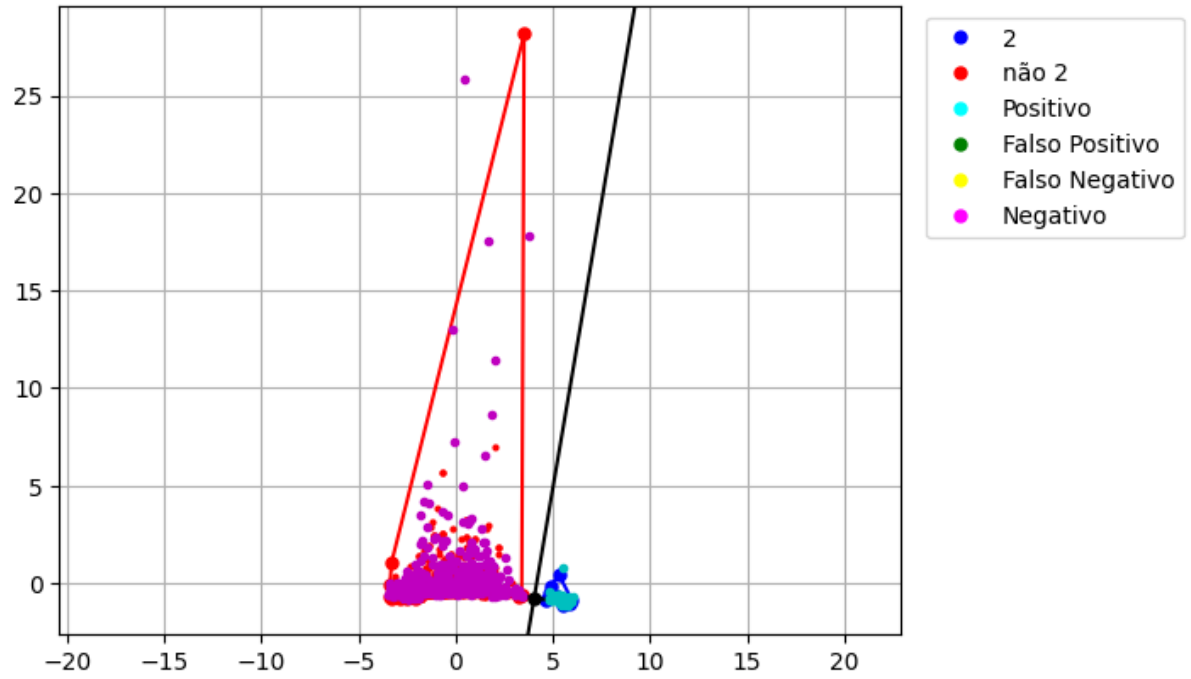
Análise dos resultados: Similar aos dados Lymphography, a envoltória de omL possui somente 3 pontos. Caso o ponto mais acima não tivesse sido escolhido para teste, a envoltória omL teria ficado dentro da outra, mas por ele ter sido de teste, o modelo obtido tinha somente 4 pontos para classificar como omL e desses 4, 2 foram falso positivos, resultando em uma precisão de 0,5. A revocação, por sua vez, fica em 1 já que os pontos omL ficam em uma região bem definida. A precisão faz com que o f1-score caia para 0.66.

Experimento 10: SEGMENT

Classes: classe 2, classe não 2

Separabilidade: Separável

Gráfico:



Métricas:

Precisão	Revocação	F1-Score
1	1	1

Análise dos resultados: O último experimento apresentou uma classificação de pontos com eficiência máxima. Apesar disso, as envoltórias convexas não estão separadas por uma distância muito grande. Podemos chegar à conclusão de que uma distância pequena entre envoltórias nem sempre é sinônimo de má qualidade na classificação. Por outro lado, o banco de dados não apresentou muitos pontos dos segmentos de classe 2 em suas instâncias. Assim, o resultado de separabilidade pode ter sido desvirtuado do real.

