# OCC e PUL

1001513 – Aprendizado de Máquina 2
Turma A – 2023/2
Prof. Murilo Naldi

naldi@ufscar.br

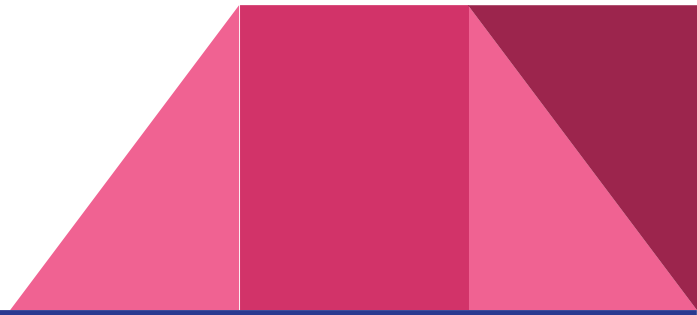
# Agradecimentos


- Pessoas que colaboraram com a produção deste material: Diego Silva, Ricado Campello, Ricardo Cerri
- Intel IA Academcy

# Na última aula...

Vimos que AM semi-supervisionado pode ser útil

- Especialmente quando levantamos a questão da dificuldade (custo) em rotular exemplos

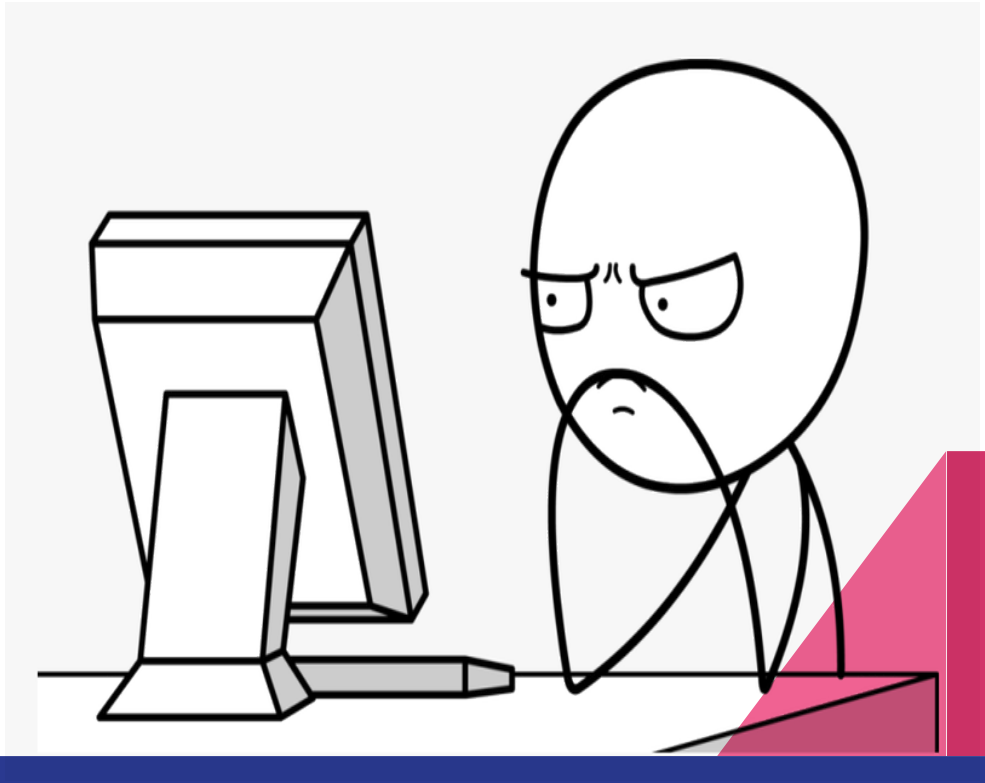


# Na última aula...

Mas, agora, vamos pensar que sequer conseguimos exemplos de todas as classes

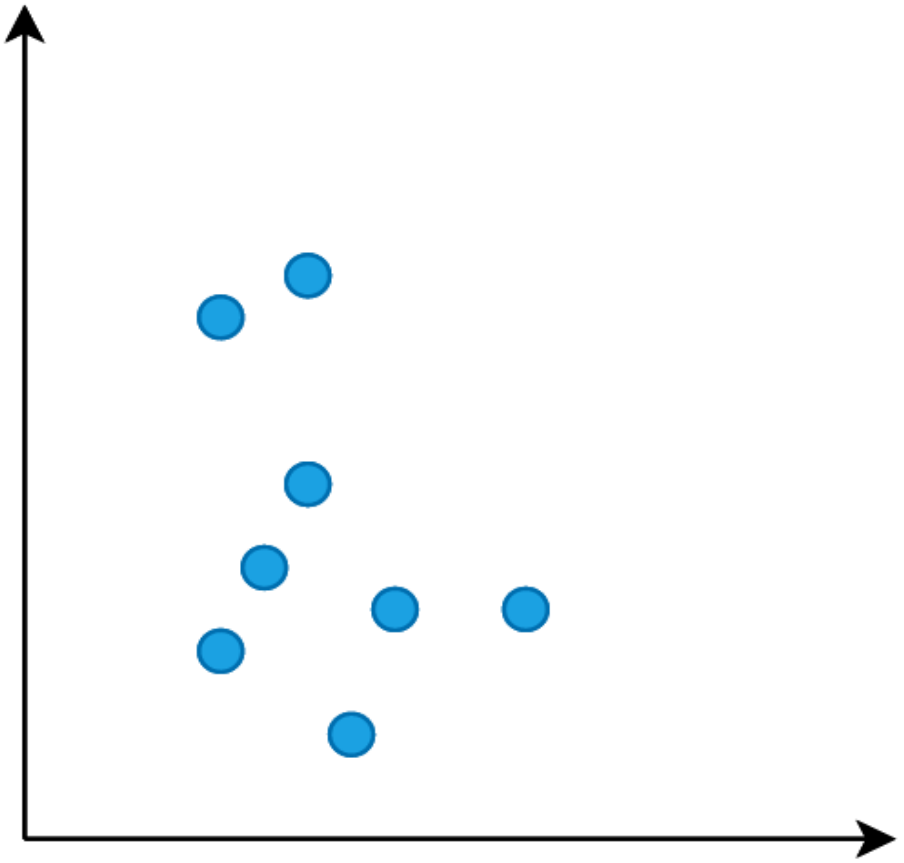
Ou pior ainda, só conseguimos para uma classe.

E agora, José?



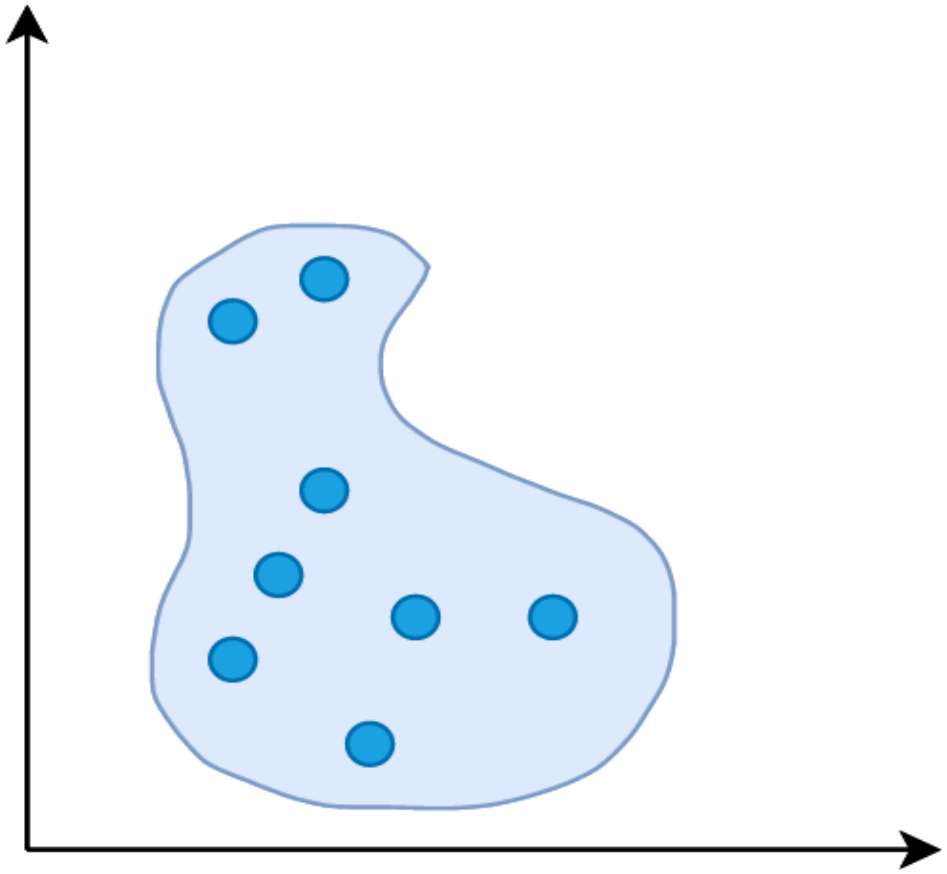
# One-class classification (OCC)

