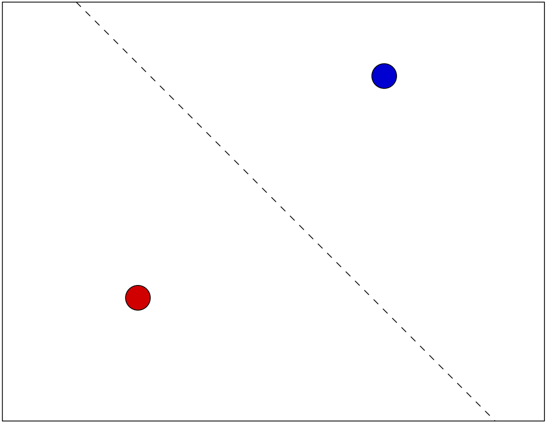
University of Pennsylvania

Graph Min-Cuts

Learning from Labeled and Unlabeled Data using Graph Mincuts

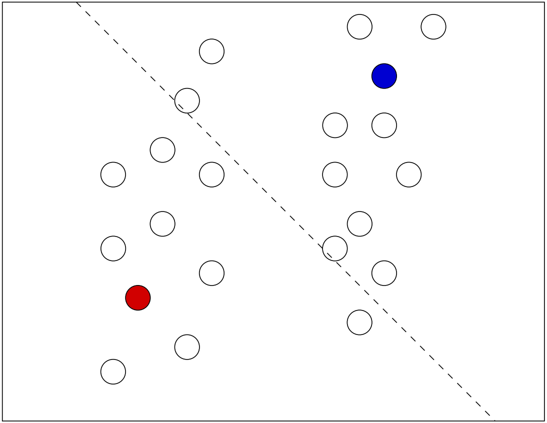
Presentation by Kuzman Ganchev

How can unlabeled data be useful?



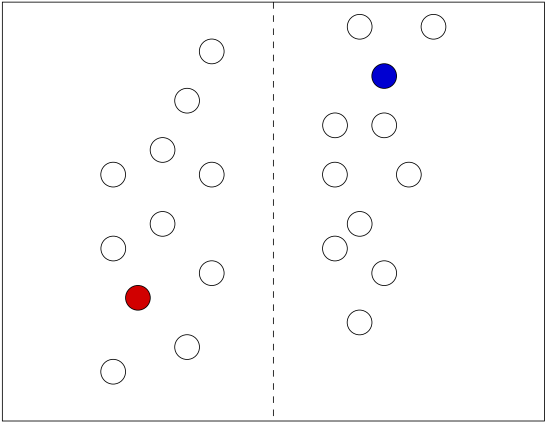
UNIVERSITY of PENNSYLVANIA

How can unlabeled data be useful?

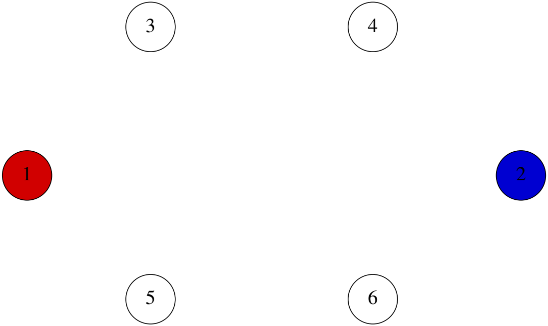


UNIVERSITY of PENNSYLVANIA

How can unlabeled data be useful?

