

Aprendizado Semi-Supervisionado

1001513 – Aprendizado de Máquina 2
Turma A – 2022/2
Prof. Murilo Naldi



naldi@ufscar.br



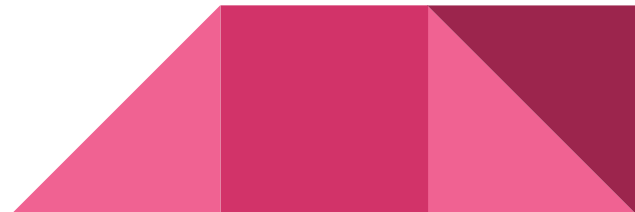
Agradecimentos

- Pessoas que colaboraram com a produção deste material: Diego Silva, Ricardo Campello, Ricardo Cerri, Moacir Ponti
- Intel IA Academy

Uma historinha pra começar a aula

Um professor e seu aluno quiseram fazer um *dataset*

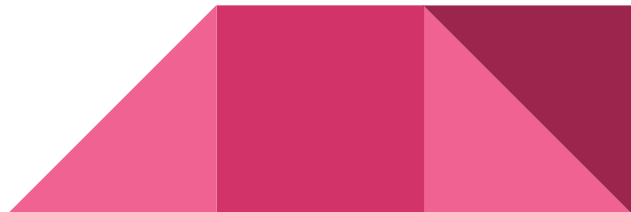
- ToLD-BR (<https://arxiv.org/abs/2010.04543>)
- Coletaram 10 milhões de *tweets*



Uma historinha pra começar a aula

Um professor e seu aluno quiseram fazer um *dataset*

- ToLD-BR (<https://arxiv.org/abs/2010.04543>)
- Coletaram 10 milhões de *tweets*
- Como rotular isso?



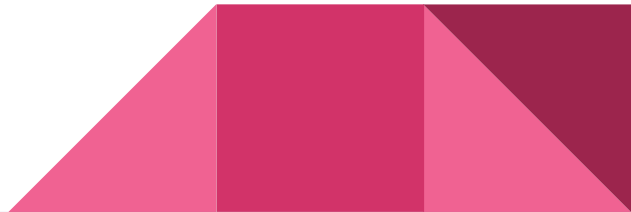
Uma historinha pra começar a aula



Uma historinha pra começar a aula

Um professor e seu aluno quiseram fazer um *dataset*

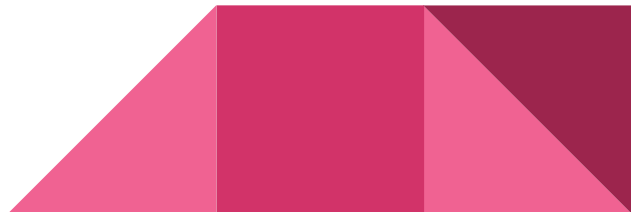
- ToLD-BR (<https://arxiv.org/abs/2010.04543>)
- Coletaram 10 milhões de *tweets*
- Conseguiram 21 mil exemplos rotulados
 - Mas e se...



Aprendizado semi-supervisionado

Grande volume de dados + poucos deles rotulados

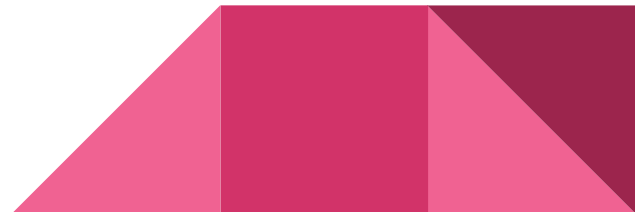
- Na verdade pode nem ser tão grande
- Pode ser classificação ou agrupamento
 - Ambos podem usar algumas informações sobre os rótulos
 - E o modelo resultante pode ser associado à criação de modelos de classificação



Aprendizado semi-supervisionado

Aprendizado semi-supervisionado é a parte de aprendizado de máquina que combina inferência de rótulos (aprendizado supervisionado) a partir da forma em que os dados são estruturados (aprendizado não-supervisionado)

