Aprendizado Semi-supervisionado (SSL)

Elaine C. R. Cândido

Pontifícia Universidade Católica de Minas Gerais

Abril 2021

Programa do Curso

1. Contextualização
   - O que é?
   - Definição do problema
2. Suposições
   - Smoothness
   - Cluster
   - Manifold
3. Algoritmos e Prática
   - Generative Gaussian Mixture
   - Self-training
   - Co-Training

Pré-requisitos

- Lógica de programação
- Programação em Python
- Nível do curso: intermediário
- Conhecimento em aprendizado de máquina

Créditos

- Python Data Science Handbook & Mastering Machine Learning Algorithms

Visão Geral

1. Contextualização do Aprendizado Semi-supervisionado
2. Fundamentos do Aprendizado Semi-supervisionado
3. Algoritmos e Prática

Definição

- Tenta resolver problemas que incluem dados rotulados e não rotulados, empregando conceitos que incluem características de métodos de agrupamento e classificação
- A existência de dados em sua grande maioria sem rótulos e a dificuldade de rotulá-los trouxe a necessidade de investigar melhores técnicas que permitem extender o conhecimento adquirido pelos dados rotulados para uma população de dados maior sem rótulos

O cenário semi-supervisionado

- Tem-se limitado número N de instâncias com rótulos
- O contexto de SSL é definido pela união dos dados com rótulos e sem rótulos
- A distribuição dos dados sem rótulos é dada como bem semelhante aos dados com rótulos
- SSL é usada quando o número de instâncias sem rótulos é bem maior do que as com rótulos

O cenário semi-supervisionado

- Se o conhecimento de X_u aumenta o conhecimento sobre a distribuição dos dados, então uma técnica semi-supervisionada irá funcionar
- Se os dados sem rótulos possuem distribuição diferente ou diferentes do que passou como dado de treinamento, então o uso de técnicas semi-supervisionadas podem levar a resultados piores

Cenários Causais

- Processo gera Y como um **efeito** da **causa** X

Cenários Causais

- Processo gera Y como um **efeito** da **causa** X
  - Podemos dizer que o conhecimento de $p(x)$ aumenta o conhecimento de $p(y|x)$?

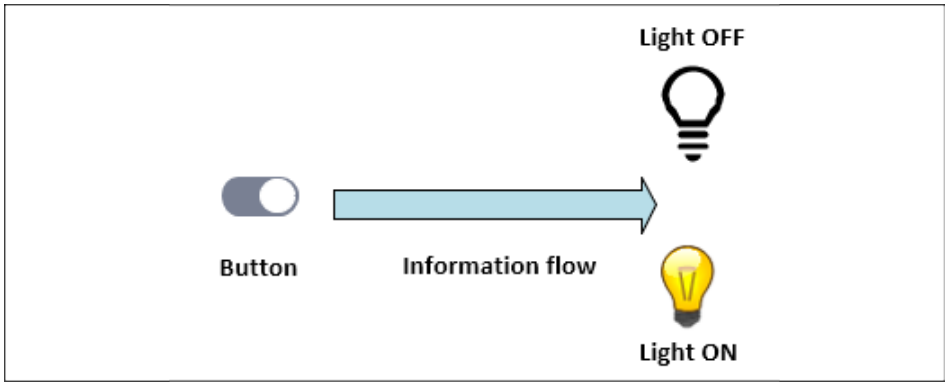
Cenários Causais

- Processo gera Y como um **efeito** da **causa** X
  - Podemos dizer que o conhecimento de $p(x)$ aumenta o conhecimento de $p(y|x)$?
  - Se foi modelado a distribuição condicional dos efeitos dado um conjunto de casos, então toda a informação necessária para decidir qual causa é a mais provável já está codificada no modelo