Nesse cenário, consideramos só possuirmos dados de uma classe



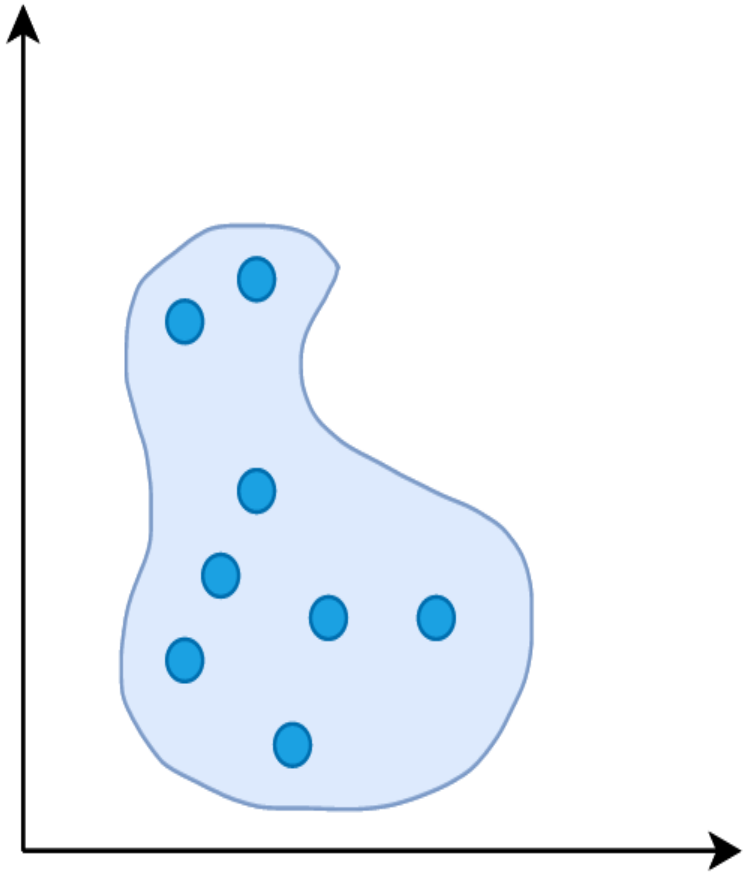
# One-class classification (OCC)

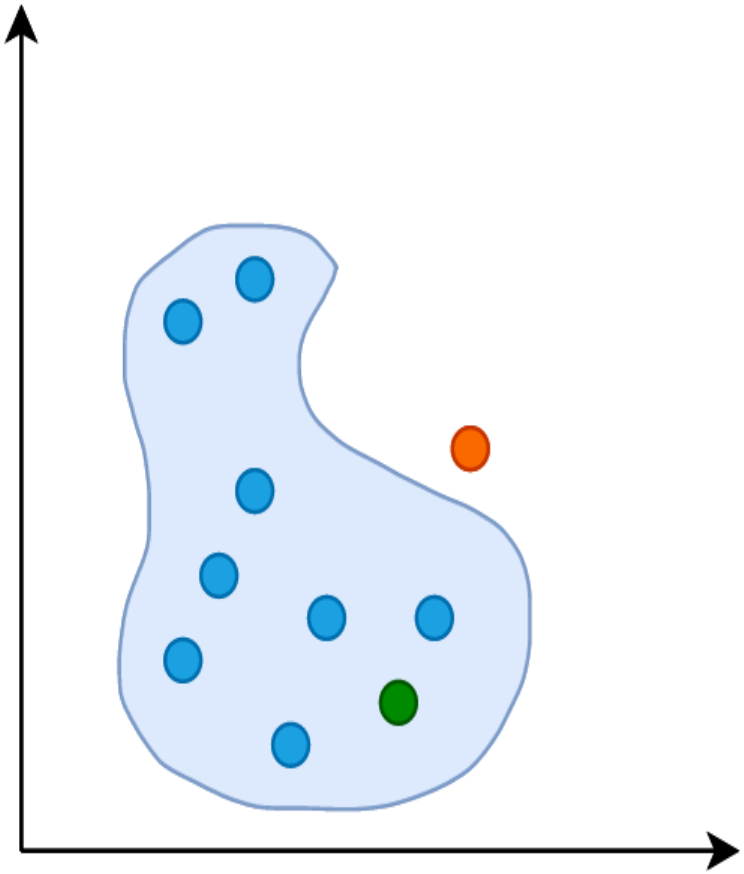
Nesse cenário, consideramos só possuirmos dados de uma classe



# One-class classification (OCC)

Nesse cenário, consideramos só possuirmos dados de uma classe





# One-class classification (OCC)

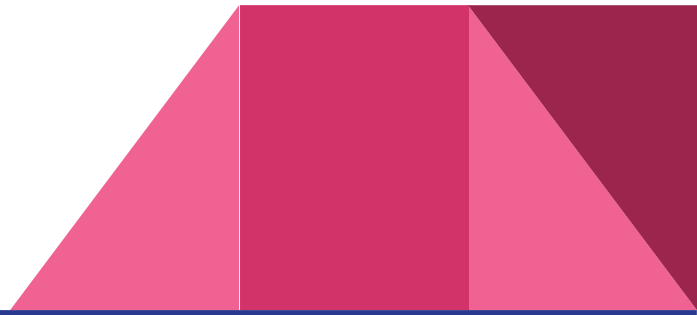
Nesse cenário, consideramos só possuirmos dados de uma classe

- Algoritmos dessa categoria são convenientes
  - Detecção de anomalias (monitorar funcionamento de um motor)
  - Detecção de *outliers*
  - É impossível coletar dados do universo (caso do sensor de insetos)

# One-class classification (OCC)

Há diversos algoritmos

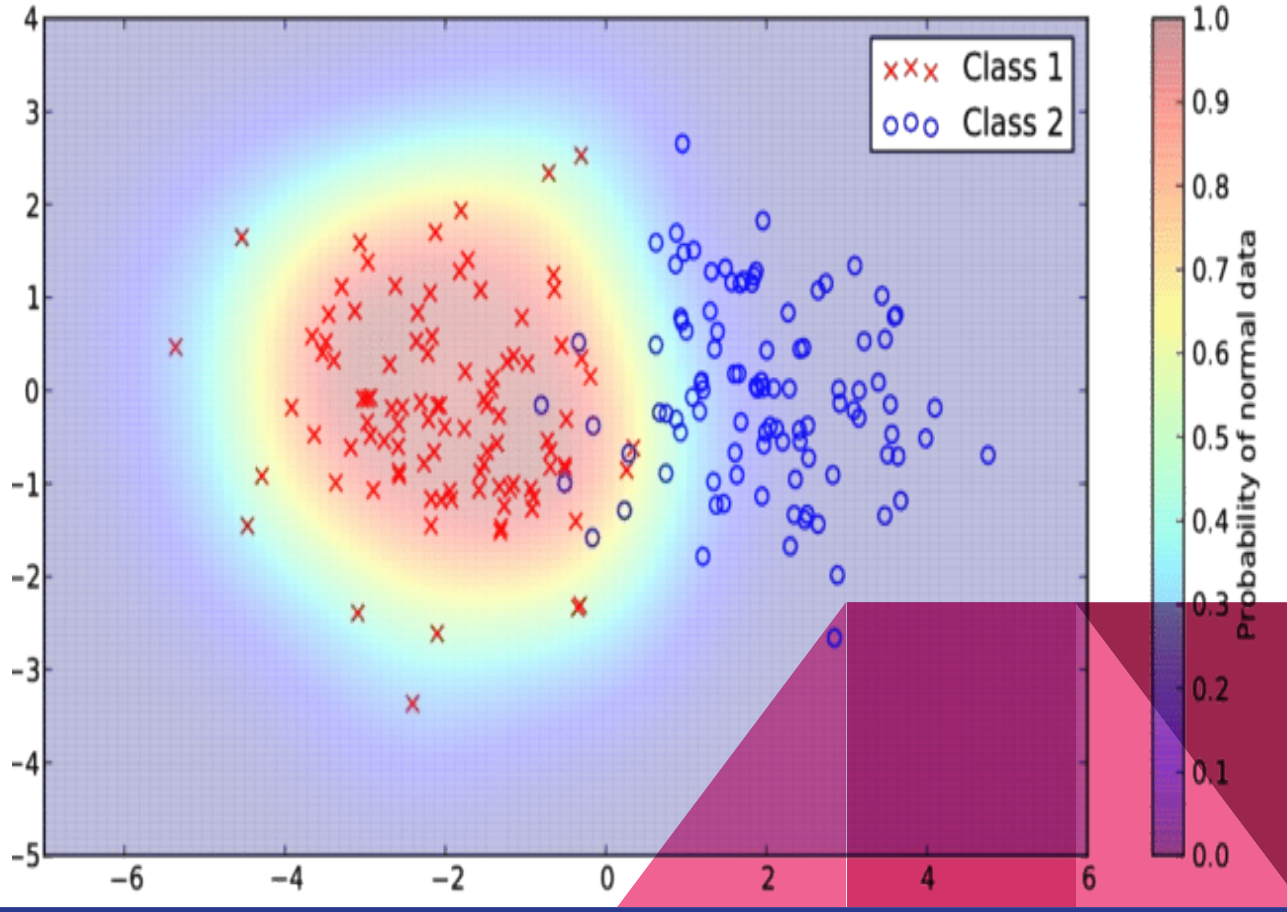
- Baseados em distribuição
- OC-SVM
- Isolation Forest
- OC-kNN



