1 Overview

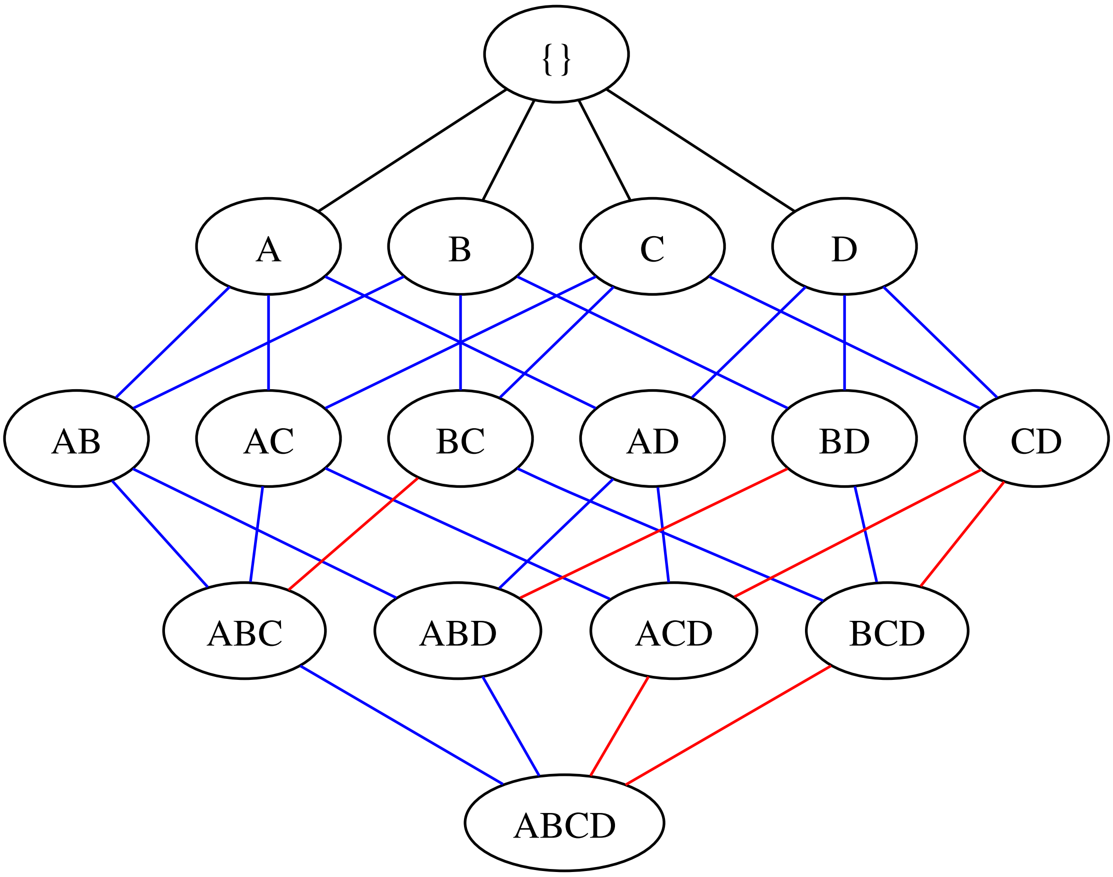
The last exercise session of the course is mainly focused on clustering.

2 Apriori Redux

Consider the lattice shown above. Going back to Apriori's candidate generation:

1. Which edges correspond to the *join* step?
2. Which edges correspond to *prune* tests?

Join-edges are blue, prune-edges red.



3 Decision Tree Building – AVC

For the dataset below, determine AVC-sets at the root node.

Age	Income	Loyalty card	Account	Responded?
≤ 30	High	No	Basic	$-$
≤ 30	High	No	Plus	$-$
$[31, 40]$	High	No	Basic	$+$
> 40	Medium	No	Basic	$+$
> 40	Low	Yes	Basic	$+$
> 40	Low	Yes	Plus	$-$
$[31, 40]$	Low	Yes	Plus	$+$
≤ 30	Medium	No	Basic	$-$
≤ 30	Low	Yes	Basic	$+$
> 40	Medium	Yes	Basic	$+$
≤ 30	Medium	Yes	Plus	$+$
$[31, 40]$	Medium	No	Plus	$+$
$[31, 40]$	High	Yes	Basic	$+$
> 40	Medium	No	Plus	$-$

	Responded	
	$+$	$-$
≤ 30	2	3
$[31, 40]$	4	0
> 40	3	2

	Responded	
	$+$	$-$
Low	3	1
Medium	4	2
High	2	2

	Responded	
	$+$	$-$
Yes	6	1
No	3	4

	Responded	
	$+$	$-$
Basic	6	2
Plus	3	3

4 Cluster Features 1

For the following two sets of points, write down the cluster feature that summarizes each set:

1. {(1,2), (2, 3), (3, 2), (2, 1)}: $\{4, (8, 8), (18, 18)\}$
2. {(2,4), (4, 3), (3, 4), (2, 2)}: $\{4, (11, 13), (33, 45)\}$

5 Cluster Features 2

Given the following cluster feature $\{5, (25, 30, 20), (145, 204, 90)\}$, do the following:

1. Compute the centroid of the cluster: $(\frac{25}{5} = 5, 6, 4)$
2. Compute the standard deviation from the centroid: $(\sqrt{\frac{145}{5} - \left(\frac{25}{5}\right)^2} = 2, 2.191, \sqrt{2})$
3. Give the Manhattan distance of the point (3, 8, 2) to the centroid: $2 + 2 + 2 = 6$

6 Clustering

Discuss the following two questions:

1. What are the differences between divisive and agglomerative clustering? What are their relative strengths and weaknesses?
   Top-down splits vs bottom-up merges. Finding “best” split more expensive than finding “best” merge. Different choices for merge (SL, CL, AL) lead to different shapes clusters (and different running times). No backtracking in division (Cobweb is special).
2. What is the key difference between k-Means clustering and EM clustering?
   Hard vs. soft assignment.
3. What does it mean that EM is a model based clustering approach?
   That there are explicit models for each cluster and we can check the likelihood that an instance was generated by that model, i.e. that it belongs to the particular cluster.

7 Thought question: Semi-supervised learning

Semi-supervised learning is a framework for addressing a lack of labeled data in predictive learning, i.e. instances that have a class label assigned to them. The question is then of course: why is there not enough labeled data? Try and think of at least two different reasons why getting labeled data might be difficult.
Last exercise we spoke about the difficulty of evaluating clustering (a.k.a. unsupervised learning) solutions. Is it more or less difficult to get correct labels for unsupervised learning, assignments of instances to groups of other instances they belong with, than for supervised learning, i.e. particular class labels?

8 Classifying Time Series with Python

Download the file `dtw.zip` from Toledo, unzip it, and start jupyter notebook.
We will be working with the bike rental system data.

1. Construct the time series of the hourly rental counts for each day and split it into the training and test sets.
2. Classify the time series from the test set as corresponding to working days or not using 1NN. Compare the performance when using Euclidean distance and DTW.
3. Explore the case when time series have varying length.