Cenários Causais

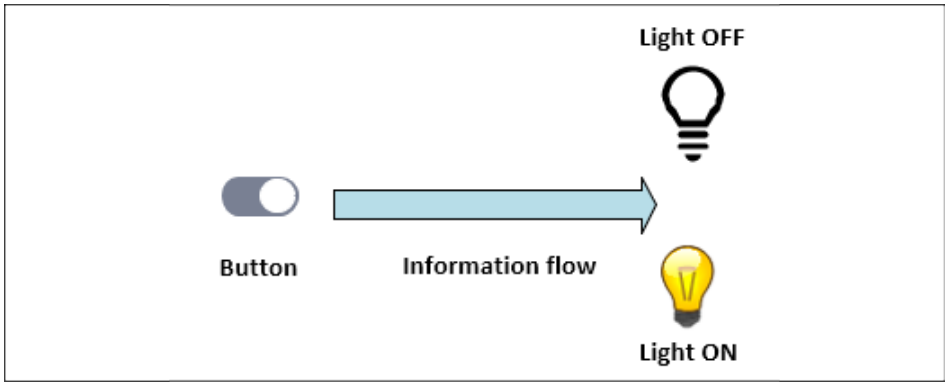
- Processo gera Y como um **efeito** da **causa** X
  - Podemos dizer que o conhecimento de $p(x)$ aumenta o conhecimento de $p(y|x)$?
  - Se foi modelado a distribuição condicional dos efeitos dado um conjunto de casos, então toda a informação necessária para decidir qual causa é a mais provável já está codificada no modelo
  - O conhecimento adicional de $p(x)$ não terá nenhum impacto de selecionar um efeito ao invés de outro, pois todo o processo é governado somente pelo conhecimento dos dados em X que trigaram todos os efeitos de Y durante o treinamento

Cenários Causais



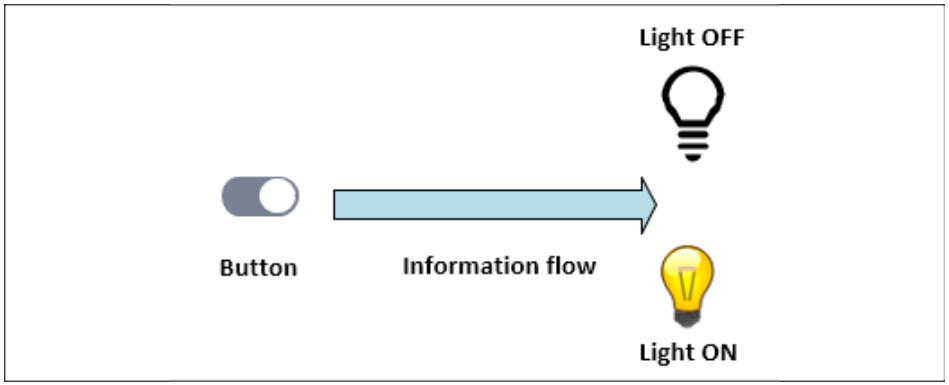
- $button \rightarrow light$.
- Training set contendo N observações
- $p(light|button = ON)$
- Se alguém diz que sabe a probabilidade exata de $p(button = ON)$?

Cenários Causais



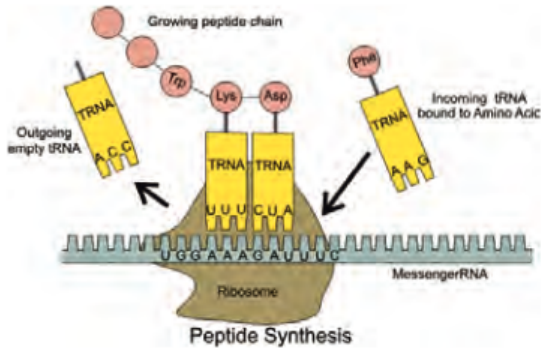
- $button \rightarrow light$.
- Training set contendo N observações
- $p(light|button = ON)$
- Se alguém diz que sabe a probabilidade exata de $p(button = ON)$?
- O status do botão já determina o status da luz (cenário causal)