# Baseados em distribuição

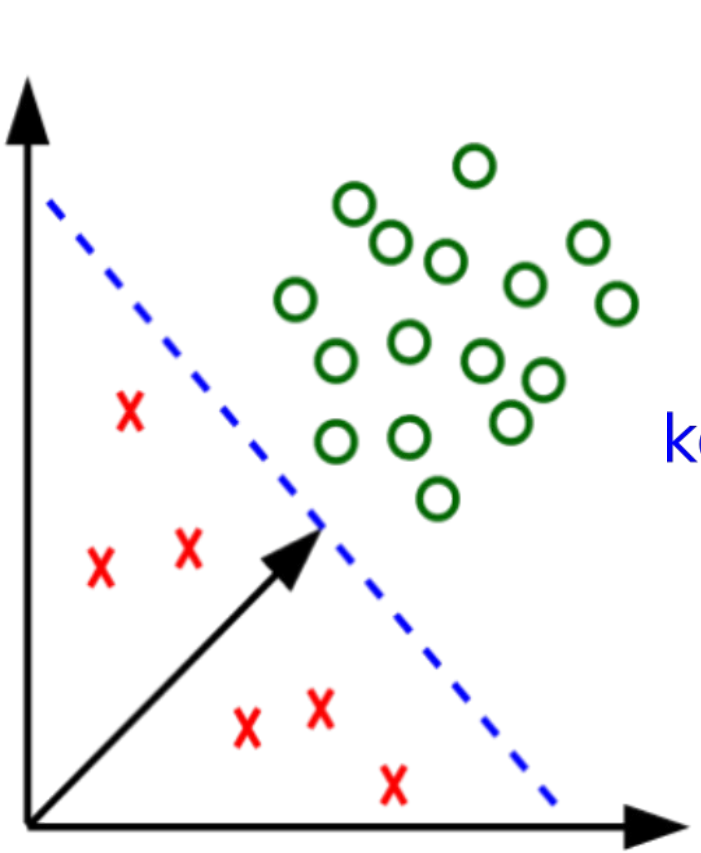
Assume que os objetos da classe se comportem segundo uma distribuição pré-definida

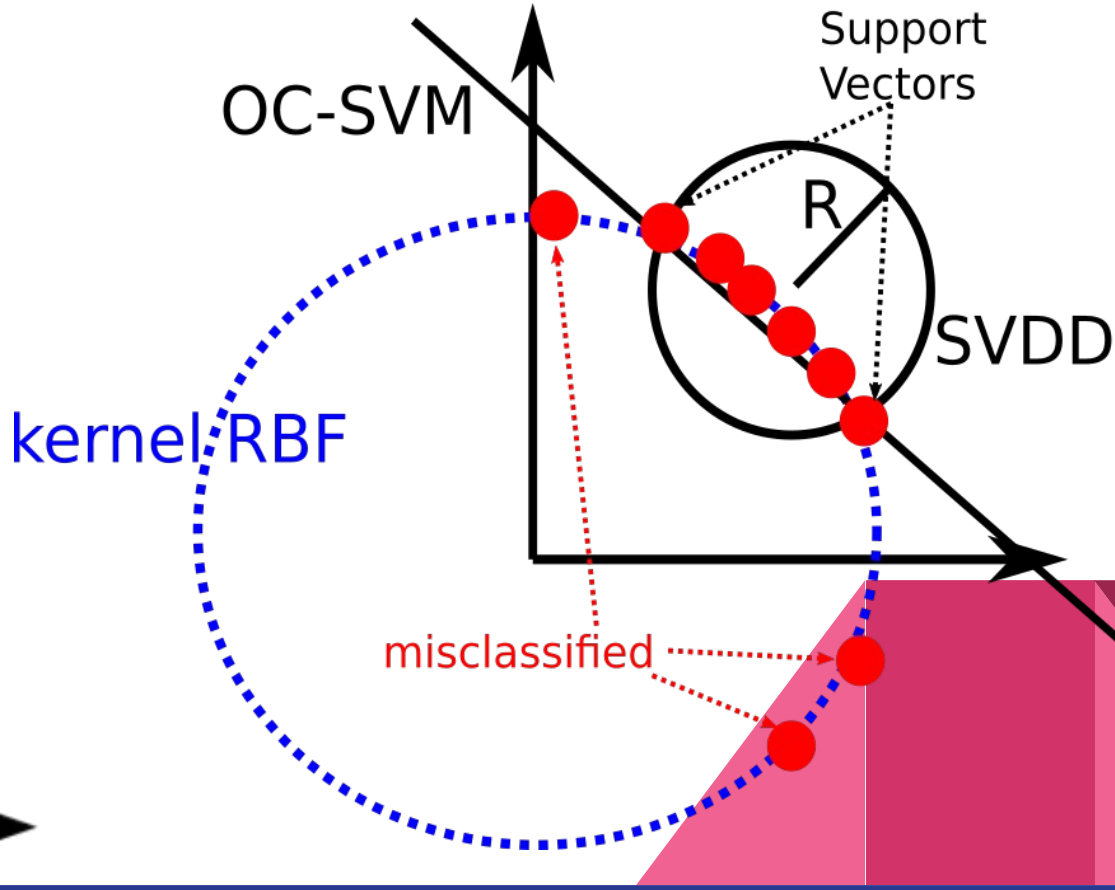
Exemplos:
Gaussiana ou Poisson



# One-Class SVM

Semelhante ao SVM, mas separa apenas a classe conhecida
Variação usa hiperesfera - SVDD(Support Vector Data Description) - Com ou sem *kernel*