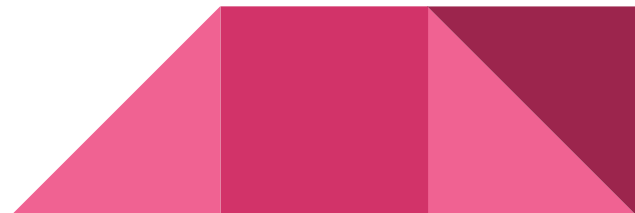
- Mistura de um pouco dos dois

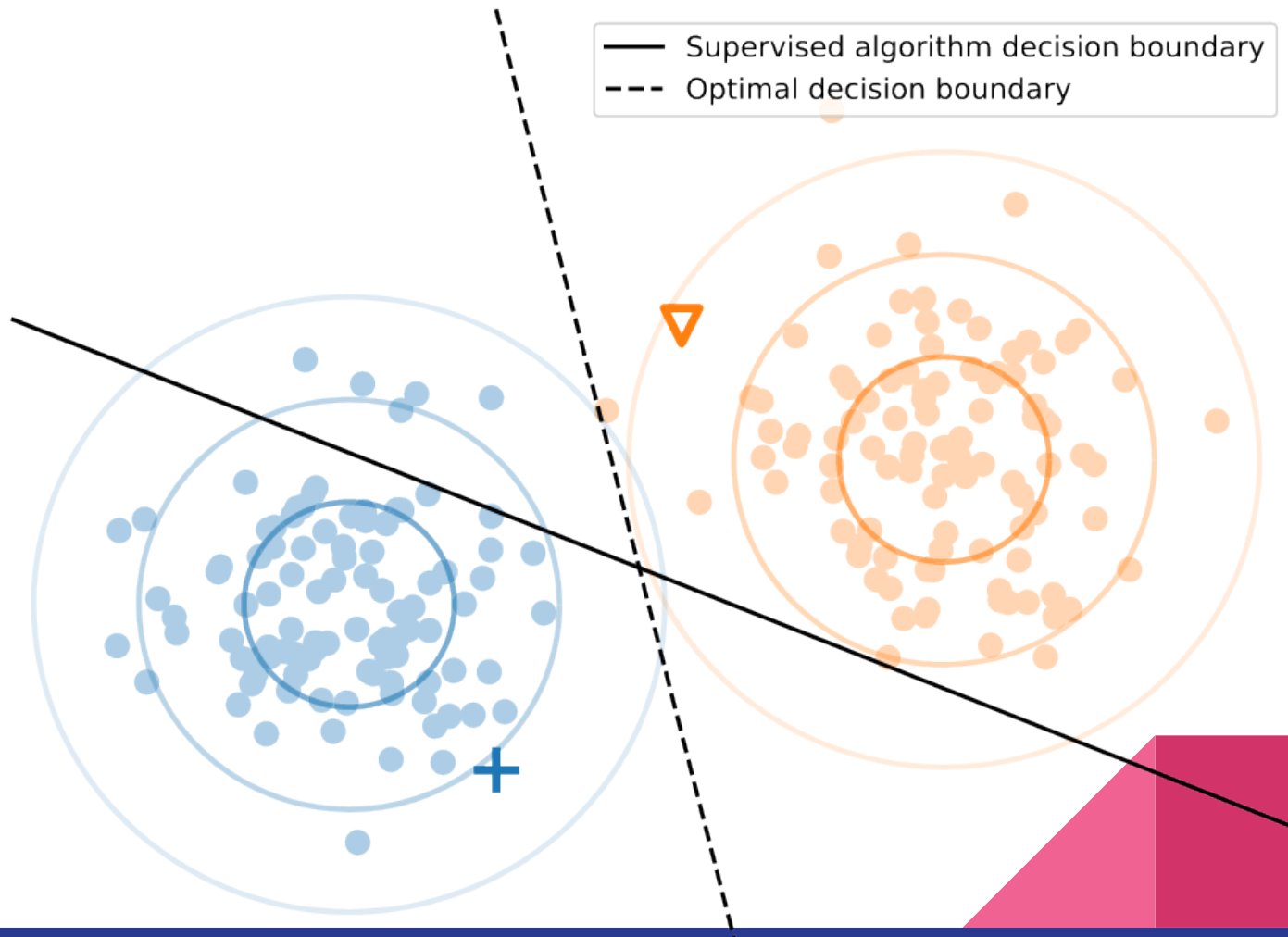


Aprendizado semi-supervisionado

A maior parte dos trabalhos é focada em classificação semi-supervisionada

- Onde dados sem rótulos são usados para melhorar o resultado de um classificador
 - Melhoram a percepção do fronteira de decisão