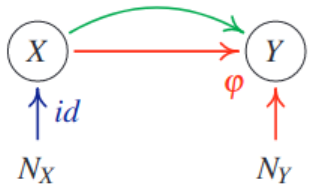
Cenários Causais

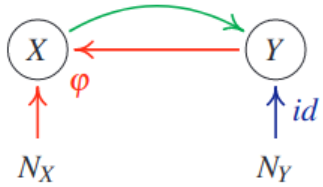


- $button \rightarrow light$.
- Training set contendo N observações
- $p(light|button = ON)$
- Se alguém diz que sabe a probabilidade exata de $p(button = ON)$?
- O status do botão já determina o status da luz (cenário causal)
- Saber a probabilidade do evento não vai mudar o conhecimento de $p(light|button = ON)$

Cenários Causais

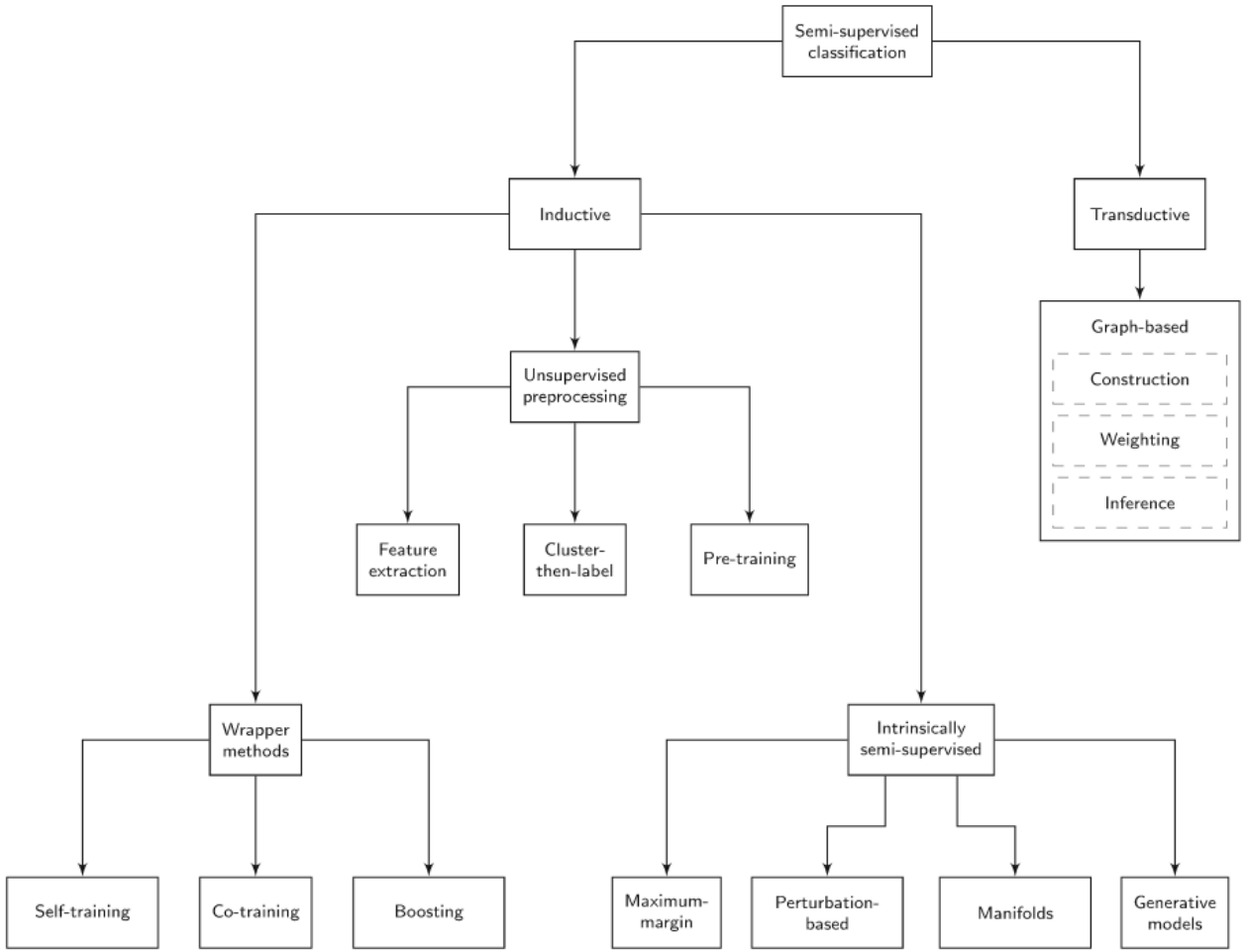






- Predizer a proteína do mRNA é um exemplo de aprendizado causal, onde a direção da predição (seta verde) é a mesma direção da causa (seta vermelha)
- No reconhecimento de dígitos, se tenta inferir o rótulo da classe da imagem produzida por um escritor, que se trata de um cenário anti-causal

Taxonomia de SSL



Transductive vs Inductive Learning

Em aprendizado de máquina temos problemas como:

- Entrada:
  - Um conjunto U de exemplos com rótulos (x_i, y_i), onde todos x_i é o vetor de entrada e y_i é o rótulo de saída correspondente
  - Um conjunto V de exemplos x'_i sem rótulos
- Saída:
  - O conjunto de rótulos esperados y'_i para todas as instâncias em V

Transductive vs Inductive Learning

Em aprendizado de máquina temos problemas como:

- Entrada:
  - Um conjunto U de exemplos com rótulos (x_i, y_i), onde todos x_i é o vetor de entrada e y_i é o rótulo de saída correspondente
  - Um conjunto V de exemplos x'_i sem rótulos
- Saída:
  - O conjunto de rótulos esperados y'_i para todas as instâncias em V
- Existem duas formar que podemos usar para resolver este problema:
  - Inductive
  - Transductive

Inductive Learning

- Considera todas as instâncias X e tenta determinar $p(x|y)$ ou uma função $y = f(x)$ que mapeia instâncias com rótulos e sem rótulos ao seus correspondentes rótulos
- Cria um modelo: regras e propriedades são induzidas através dos dados

Inductive Learning

- Considera todas as instâncias X e tenta determinar $p(x|y)$ ou uma função $y = f(x)$ que mapeia instâncias com rótulos e sem rótulos ao seus correspondentes rótulos
- Cria um modelo: regras e propriedades são induzidas através dos dados
- É a tentativa de descoberta de regras e generalizações baseado na análise dos dados coletados.
  - **Tentativa**, pois as generalizações não são fatos, mas sim aproximações baseadas na evidência que foi coletada.

Inductive Learning

- No exemplo anterior:

Inductive Learning

- No exemplo anterior:
  - Entrada: U para aprendizado supervisionado e V se for usado aprendizado semi-supervisionado

Inductive Learning

- No exemplo anterior:
  - Entrada: U para aprendizado supervisionado e V se for usado aprendizado semi-supervisionado
  - Saída: Um conjunto de rótulos y'_i para todas as instâncias em V

Inductive Learning

- No exemplo anterior:
  - Entrada: U para aprendizado supervisionado e V se for usado aprendizado semi-supervisionado
  - Saída: Um conjunto de rótulos y_i' para todas as instâncias em V
  - Técnica: Computar uma função f, usando a informação em U, tal que $f(x_i)$ fique o mais próximo possível de y_i em todas as instâncias de U.

