Machine Learning (CS60050)
Spring, 2020-2021

Instructor: Aritra Hazra
Scribed by:
Saurav Koranga (20CS60R65)
Trapti Singh (20CS60R68)
Friday, 9th April 2021

1. Recap

We have already seen three types of learning: Supervised, Unsupervised and Reinforcement Learning.
In Supervised Learning all the training examples are labelled i.e. all the data that are accessed have a label corresponding to the values of the attributes <x,y>.
In unsupervised learning, we get the raw data without any expert i.e. we don't have the label <x,?>. Here we classify or cluster similar oriented data.
In reinforcement learning, data comes with the deferred label i.e the data comes from the environment and training algorithm that we are making sense data from the environment and the actions will be taken in that environment only.

But in the real world there are few labelled examples as well as some unlabelled data points. For example: in computer vision, medical imaging all data is not that much clear.
Techniques that best utilize data, minimizing need for expert/human intervention. Paradigms where there has been great progress,and to make that more clear and useful we deal with two kind of learnings that are:

1. Semi-supervised Learning
2. Active(Interactive) Learning

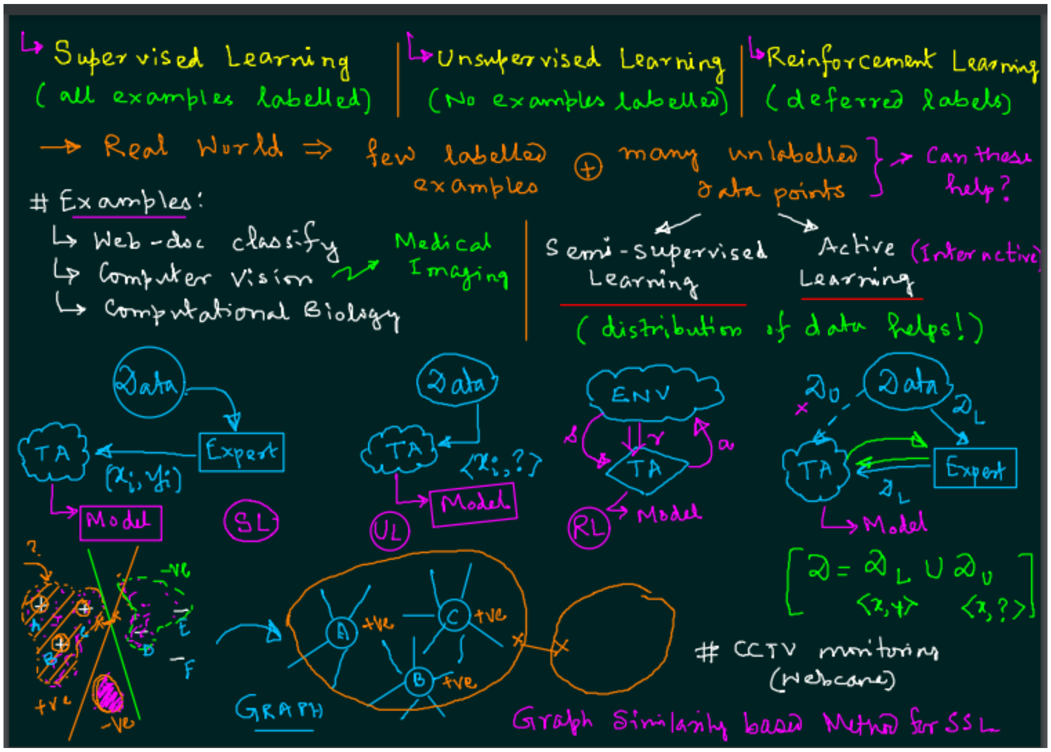

Figure.1

2. Semi-supervised Learning

Semi-supervised learning (SSL) lies somewhere between supervised and unsupervised learning. It combines the positives of both supervised and unsupervised learning. Here only a small amount of the training set D is labeled while a relatively large fraction of the training data is left unlabeled. The goal of a SSL algorithm is to learn a function f: X -> Y, given a training set {D = E, F} ; where E represents the labeled portion of the training data while F represents the unlabeled samples.