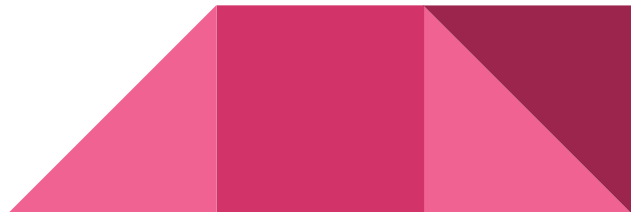




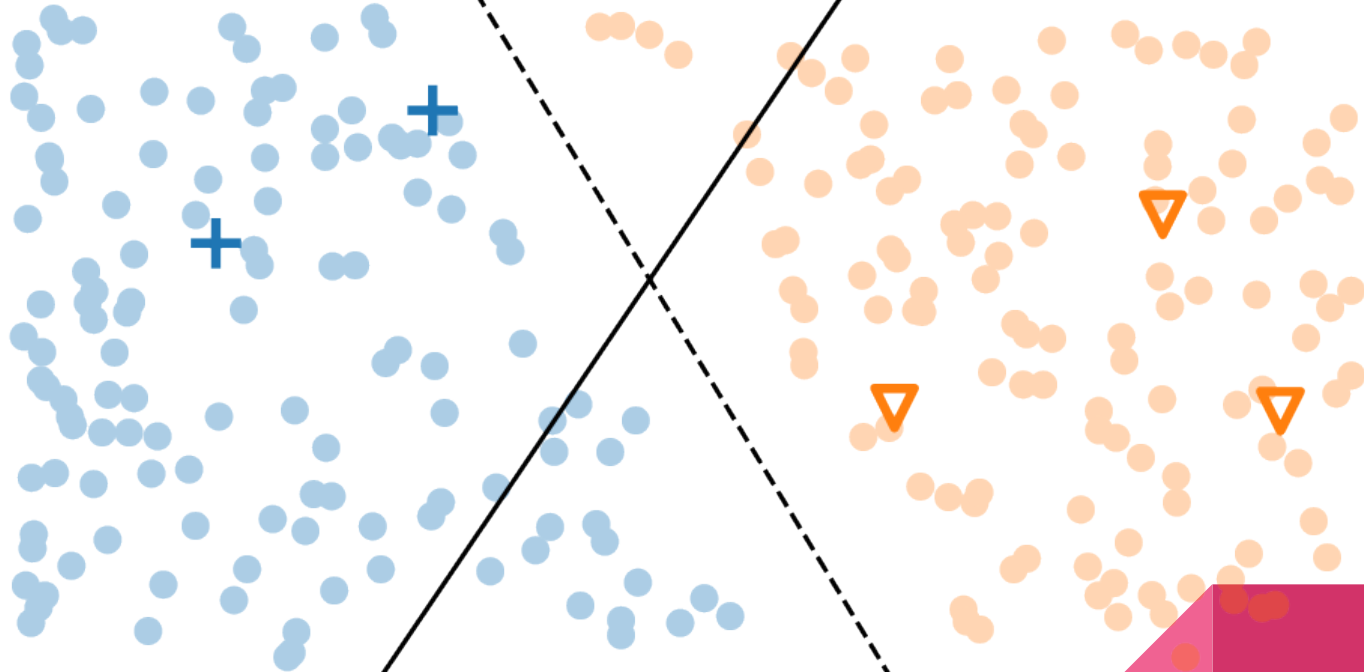
Suposições

Algumas suposições são importantes :

- Suposição de suavidade: um objeto próximo de um objeto rotulado tende a possuir o mesmo rótulo
- Suposição de baixa densidade: a fronteira de decisão deve passar por uma região de baixa densidade de dados



— Supervised algorithm decision boundary
- - - Optimal decision boundary

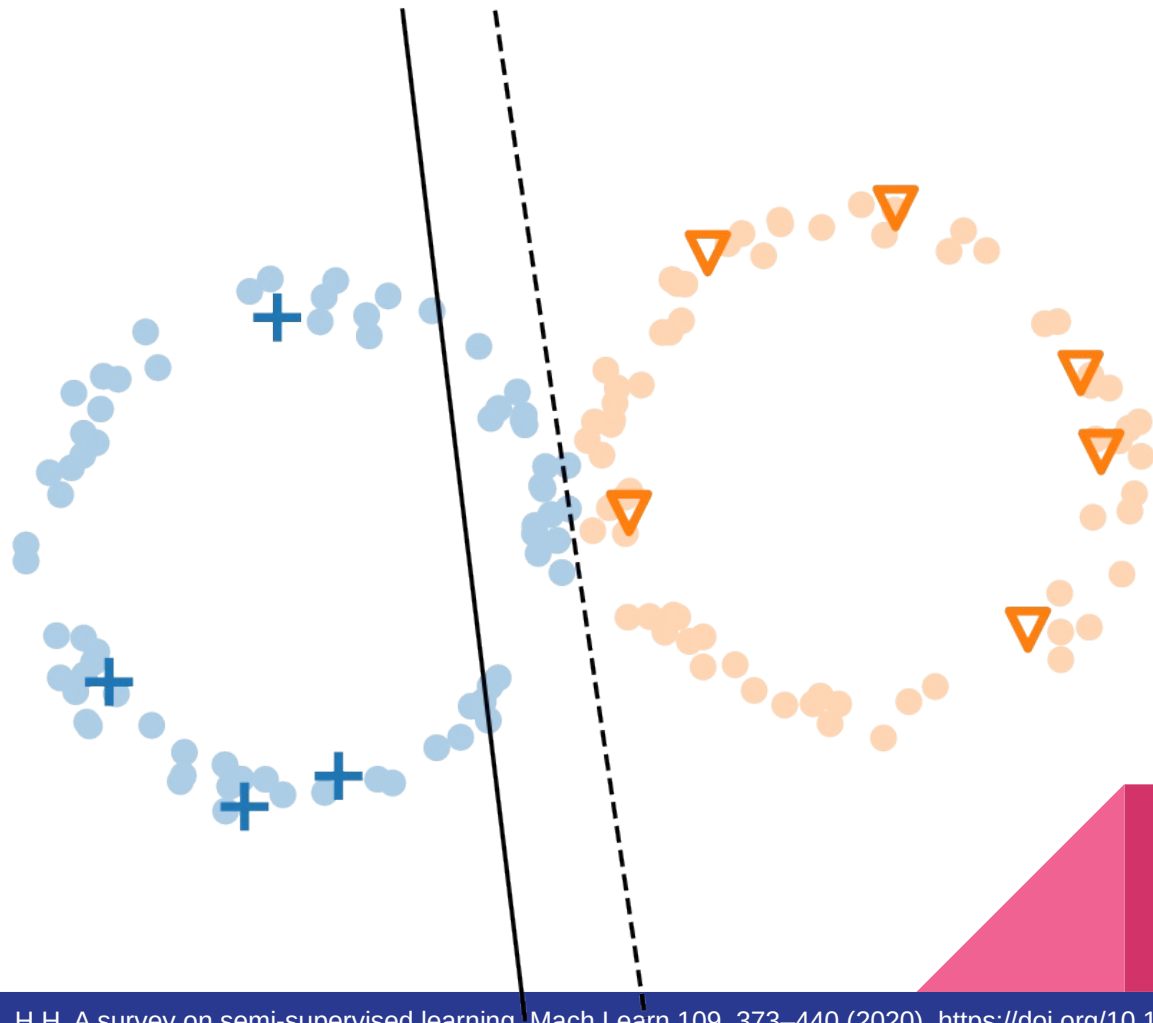


Suposições

Algumas suposições são importantes :