$$y_i' = f(x_i')$$

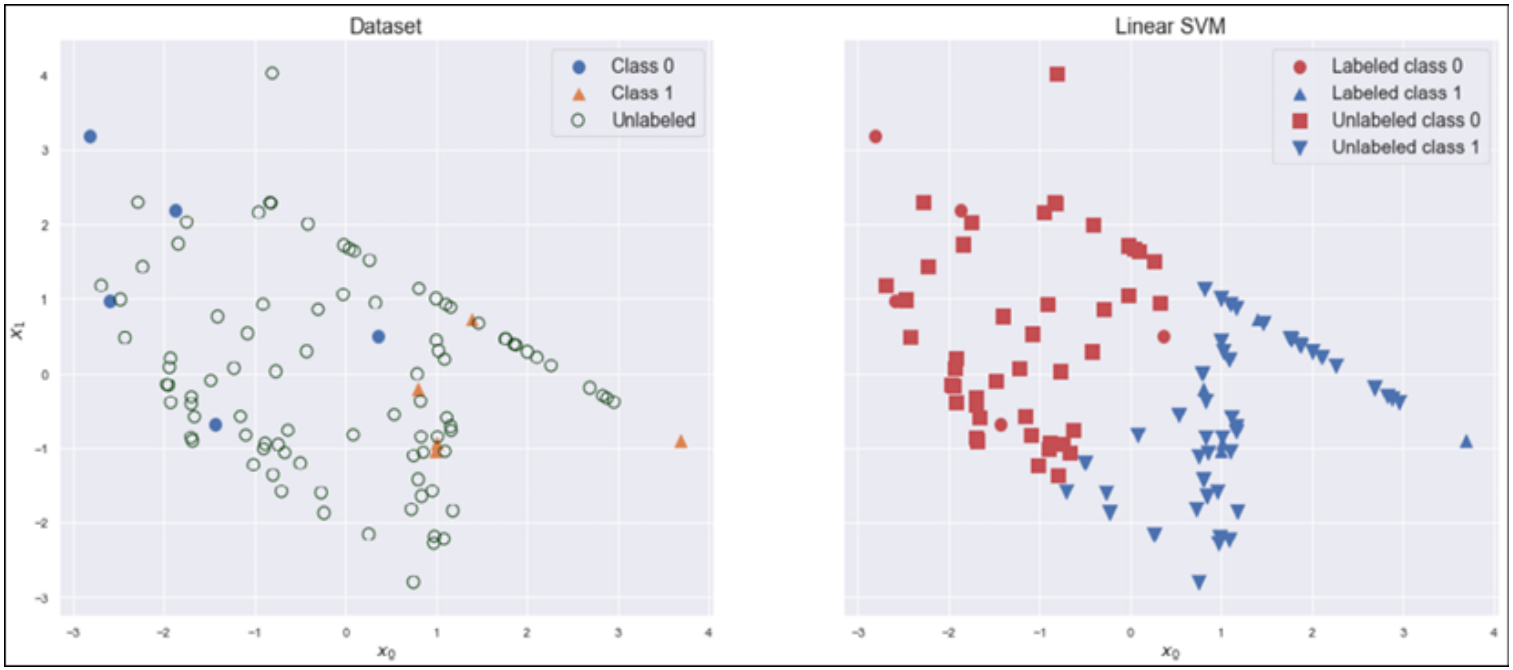
Transductive Learning

- Introduzido por Vladmir Vapnik, com o pensamento:

  "When trying to solve some problem, one should not solve a more difficult problem as an intermediate step."

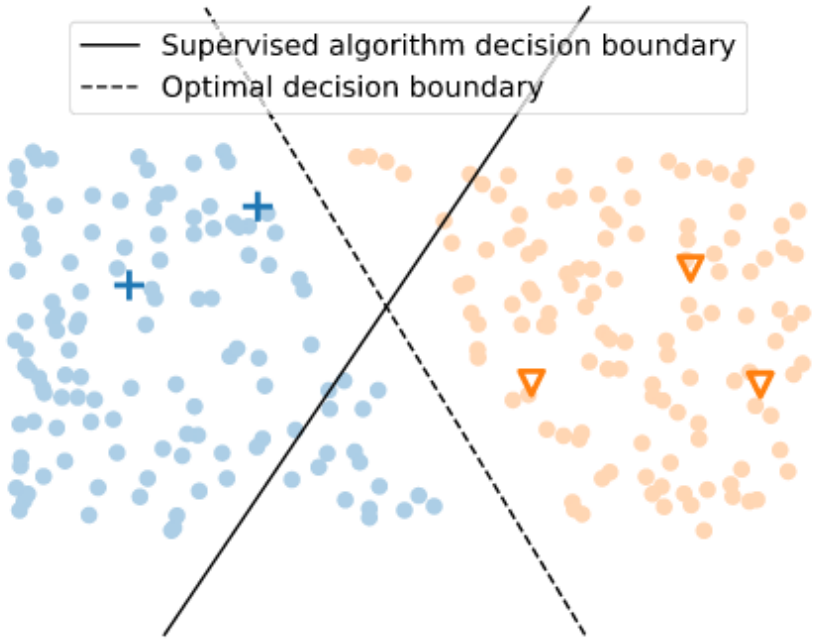
- Mais complexo e requer maior tempo computacional
- Principal desvantagem:
  - Retreinar todo novo conjunto de instâncias