Key Insight/Underlying Fundamental Principle:
Unlabeled data is useful if we have a bias/belief not only about the form of the target, but also about its relationship with the underlying data distribution.

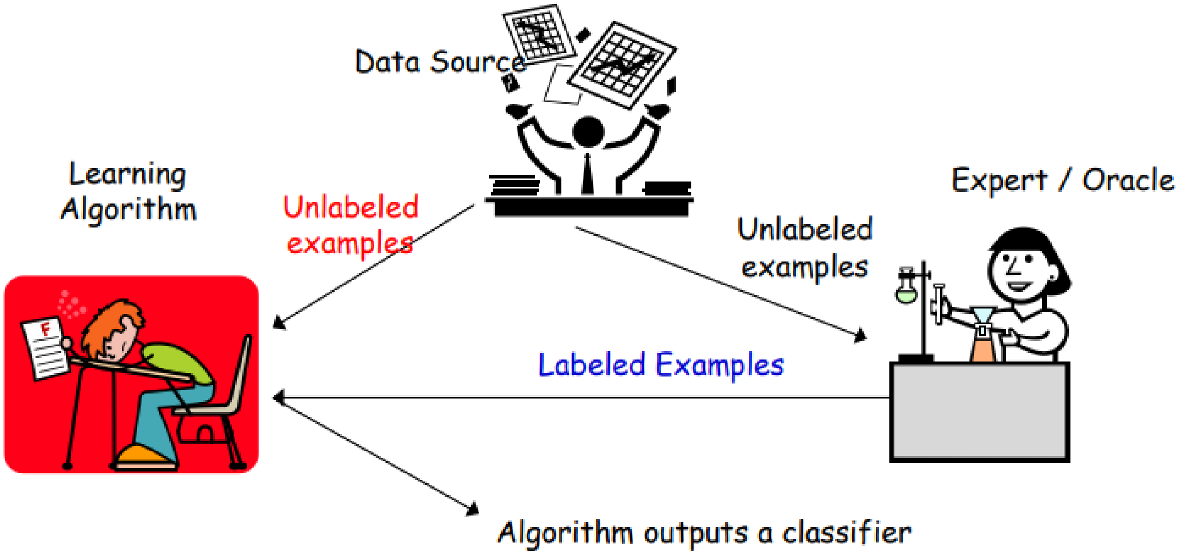

Figure.2: Semi-supervised Learning

3. Graph-based Active and Semi-Supervised Methods

Assume we are given a pairwise similarity function and that very similar examples probably have the same label. If we have a lot of labeled data, this suggests a Nearest-Neighbor type of algorithm. If you have a lot of unlabeled data, perhaps you can use them as "stepping stones".

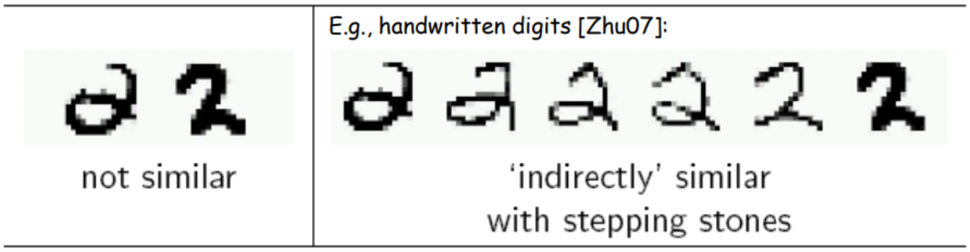

Figure.3

Idea: construct a graph with edges between very similar examples. Unlabeled data can help "glue" the objects of the same class together.
Often, transductive approach. (Given L + U, output predictions on U). Are allowed to output any labeling of $L \cup U$.