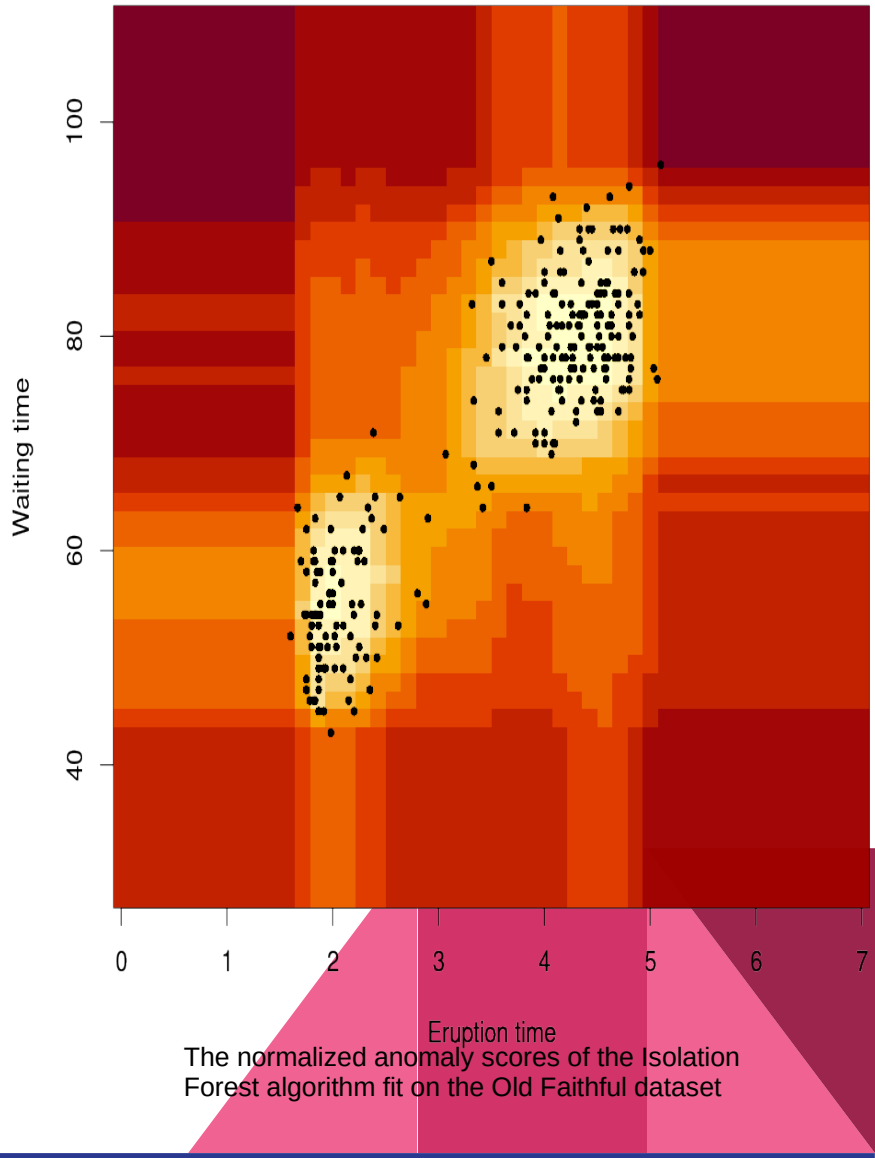




# Isolation Forest

Desenvolvido originalmente para detecção de anomalias

- Isolation Tree
  - Como uma árvore de decisão, mas para separar um espaço
  - Quando pronta, verifica as folhas
    - Anomalia estarão sozinhos em nó folha mais próximos da raiz
    - Mais fácil de isolar




The normalized anomaly scores of the Isolation Forest algorithm fit on the Old Faithful dataset

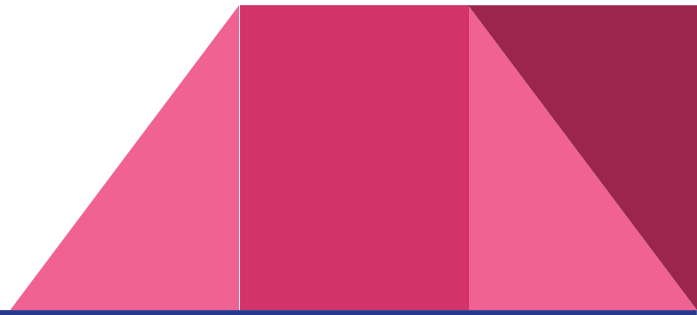
# One-class kNN

Como kNN, mas para uma classe só????

# One-class kNN

Como kNN, mas para um classe só

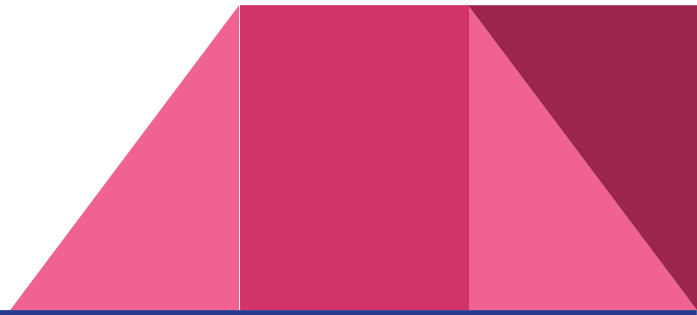
- d(z,y): distância entre duas amostras z e y
- *k*NN(y): *k*-ésimo vizinho mais próximo da amostra y
- Amostra z, encontra *k*NN(z) = y, classifica z como pertencente à classe alvo quando $d(z,y)/d(y,kNN(y)) < \delta$
  - δ pré definido
- Pode ser aplicado com médias e outras medidas



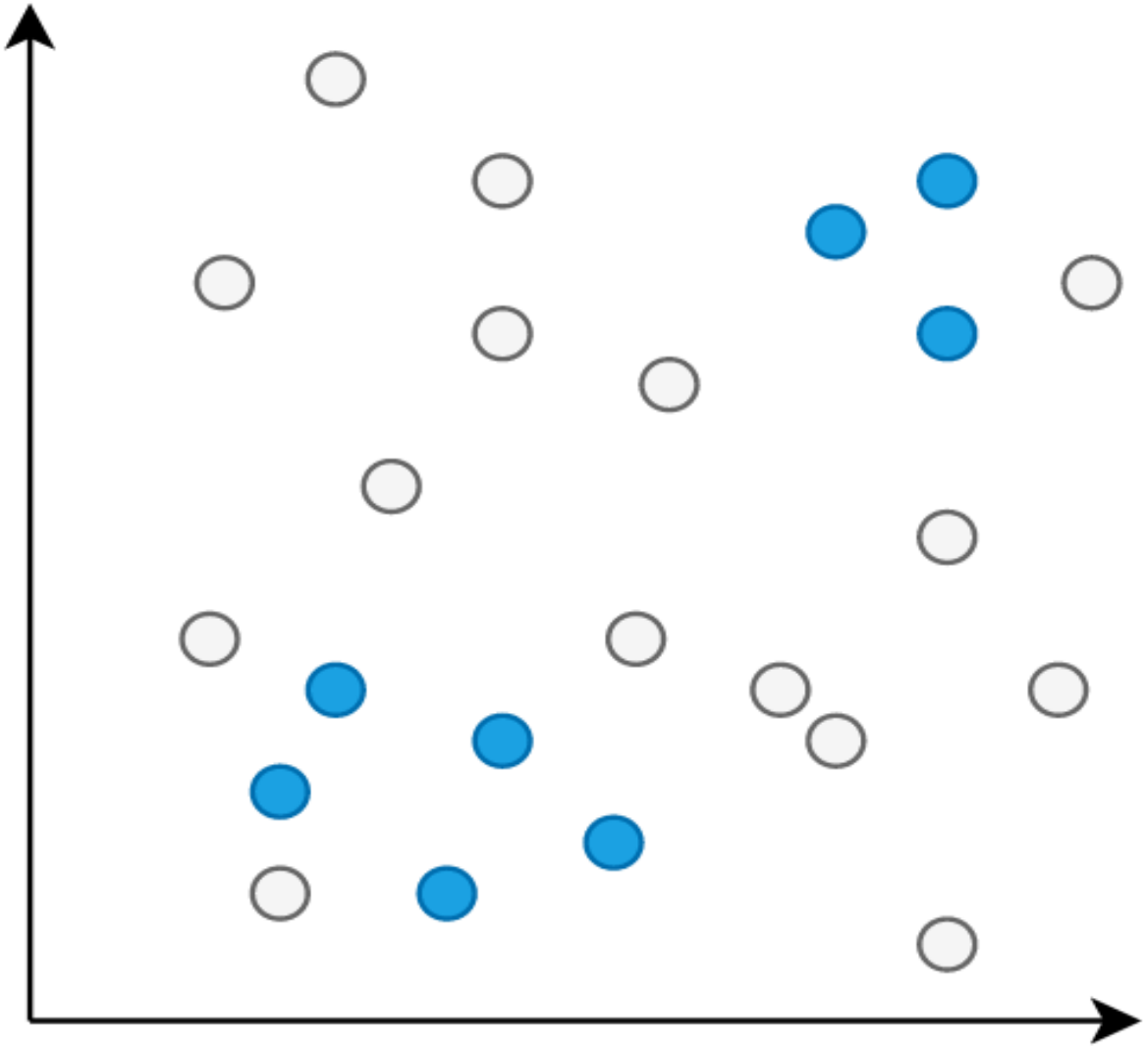
# Positive and Unlabeled Learning (PUL)

Em alguns cenários podemos considerar que:

- não conseguimos rótulos além da classe positiva
- é fácil conseguir exemplos não rotulados