Transductive Learning



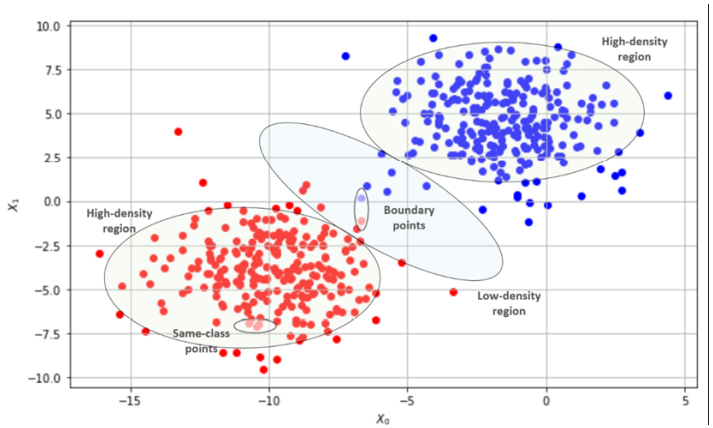
SSL Assumptions - Smoothness

- Para dois pontos x e $x' \in X$ que estão próximos no espaço, seus rótulos deveriam ser os mesmos.



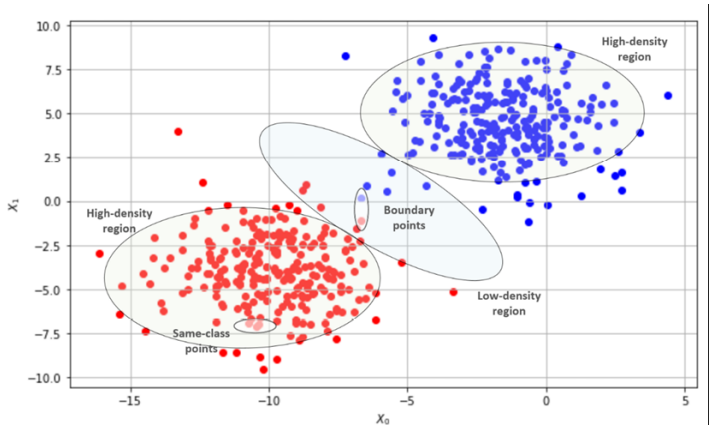
SSL Assumptions - Cluster or Low-Density

- A fronteira de decisão deve, preferencialmente, passar em uma região de pouca densidade.



SSL Assumptions - Cluster or Low-Density

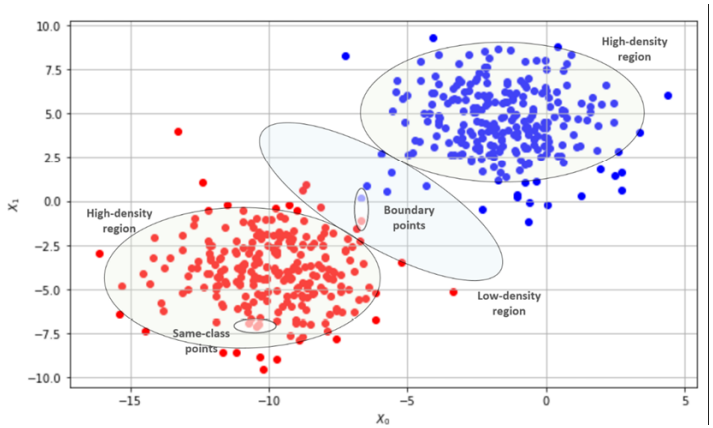
- A fronteira de decisão deve, preferencialmente, passar em uma região de pouca densidade.