- Suposição de suavidade: um objeto próximo de um objeto rotulado tende a possuir o mesmo rótulo
- Suposição de baixa densidade: a fronteira de decisão deve passar por uma região de baixa densidade de dados
- Suposição de variedade: as classes estão estruturadas um espaço topológico que se parece localmente com um espaço euclidiano nas vizinhanças de cada ponto (variedade)
 - Exemplo: duas esferas podem ser divididas em uma variedade de círculos

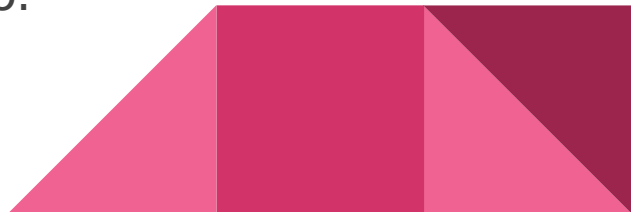




Conexão com agrupamento

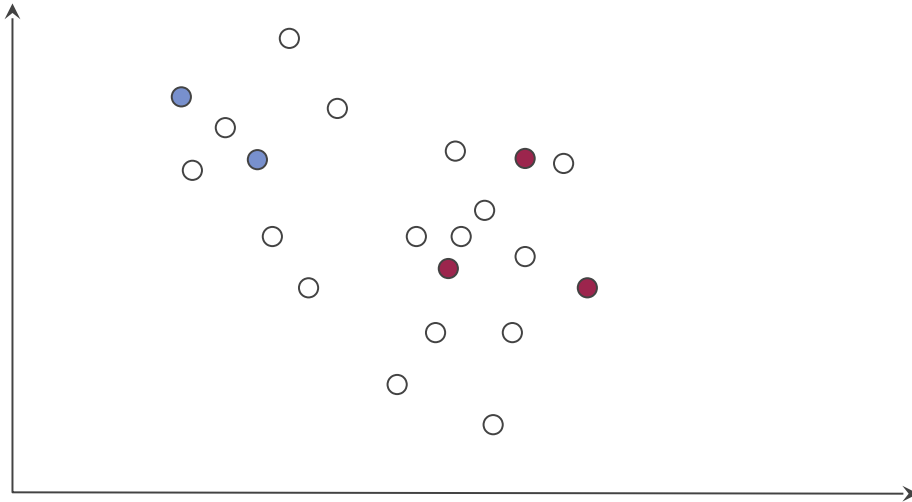
As suposições anteriores podem ser generalizadas como a “*suposição de agrupamento*”, ou seja, que os dados e suas classes se organizam como grupos:

- Conceito de grupo por similaridade
- Se os dados (não rotulados e rotulados) não puderem ser agrupados, não é possível que um método de aprendizado semi-supervisionado possa melhorar o resultado em relação a um método de aprendizado supervisionado.