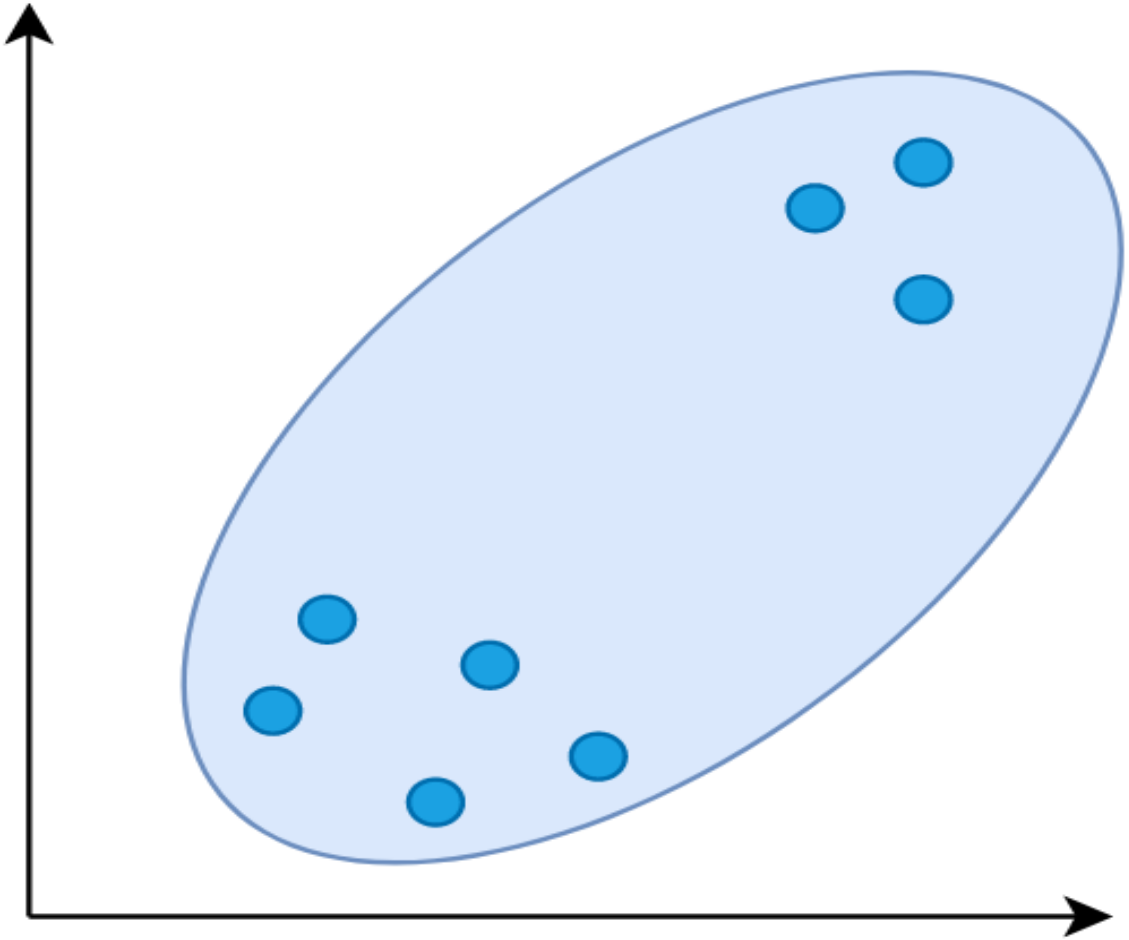


# Positive and Unlabeled Learning (PUL)



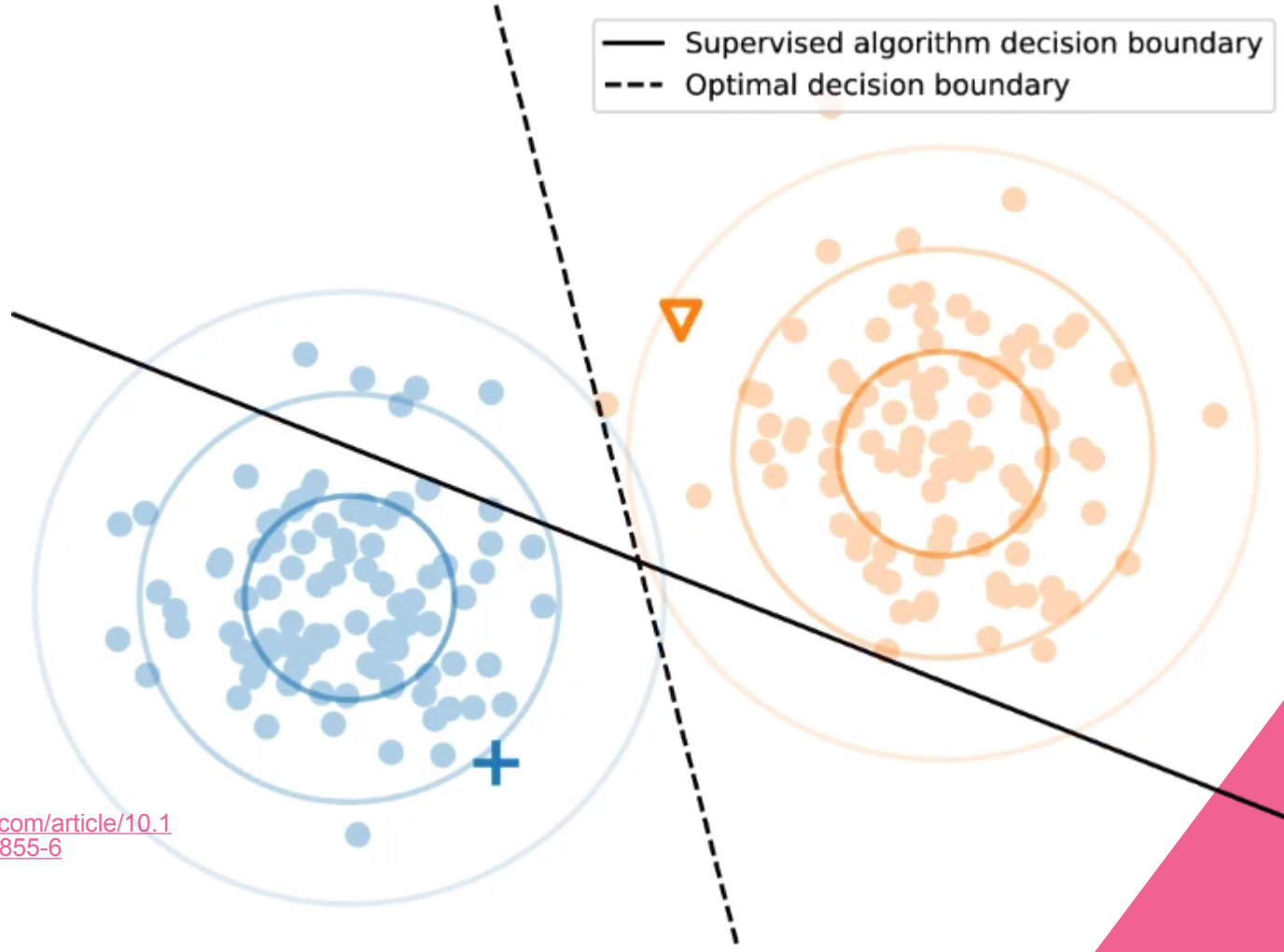
# Positive and Unlabeled Learning (PUL)