- Se smoothness é verdadeiro, então qualquer dois pontos próximos tem o mesmo rótulo

SSL Assumptions - Cluster or Low-Density

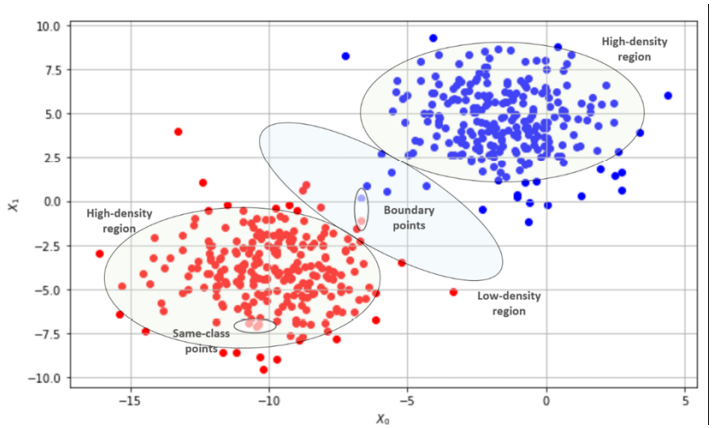
- A fronteira de decisão deve, preferencialmente, passar em uma região de pouca densidade.



- Se smoothness é verdadeiro, então qualquer dois pontos próximos tem o mesmo rótulo
- Em qualquer área desamente populada, espera-se que os pontos tenham mesmo rótulos

SSL Assumptions - Cluster or Low-Density

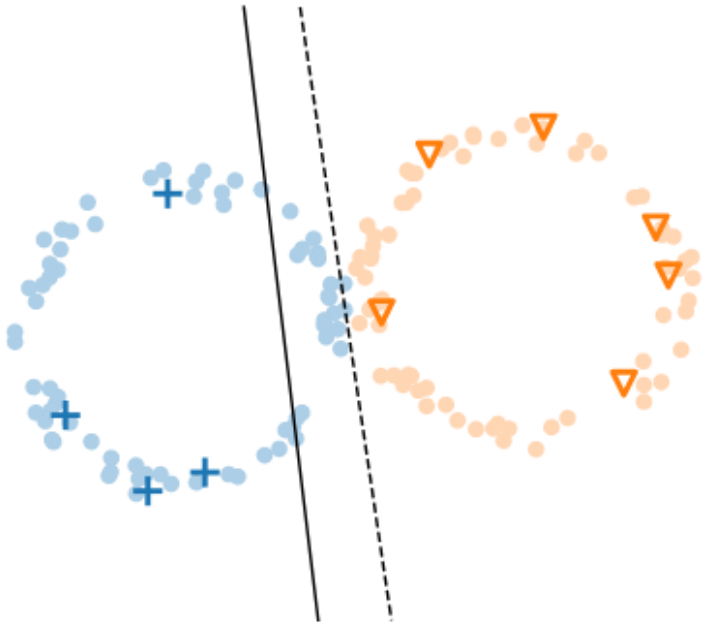
- A fronteira de decisão deve, preferencialmente, passar em uma região de pouca densidade.



- Se smoothness é verdadeiro, então qualquer dois pontos próximos tem o mesmo rótulo
- Em qualquer área desamente populada, espera-se que os pontos tenham mesmo rótulos
- Uma fronteira de decisão, pode então ser construtída passando pela região de baixa densidade (low-density)

SSL Assumptions - Manifold

- Se for possível determinar quals manifolds existem e quais pontos estão em cada manifold, os rótulos dos dados, previamente sem rótulos, podem ser inferidos dos dados com rótulos no mesmo manifold.