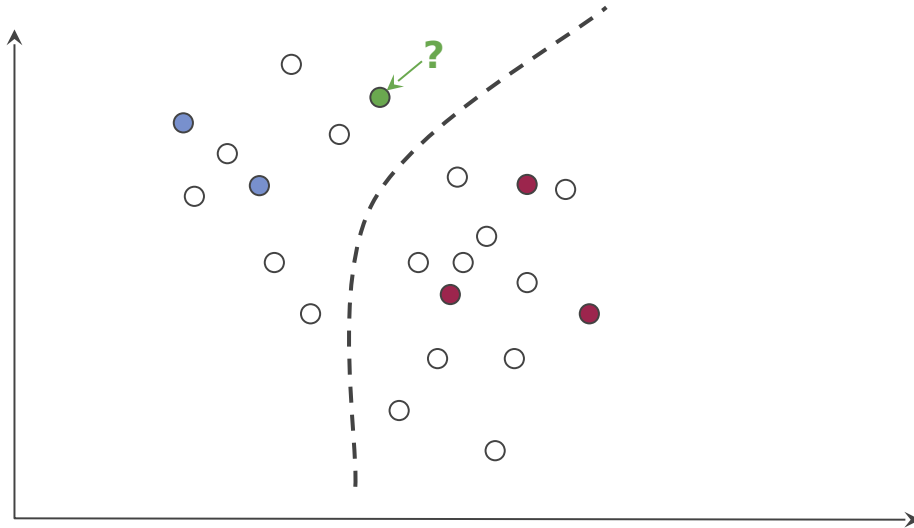
Aprendizado semi-supervisionado

Aprendizado (semi-supervisionado) indutivo



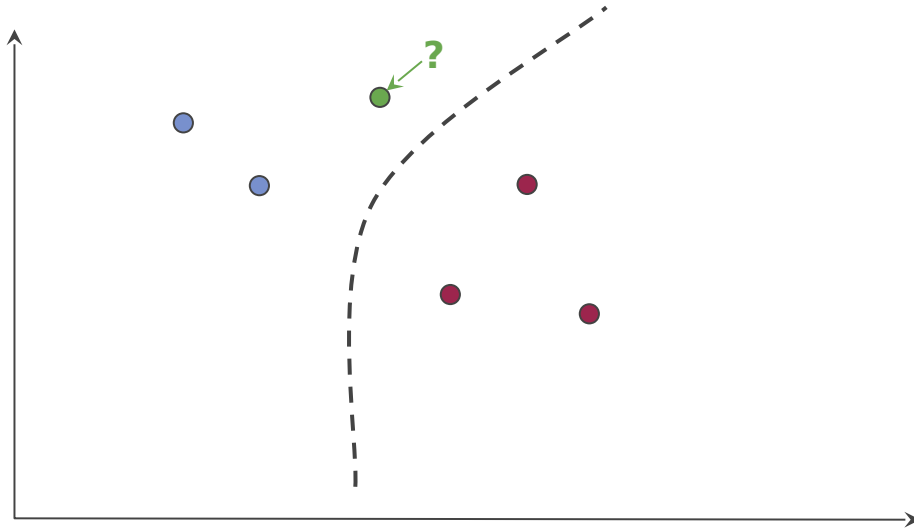
Aprendizado semi-supervisionado

Aprendizado (semi-supervisionado) indutivo



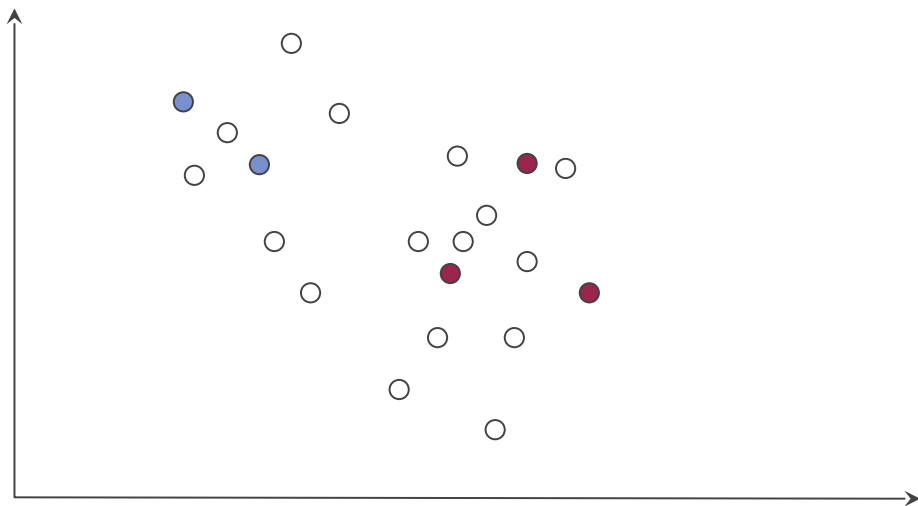
Aprendizado semi-supervisionado

Aprendizado (semi-supervisionado) indutivo



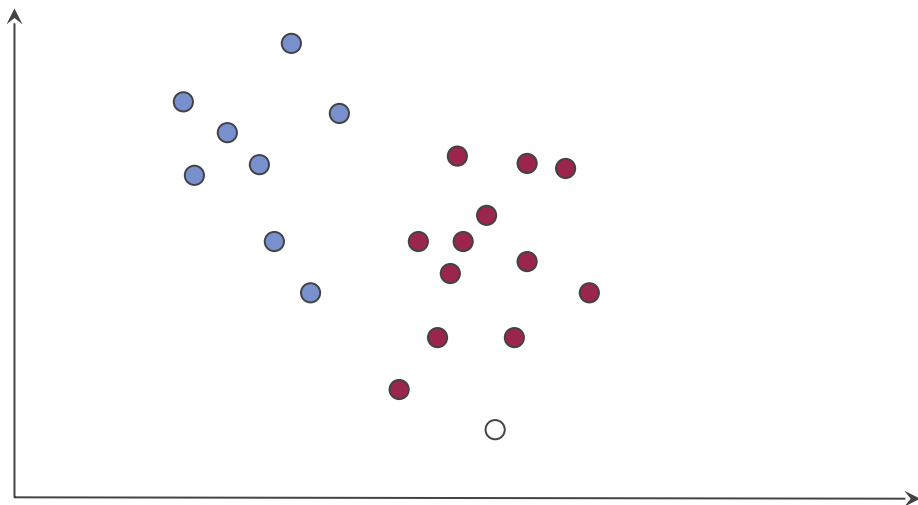
Aprendizado semi-supervisionado

Aprendizado (semi-supervisionado) transdutivo



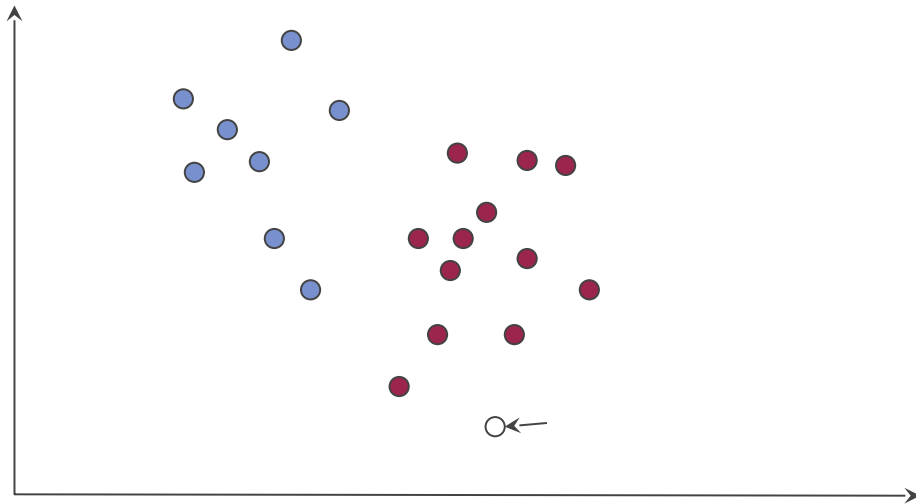
Aprendizado semi-supervisionado

Aprendizado (semi-supervisionado) transdutivo