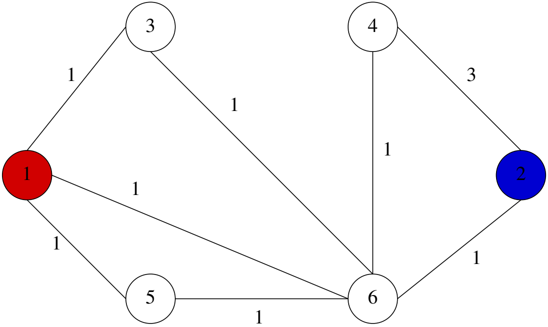


Graph Min-Cuts Example



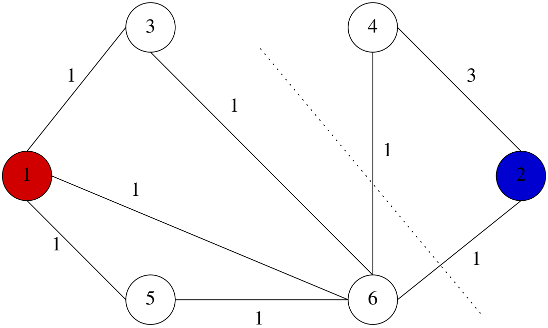
- Instances: $\{1 \ldots 6\}$
- Labels: $\{Red, Blue\}$
- Training Data: 1 is *Red*, 2 is *Blue*

Graph Min-Cuts Example



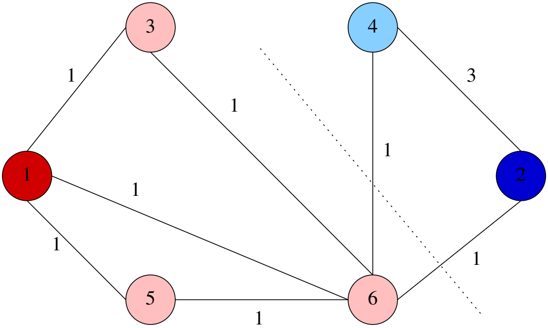
- Create graph with weighted edges

Graph Min-Cuts Example



- Find Min-Cut in graph

Graph Min-Cuts Example



- Label vertices according to partition

Notation

- L: labeled instances; $L = L_+ \cup L_-$
- U: unlabeled instances
- $w(e)$: weight of edge e; $e \notin E \Rightarrow w(e) = 0$
- $w(x, y)$: weight of edge $e = (x, y)$

Theoretical Results

- Min-Cut can create data easy for Nearest Neighbor
- Min-Cut can create data easy for Averaged Nearest Neighbor
- Min-Cut can create data easy for Symmetric Weighted Nearest Neighbor
- If the data is generated from well-shaped regions with gap between them then Min-Cut performs well

Only the last of these seems actually convincing

Min-Cut can create data easy for Nearest Neighbor

The labeling we produce for U is such that the leave-one-out cross-validation error for a nearest neighbor algorithm on $L \cup U$ is minimal.

- Let nn_{xy} be the indicator of "y is the nearest neighbor of x"
- Let $w(x, y) = nn_{xy} + nn_{yx}$

LOOCV error is: count of $x \in L \cup U$ with label of x different from label of its nearest neighbor. This is also the value of the cut.

University of Pennsylvania

Min-Cut can create data easy for k averaged Nearest Neighbor

The labeling we produce for U is such that the leave-one-out cross-validation error for a k averaged nearest neighbor algorithm on $L \cup U$ is minimal.

- define $w(x, y)$ as $w_{xy} + w_{yx}$ where w_{xy} is the weight of the label of y when classifying x
- Proof is just algebra

UNIVERSITY of PENNSYLVANIA

Min-Cut is good for a particular generative framework

Notation:

- δ-interior of R: set of points in R at least δ from the boundry
- δ-tendrils of R: set of points (in R?) at least δ from the δ-interior
- R is (ϵ, δ)-round iff:
    - at most ϵ fraction of its volume is in its δ-tendrils
    - its δ-interior is connected and non-empty
- V_r: the volume of a ball of radius r

Min-Cut is good for a particular generative framework

Result. If data is generated uniformly at random from:

- k $(\epsilon, \delta/4)$-round regions
- distance between any two regions is at least δ

Then we can classify with $1 - O(\epsilon)$ accuracy if we have:

- $O(\frac{k \log k}{\epsilon})$ labeled examples and
- $O(\frac{-\log V_{\delta/8}}{V_{\delta/4}})$ unlabeled examples

Experimental results