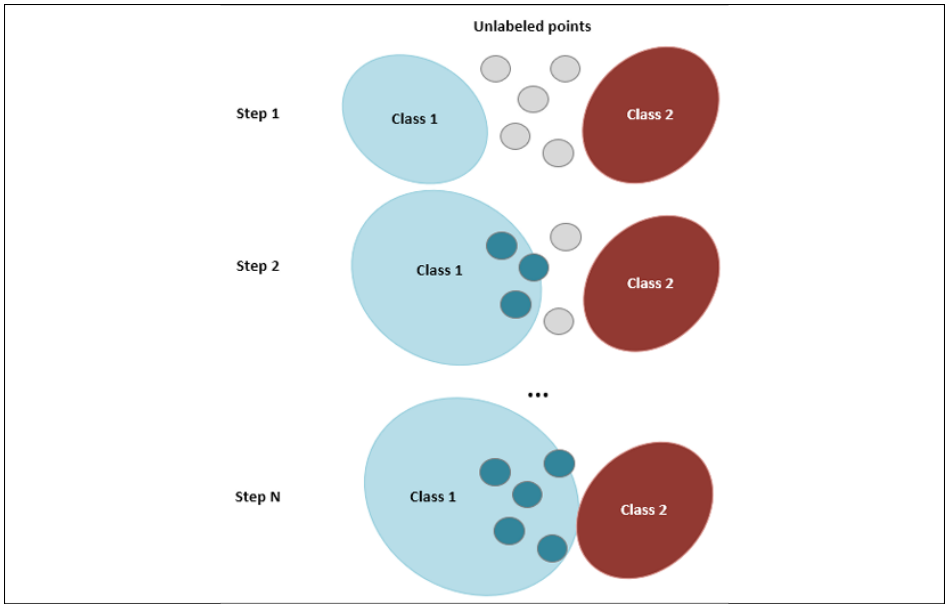
Generative Gaussian Mixture - GMM

- Algoritmo Indutivo (Inductive)
- Modela a probabilidade condicional $p(x, y)$
- Explica a estrutura dos dados existentes
- Retorna a probabilidade dos dados pertecerem as classes

Self Training

- Aplica smoothness e cluster assumptions
- Escolha válida quando os dados com rótulos contém informação suficiente
- Consiste de um classificador supervisionado que é iterativamente treinado em dados com rótulos e dados com pseudo-rótulos (rotulados em iterações anteriores do algoritmo)
  - No início o classificador é treinado apenas com dados rotulados, após o classificador faz a predição nos dados sem rótulo e então as predições mais confiantes são adicionadas aos dados de treino e o classificador é retreinado com dados de treino + dados pseudo-rotulados

Self Training



Co-training

- Efetivo quando os dados podem ser teoricamente classificados usando somente uma parte das features
- Extensão do self training
- Dois ou mais classificadores são iterativamente treinados nos dados de treino, adicionando as predições mais confidentes aos dados com rótulos a cada iteração
- *Disagreement-based methods*: Os base learners não podem ser muito fortemente correlacionados em suas predições

Referências

- Livros
  - Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf. Elements of Causal Inference Foundations and Learning Algorithms. The MIT Press, 2017 Disponível em: `https://library.oapen.org/bitstream/handle/20.500.12657/26040/11283.pdf?sequence=1&isAllowed=y`
  - VanderPlas, Jake. Python data science handbook: Essential tools for working with data. O'Reilly Media, Inc., 2016.

Referências

- Artigos:
  - Introduction to Semi-Supervised Learning. Disponível em:http://mitp-content-server.mit.edu:18180/books/content/sectbyfn?collid=books_pres_0&id=6173&fn=9780262033589_sch_0001.pdf
  - A small and easy introduction to Transductive Learning. Disponível em https://codesachin.wordpress.com/2016/07/03/a-small-and-easy-introduction-to-transductive-learning/
  - Jesper E. van Engelen and Holger H. Hoos. A survey on semi-supervised learning. Disponível em: https://link.springer.com/content/pdf/10.1007/s10994-019-05855-6.pdf