Aprendizado semi-supervisionado

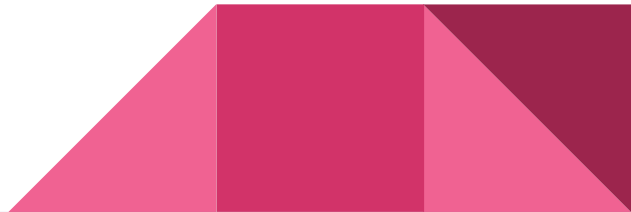
Aprendizado (semi-supervisionado) transdutivo



Label propagation

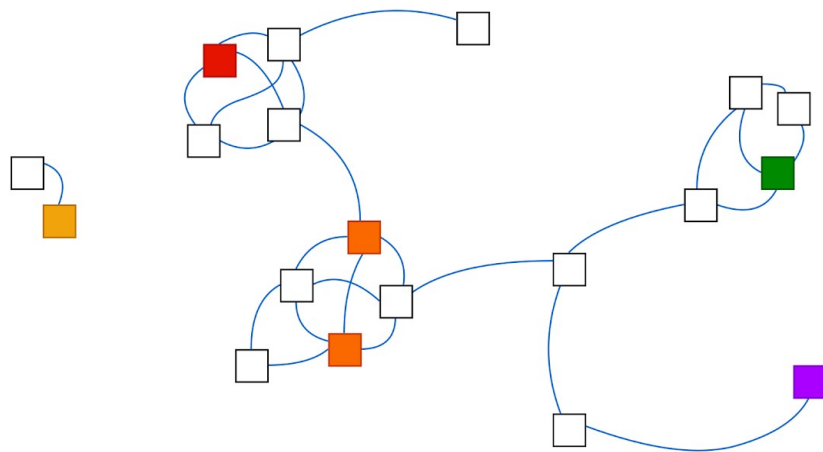
Dado um grafo,

- 1) Cada nós tem seu rótulo correspondente
- 2) O rótulo denota a comunidade à qual esse nó pertence
- 3) Através da iteração, cada nó atualizará seu rótulo com base nos rótulos dos nós vizinhos
 - 1) O rótulo atualizado de cada nó será o mais presente dentre os vizinhos do nó
- 4) Eventualmente, nós densamente conectados alcançam uma comunidade de rótulos comum