Se eu atacasse como OCC



# Positive and Unlabeled Learning (PUL)

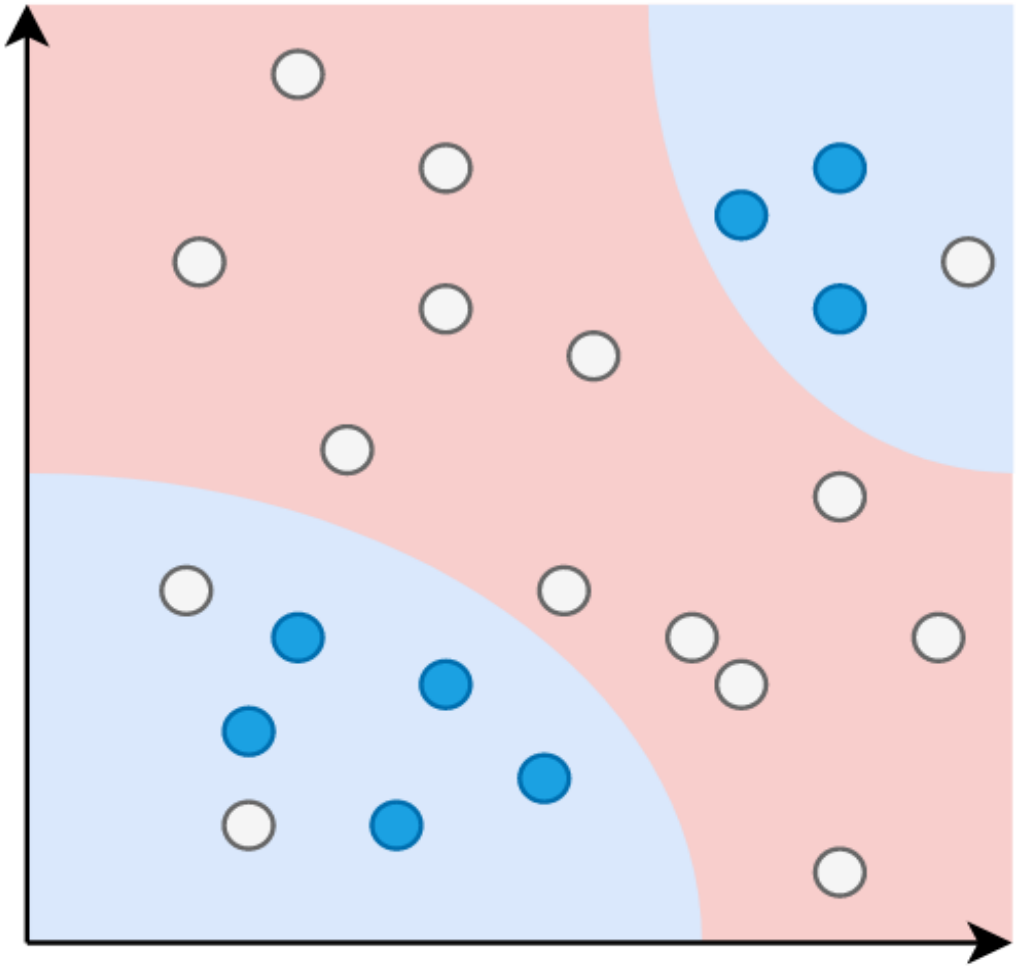
Um detalhe: aprendizado indutivo semi supervisionado




https://link.springer.com/article/10.1007/s10994-019-05855-6

# Positive and Unlabeled Learning (PUL)

Então, usar os não rotulados deve ser bom aqui também

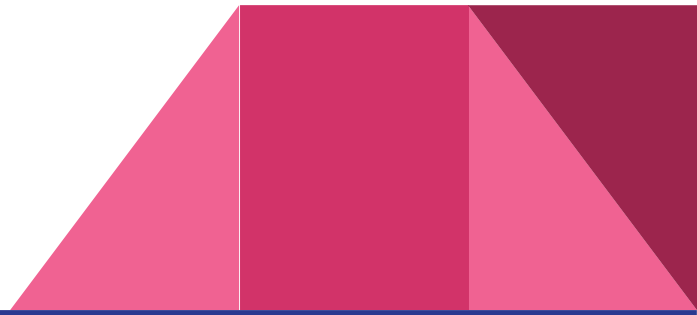


# Positive and Unlabeled Learning (PUL)

É possível enxergar um exemplo não rotulado como sendo uma entre duas opções:

1. Um exemplo negativo (ou)
2. Um exemplo positivo que não foi selecionado pelo mecanismo de rotulação

Sendo assim, é preciso fazer suposições sobre o mecanismo de rotulação, a distribuição das classes, ou ambos...



# Positive and Unlabeled Learning (PUL)

Formalmente,na classificação binária, temos:

$$\mathbf{x} \sim f(x)$$
$$\sim \alpha f_+(x) + (1 - \alpha) f_-(x)$$

em que $\alpha = Pr(y = 1)$ e a função de probabilidade de densidade da distribuição da classe positiva é $f_+$ e $f_-$ é a função da classe negativa

# Positive and Unlabeled Learning (PUL)

A distribuição que rege os rótulos é uma versão enviesada da distribuição positiva:

$$f_l(x) = \frac{e(x)}{c} f_+(x)$$

em que e(x) é uma função de propensão e($x$) = Pr($s = 1|y = 1$, $x$) e c é a frequência de rótulos $c$ = Pr($s = 1|y = 1$)

# Positive and Unlabeled Learning (PUL)

No cenário de treino único, assume-se que os dados positivos e não rotulados pertencem ao mesmo conjunto

$$\mathbf{x} \sim f(x)$$
$$\sim \alpha f_+(x) + (1 - \alpha) f_-(x)$$
$$\sim \alpha c\, f_l(x) + (1 - \alpha c) f_u(x)$$

em que $f_u$ e a função de probabilidade de densidade da distribuição dos objetos não rotulados

# Positive and Unlabeled Learning (PUL)

No cenário de controle de caso, assume-se que os dados positivos e não rotulados são de conjuntos distintos

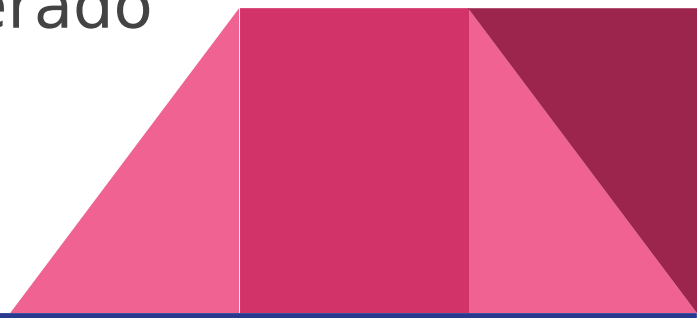
$$\mathbf{x|s} = 0 \sim f_u(x)$$
$$\sim f(x)$$
$$\sim \alpha f_+(x) + (1 - \alpha) f_-(x)$$

Nesse cenário é considerado que o conjunto de dados positivos somente terá dados positivos

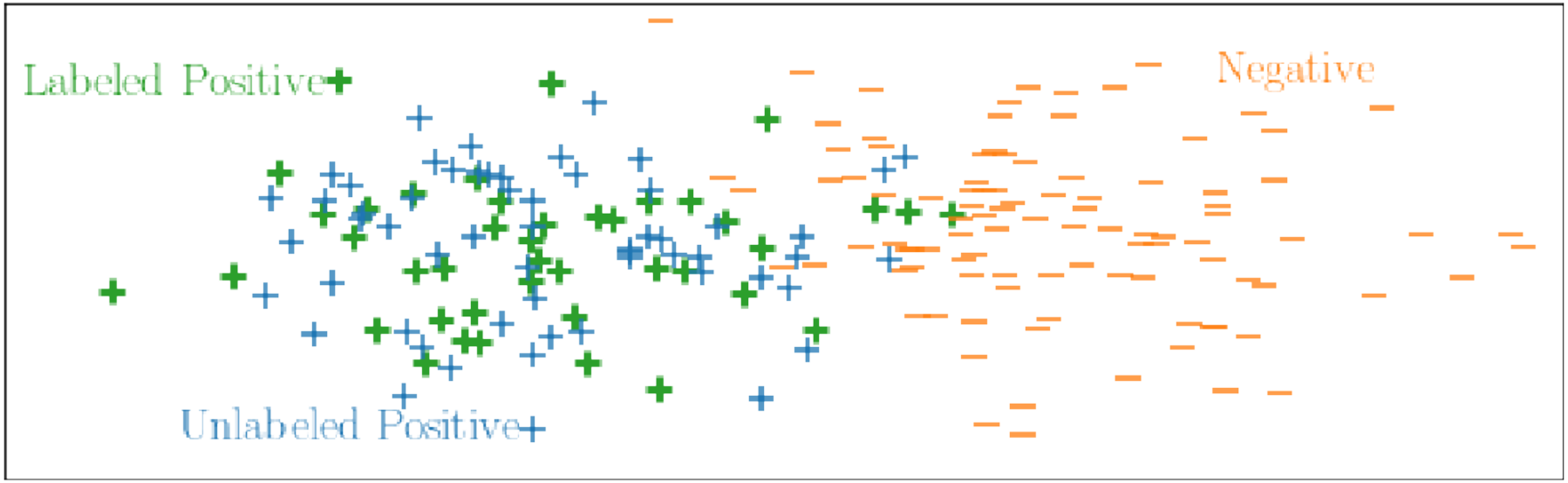
# Positive and Unlabeled Learning (PUL)

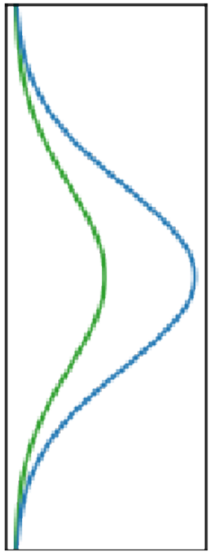
Podemos criar variações de algoritmos supervisionados