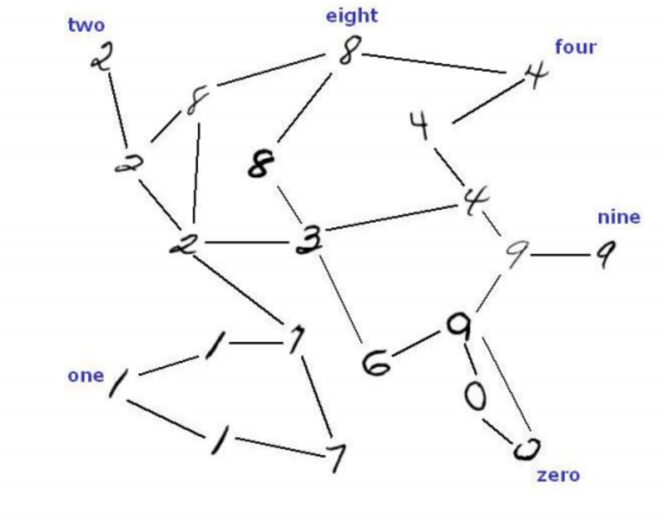

Figure.4

Main Idea: Might have also glued together in G examples of different classes. Often, transductive approach. (Given L + U, output predictions on U). Are allowed to output any labeling of $L \cup U$. Construct graph G with edges between very similar examples. Run a graph partitioning algorithm to separate the graph into pieces.

Several methods are:

- Minimum/Multiway cut
- Minimum “soft-cut”
- Spectral partitioning

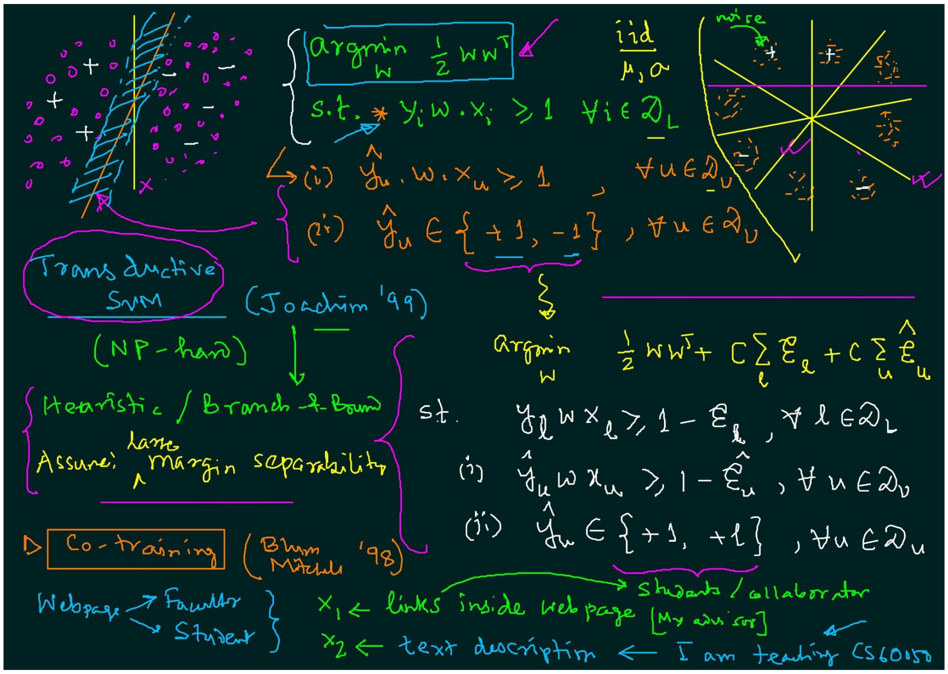
$\arg\min_w \frac{1}{2} ww^T$
s.t. $y_i w \cdot x_i \geq 1 \quad \forall i \in \mathcal{D}_L$
(i) $\hat{y}_u \cdot w \cdot x_u \geq 1, \quad \forall u \in \mathcal{D}_U$
(ii) $\hat{y}_u \in \{+1, -1\}, \forall u \in \mathcal{D}_U$
iid
μ, σ
noise
Transductive SVM
(Joachim '99)
(NP-hard)
Heuristic / Branch & Bound
Assume: large margin separability
$\arg\min_w \frac{1}{2} ww^T + C \sum_l \varepsilon_l + C \sum_u \hat{\varepsilon}_u$
s.t. $y_l w x_l \geq 1 - \varepsilon_l, \forall l \in \mathcal{D}_L$
(i) $\hat{y}_u w x_u \geq 1 - \hat{\varepsilon}_u, \forall u \in \mathcal{D}_U$
(ii) $\hat{y}_u \in \{+1, +1\}, \forall u \in \mathcal{D}_u$
Co-training
(Blum Mitchell '98)
Webpage → Faculty
Student
x_1 ← links inside web page
students / collaborator
[My advisor]
x_2 ← text description ← I am teaching CS60050