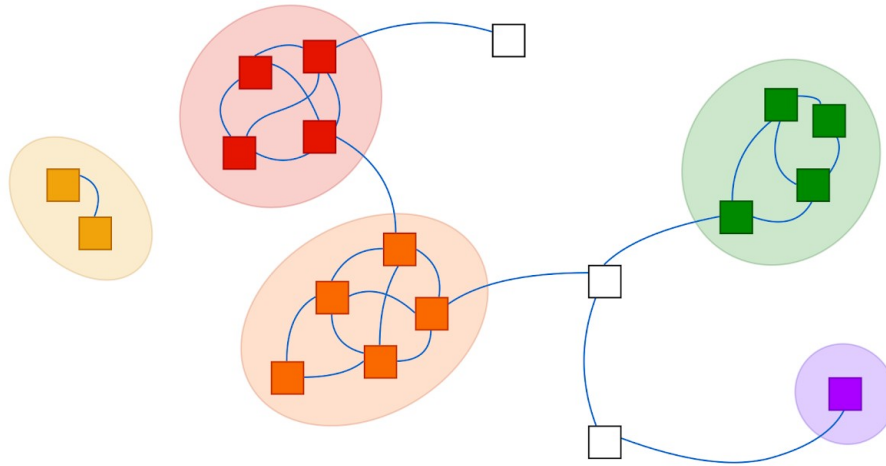
Label propagation

Vamos começar pelo conceito geral



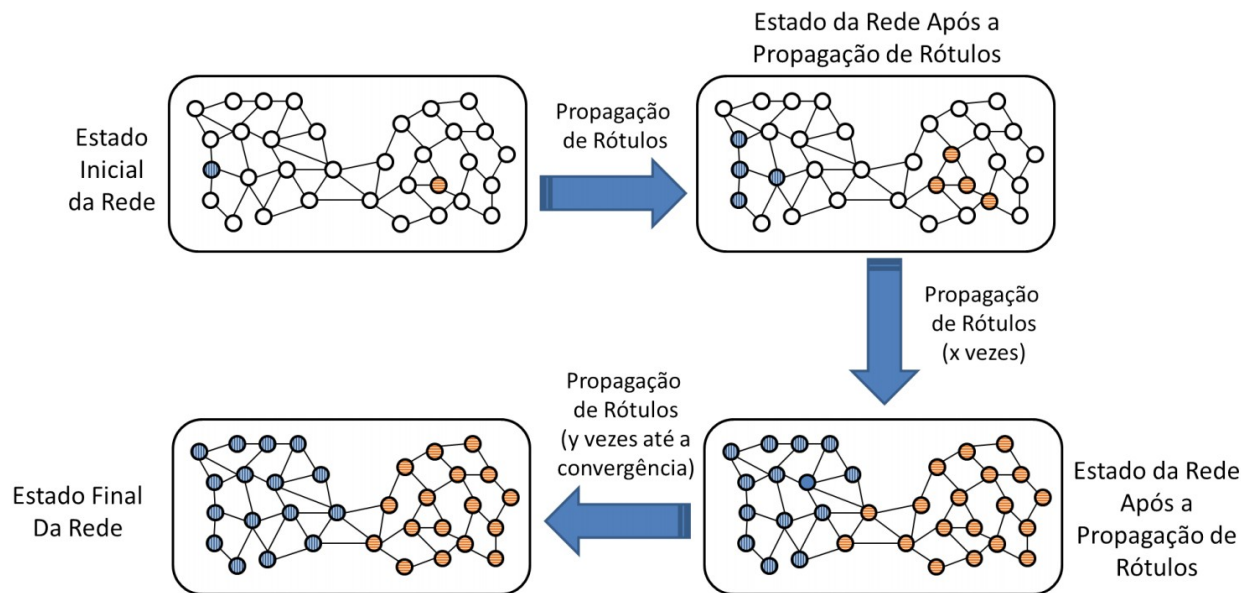
Label propagation

Vamos começar pelo conceito geral



Label propagation

Vamos começar pelo conceito geral

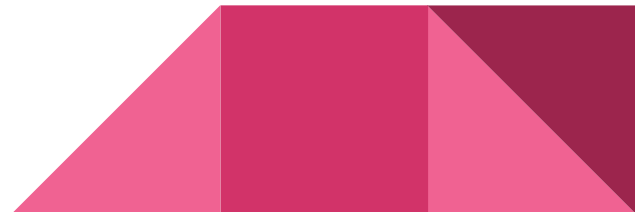


https://www.teses.usp.br/teses/disponiveis/55/55134/tde-05042016-105648/publico/VersaoRevisada_RafaelGeraldiniRossi.pdf

Label propagation

Construção do grafo – Gaussiana ou RBF sobre grafo completo

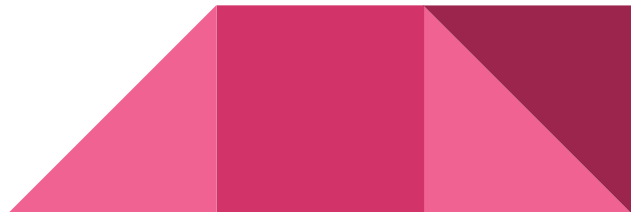
- Faz-se o grafo completo
- Ajusta-se Funções Gaussianas ou RBF nos dados
 - Arestas são ponderadas de acordo com a distribuição
- Aplica-se label propagation



Label propagation

Construção do grafo – k NNG

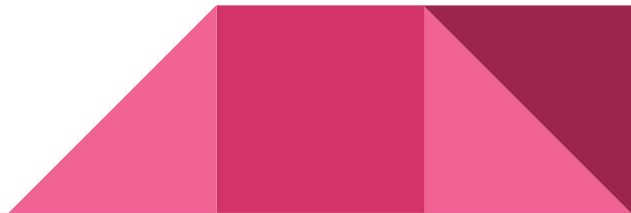
- Dado um valor de k , conecta-se os k -vizinhos mais próximos
- Depois label propagation
- Alternativamente, ϵ NNG, onde ϵ define uma dissimilaridade limite para conectar vizinhos (constante)



Label propagation

Construção do grafo – k MNNG (mútuo)

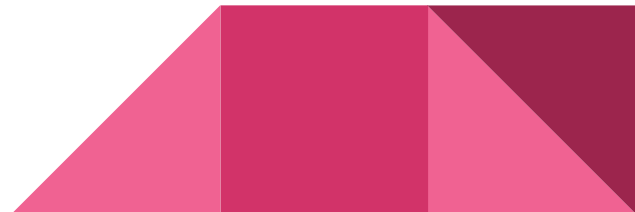
- Mesma ideia do k NNG, porém as conexões só ocorrem entre k vizinhos mais próximos que seja mútuos
 - Tende a gera menos *hubs*
- Alternativamente, também possui versão ϵ MNNG



Label propagation

Algoritmo

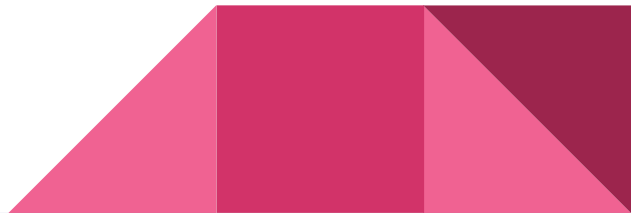
- l e u são o número de exemplos rotulados e não rotulados
- Y é uma matriz $(l+u) \times C$ com a distribuição de probabilidade dos labels