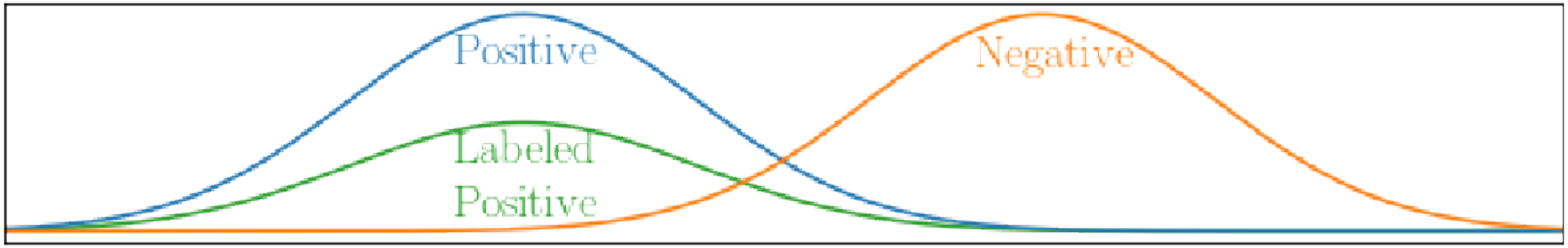
- Probabilístico que modela P e U
- Baseados em otimização
  - Maximizar o número de U fora da região, mas garantindo todos os positivos nela
- Atribuir pesos aos U e ajustar uma curva
  - Regressão logística com erro ponderado



# Positive and Unlabeled Learning (PUL)



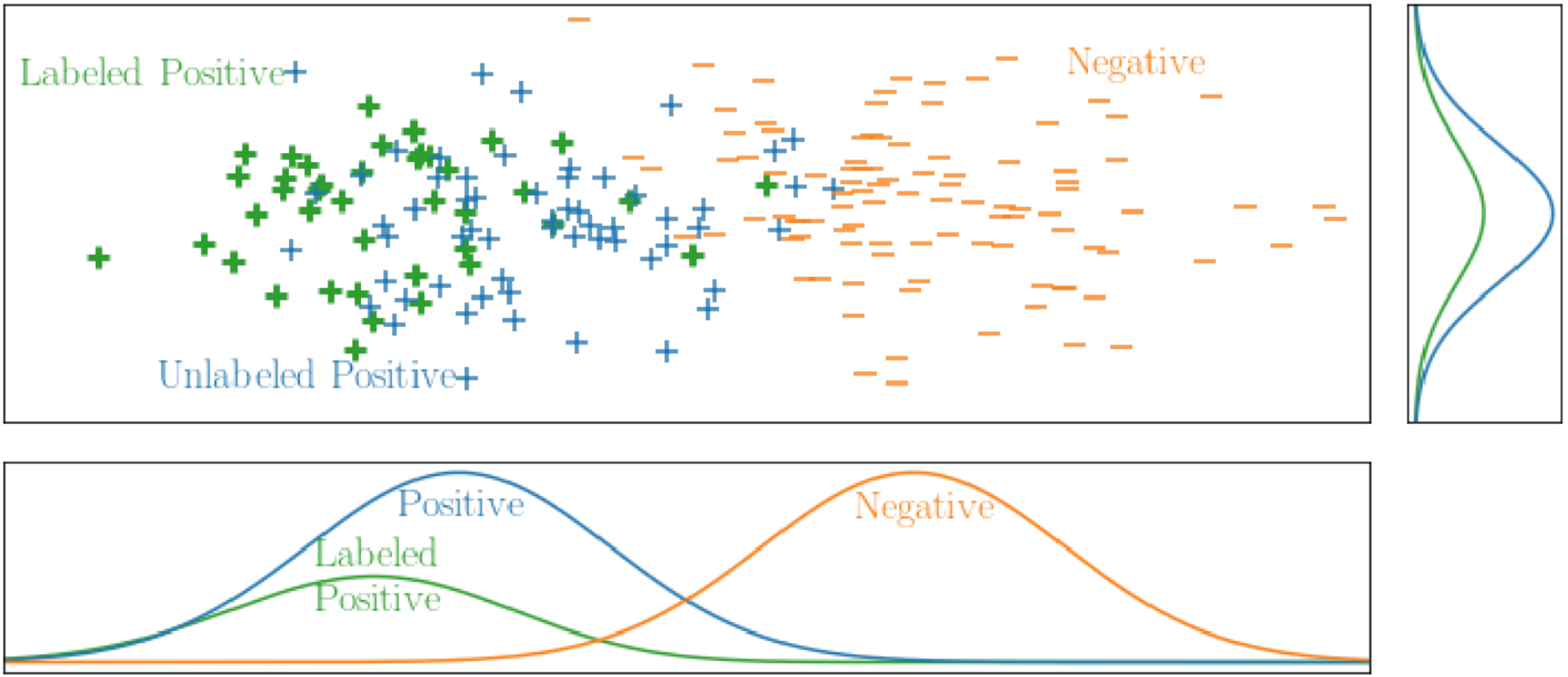






- Exemplo em que o rotulador escolhe os amostras de forma uniforme

Bekker, J., Davis, J. Learning from positive and unlabeled data: a survey. Mach Learn 109, 719–760 (2020). https://doi.org/10.1007/s10994-020-05877-5

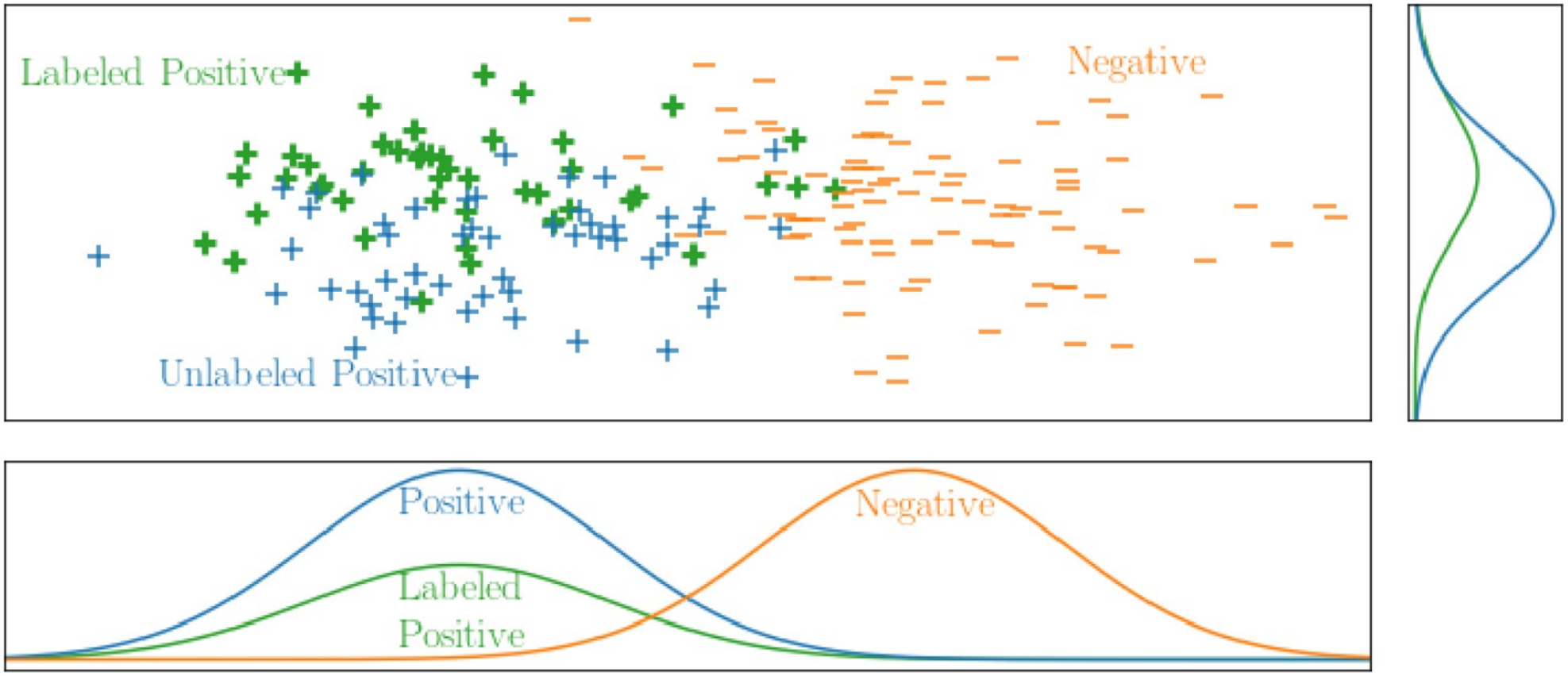
# Positive and Unlabeled Learning (PUL)