Conclusão:

O método de classificação estudado é um exemplo de abordagem geométrica e pode ser útil em casos específicos, mas sua eficácia depende da natureza dos dados e da suposição de separabilidade linear entre as classes.

Banco de dados:

Breve resumo dos bancos de dados utilizados

ZOO (https://sci2s.ugr.es/keel/dataset.php?cod=69): Este banco de dados é bem simples onde a tarefa é classificar os animais em sete classes predefinidas e a maioria dos atributos tem valor booleano. O conjunto de dados contém 7 classes de 101 instâncias cada.

LYMPHOGRAPHY (https://sci2s.ugr.es/keel/dataset.php?cod=64): Este é um domínio fornecido pelo Instituto de Oncologia que tem aparecido repetidamente na literatura de aprendizado de máquina. A tarefa é detectar a presença de um linfoma e seu estado atual. O conjunto de dados contém 4 classes de 148 instâncias cada.

SEGMENT (https://sci2s.ugr.es/keel/dataset.php?cod=107): Este banco de dados contém instâncias extraídas aleatoriamente de um banco de dados de 7 imagens externas (classes). As imagens foram segmentadas manualmente para criar uma classificação para cada pixel. Cada instância codifica uma região 3x3. A tarefa é determinar o tipo de superfície de cada região.

DERMATOLOGY (https://sci2s.ugr.es/keel/dataset.php?cod=60): No conjunto de dados construído para este domínio, o recurso histórico familiar tem valor 1 se alguma dessas doenças foi observada na família e 0 caso contrário. O conjunto de dados contém 6 classes de 366 instâncias cada.

IRIS (https://sci2s.ugr.es/keel/dataset.php?cod=18): Este é talvez o banco de dados mais conhecido encontrado na literatura de reconhecimento de padrões. O conjunto de dados contém 3 classes de 50 instâncias cada, onde cada classe se refere a um tipo de planta de íris. Uma classe é linearmente separável das outras 2; os últimos NÃO são linearmente separáveis entre si.

THYROID (https://sci2s.ugr.es/keel/dataset.php?cod=66): Este conjunto de dados é uma das diversas bases de dados sobre Tireóide disponíveis no repositório da UCI. A tarefa é detectar se um determinado paciente é normal (1) ou sofre de hipertireoidismo (2) ou hipotireoidismo (3). O conjunto de dados contém 3 classes de 215 instâncias cada.

WINE (https://sci2s.ugr.es/keel/dataset.php?cod=31): Este banco de dados contém o resultado de uma análise química de vinhos cultivados na mesma região da Itália, mas derivados de três cultivares diferentes. O conjunto de dados contém 3 classes de 178 instâncias cada.

ECOLI (https://sci2s.ugr.es/keel/dataset.php?cod=61): Este conjunto de dados é composto por algumas medidas sobre as células (citoplasma, membrana interna, perisplasma, membrana externa, lipoproteína da membrana externa, lipoproteína da membrana interna, membrana interna, sequência sinal clivável). A tarefa é prever o local de localização das proteínas utilizando estes dados. O conjunto de dados contém 8 classes de 336 instâncias cada.

BANANA (https://sci2s.ugr.es/keel/dataset.php?cod=182): Este banco de dados contém instâncias pertencentes a vários clusters em formato de banana. A tarefa é diferenciar o formato da banana, dentre os 2 tipos presentes nos dados. O conjunto de dados contém 2 classes de 5300 instâncias cada.

BREAST CANCER (https://sci2s.ugr.es/keel/dataset.php?cod=97): Este é um dos três domínios fornecidos pelo Instituto de Oncologia que tem aparecido repetidamente na literatura de aprendizado de máquina. A tarefa é detectar a presença de um linfoma e seu estado atual. Este conjunto de dados inclui 201 instâncias de uma classe e 85 instâncias de outra classe.