Label propagation

Algoritmo

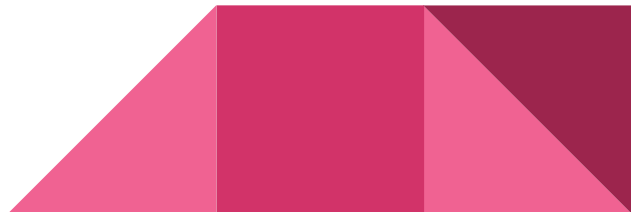
- Definimos T , uma matriz de prob. de transição $(l+u) \times (l+u)$



Label propagation

Algoritmo

1. Propagamos os rótulos: $Y \leftarrow TY$
2. Normalizamos Y (por linha)
3. Asseguramos o rótulo dos inicialmente rotulados



Label propagation

Detalhes

- Convergência
- Parâmetro σ (RBF)
- Rebalanceamento das classes

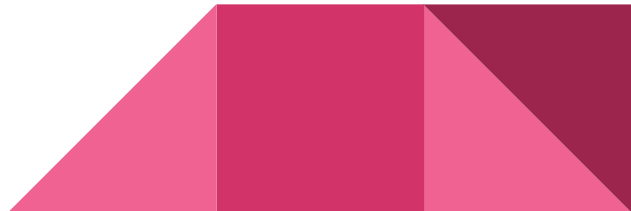
<http://pages.cs.wisc.edu/~jerryzhu/pub/CMU-CALD-02-107.pdf>

- Variações: GFHF e LLGC

Agrupamento Semi-supervisionado

A forma que os rótulos são aplicados no agrupamento é diferente da forma da tarefa de classificação

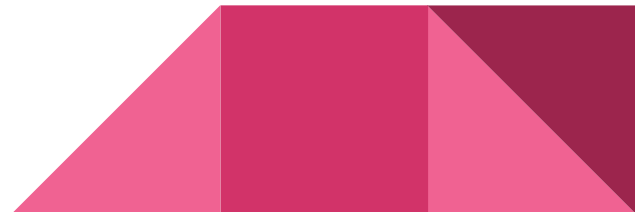
- Na classificação os rótulos são usados para definir rótulos dos objetos não rotulados por transdução e melhorar a indução do modelo
- No agrupamento os rótulos servem para definir “o grupo” do objeto e só faz sentido se houver dois ou mais (porque?)



Agrupamento Semi-supervisionado

Um objeto rotulado “define” o rótulo do grupo, portanto:

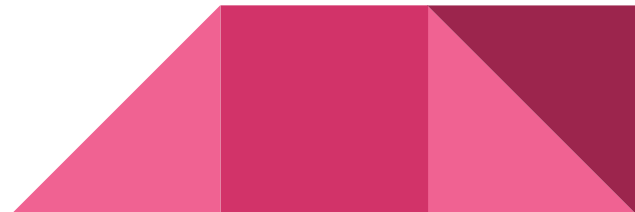
- Um grupo não deve ter dois ou mais objetos com rótulos distintos
- Um grupo deve possuir todos os objetos que possuem o mesmo rótulo



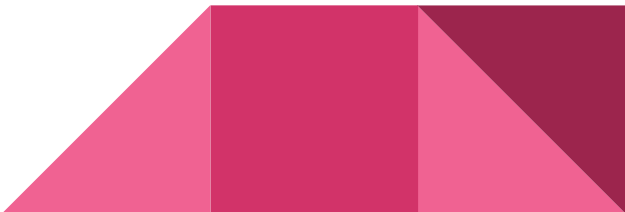
Agrupamento Semi-supervisionado

Em outras palavras:

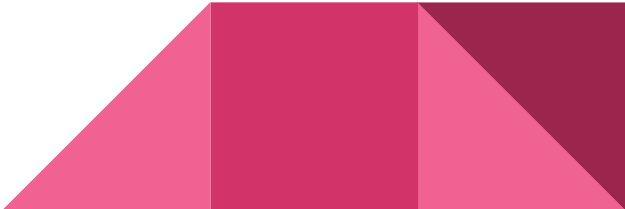
- Os rótulos servem para definir relações *must-link* e *cannot-link* no processo de agrupamento
- Os algoritmos devem ser adaptados para respeitar essas restrições durante a construção do modelo



Exemplo: k -médias

- 1) Escolher um número k de protótipos (centros) para os grupos
 - 2) Atribuir cada objeto para o grupo de centro mais próximo (segundo alguma distância, e.g. Euclidiana)
 - 3) Mover cada centro para a média (centróide) dos objetos do grupo correspondente
 - 4) Repetir os passos 2 e 3 até que algum critério de convergência seja obtido
- 

Exemplo: k -médias com restrições

- 1) Escolher um número k de protótipos (centros) para os grupos
 - 2) Atribuir cada objeto para o grupo de centro mais próximo, obedecendo a lista de *cannot-link*
 - 3) Mover cada centro para a média (centróide) dos objetos do grupo correspondente
 - 4) Repetir os passos 2 e 3 até que algum critério de convergência seja obtido
 - 5) Aglomerar grupos com objetos *must-link*
- 

Isso vai looooooooooooooonge

LIGUE OS PONTOS NA ORDEM CORRETA,
E AJUDE A ARIEL A DESCOBRIR UM NOVO AMIGO!