Figure.5

More observations from Joachim Paper:

He has proven that this approach will converge and frame optimal solutions. The primary thing here is to note that due to introduction of variables for labeled and unlabeled data and because of non-deterministic factor {+1, -1} the problem is NP-hard.

Earlier, when we didn't have non-determinism in consideration we solved this using quadratic programming(a mathematical technique to solve constraint optimisation problems). Now since the problem is NP-hard a deterministic polynomial time algorithm can not solve it. Joachim proposed a heuristic to guide through the choice of non-deterministic value for yu and a branch and bound approach(also used to solve 8 queens problem).

This method works well when categorizing news, i.e. whether sports article, politics, etc.

Demerit:

Assumption: Margin separability, that our data is linearly separable and we have large margin to establish this, also known as large margin assumption.
There could be many possibilities to cluster the data.

Co-training - by Blum and Mitchell ‘1998

suppose you want to classify whether the Webpage is of a faculty or is of a student. Let x1 and x2 are the 2 features. x1 looks for the links inside a webpage, as for a faculty the links could be my students or collaborator while for a student a link could be my advisor.

x2 is text description, like for a teacher a phrase could be "I am teaching" while for a student a phrase could be "course is taken".

In the paper, depending on the attribute x1 and x2 weak learners are learnt depending upon the restricted attribute set.