- Exemplo em que o rotulador escolhe amostras com viés positivo, afastando a chance de escolher negativos

Bekker, J., Davis, J. Learning from positive and unlabeled data: a survey. Mach Learn 109, 719–760 (2020). https://doi.org/10.1007/s10994-020-05877-5

# Positive and Unlabeled Learning (PUL)



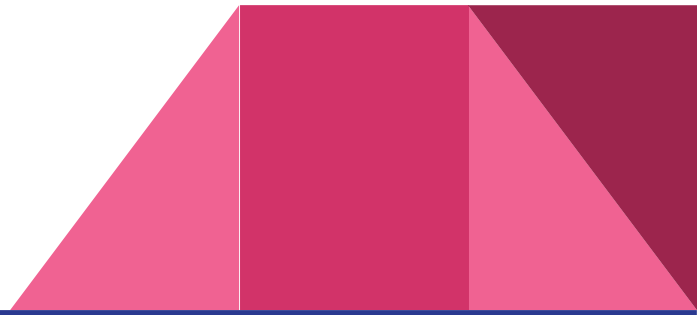

- Exemplo em que o rotulador independente, com viés próprio

Bekker, J., Davis, J. Learning from positive and unlabeled data: a survey. Mach Learn 109, 719–760 (2020). https://doi.org/10.1007/s10994-020-05877-5

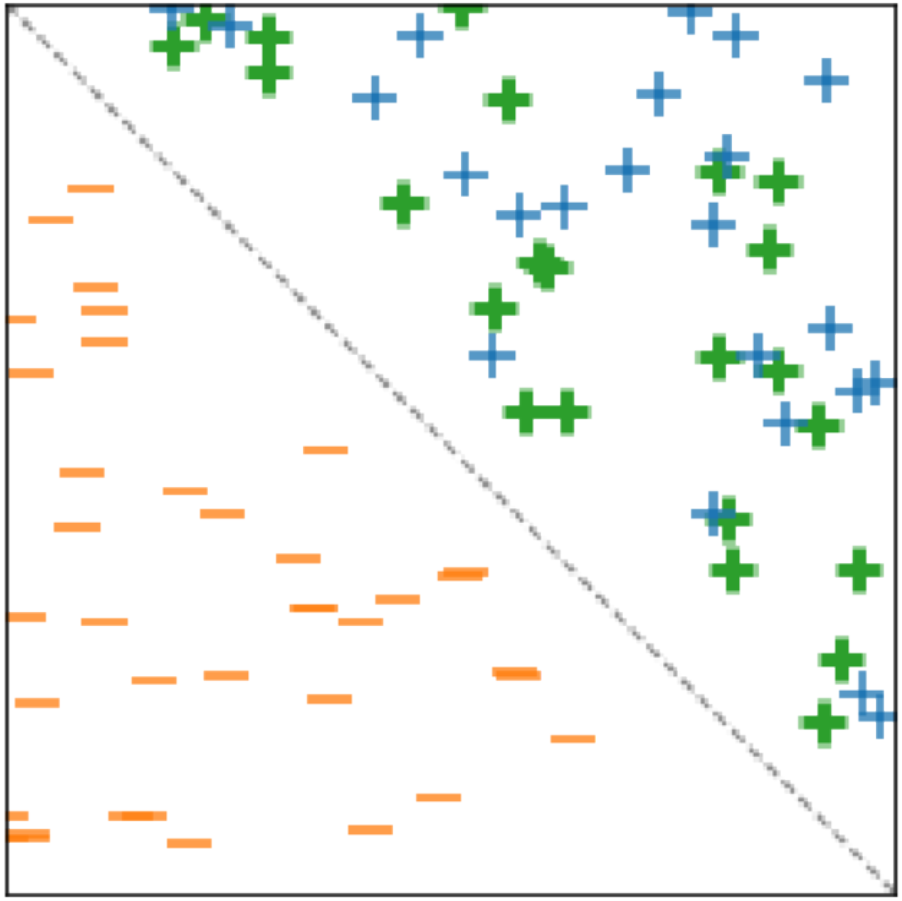
# Positive and Unlabeled Learning (PUL)

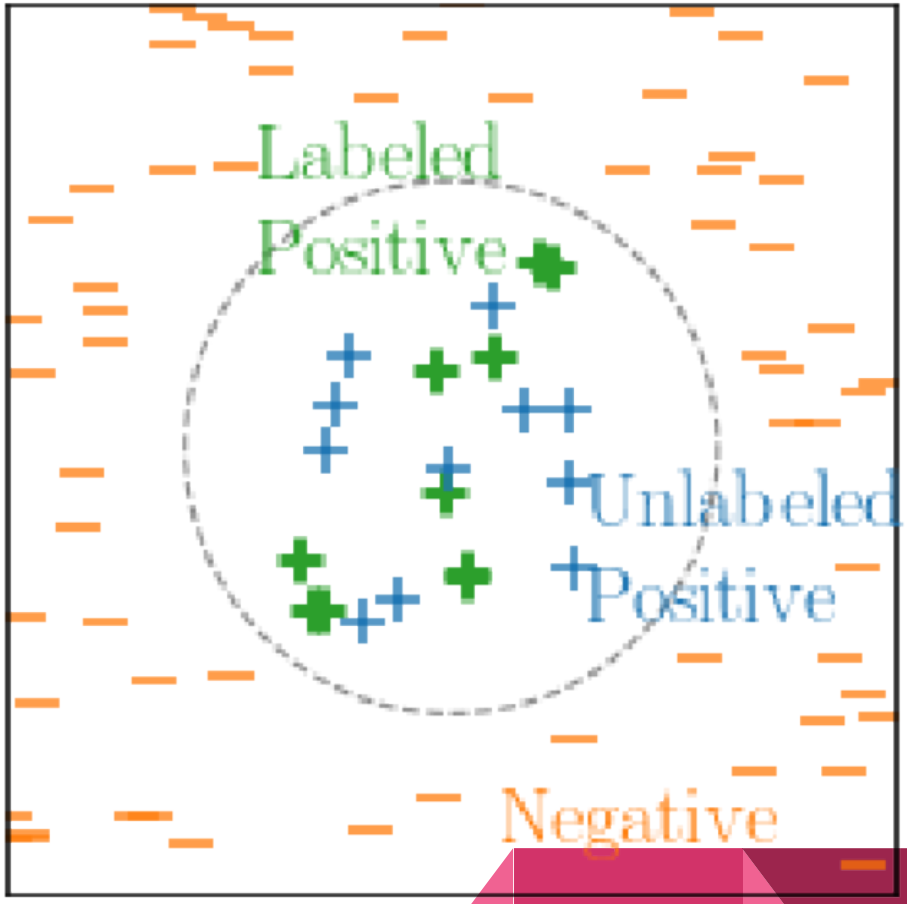
Uma possível estratégia alternativa é utilizar duas etapas

- Encontrar rótulos negativos “confiáveis”
- Constrói um classificador a partir deles e dos positivos
  - Ou faz um Label Propagation, por exemplo (PU-LP)



# Positive and Unlabeled Learning (PUL)

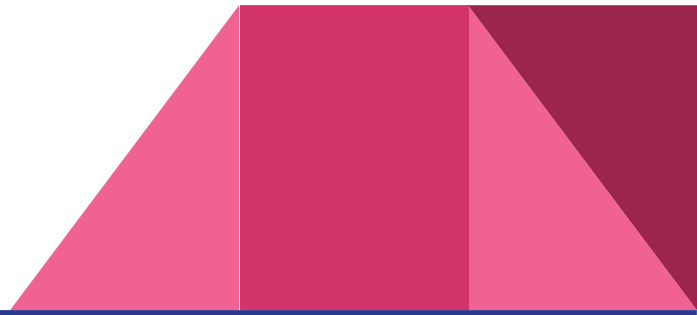





Bekker, J., Davis, J. Learning from positive and unlabeled data: a survey. Mach Learn 109, 719–760 (2020). https://doi.org/10.1007/s10994-020-05877-5

# Nota sobre avaliação

- Tem que conhecer o rótulo dos objetos de teste da classe positiva
- Desempenho em tarefa fim



# O que importa é ser positivo