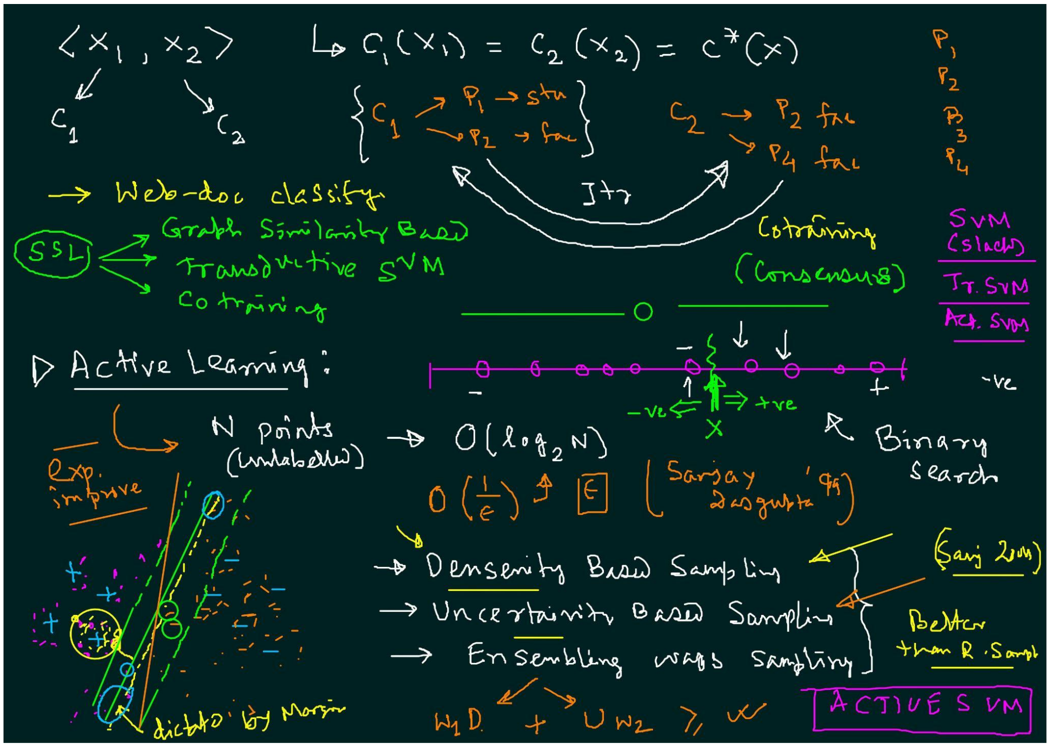

Figure.6

Assume that these weak learners C1(x1) = C2(x2) = C*(x), weak learner 1 classify based on attribute set x1 and weak learner 2 classify based on attribute set x2 and these two are the same with respect to the optimal classification.

Since we have a weak learner, we need to produce a strong learner from the weak learner. We can use Ensembling for this.

Let's say we have 4 pages(p1, p2, p3, p4), C1 predicts p1->student, p2->faculty, C2 predicts p2->faculty and p4->faculty.

These weak predictors are not good overall but good for a certain position.

C1 labels will be used to retrain c2 and similarly c2 labels will be used again to retrain c1 iteratively until we don't increase the number of labels. This is the co-training part of the paper. Some weak learners can classify some clusters well and other weak learners can classify other cluster well.
This performs very well for Web document classification.

Summary for semi-supervised learning:
We learnt three approaches:
1. Graph similarity based
2. Transductive SVM
3. Co-training

Active Learning

suppose you are dealing with circuit domain. In low-level circuits the simulation is very time consuming. Therefore one needs to be very meticulous about the examples that need to be chosen so that the classification gives the best result. This is very important for the medical domain as well.
Consider you have a number line with N unlabeled points with two classes (+ , -) and we want to find the point that divides the number line such that all (-) are one side and (+) are other. A simple binary search can help us with that. For this problem, we need to ask the expert for $O(\log_2 n)$ points, an exponential improvement. From Sanjay Das Gupta paper.
suppose we have some samples as given bottom left of fig, if we use SVM and classify using the labeled points. There are 3 things that we can do:

1. **Density based sampling**: In density based sampling we select the highly dense set of points and assign it the class which is closest to it.
2. **Uncertainty based sampling:** In uncertainty based sampling the points that fall inside a margin are uncertain points as they can dictate to shift the boundary to different positions. These uncertain points are chosen in uncertainty based sampling that are close to margin and scattered away from given labels.
3. **Ensembling based sampling:** In ensembling both density and uncertainty are taken into consideration with a weight factor to ensemble. W_1D and W_2U, whichever is the max the next point is labeled as such.

This is from the Sanjay DasGupta paper from 2004. This paper also probes that these sampling methods are better than random sampling. They have also given proof for what number of points you should take when you are going for semi-supervised learning.
This technique is Active SVM.

Online learning:

In online learning the only difference is that there is no provision of a training algorithm to judge which example you need next. There is a set where labels are given and a set where labels are not given, gradually we sample from which labeled point we want and which unlabeled point we want but do not have the choice to ask which next point that is needed.

References:

1. This scribe is based on lecture taught Prof. Aritra Hazra on 9-April-2021 in Machine Learning(CS60050) course.
2. Figure 1, 5 and 6 are taken from the handouts 16+17.
3. Figure 2,3 and 4 are taken from slide 17.