Google Cloud

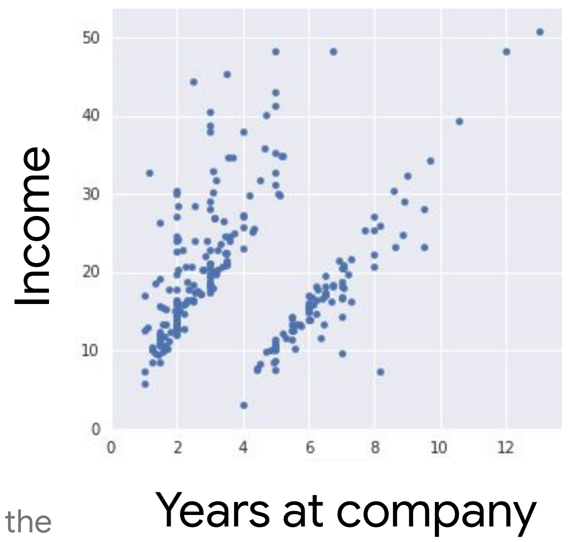
ASL Tech Talk:

Semi-supervised Learning

Unsupervised and supervised learning are the two main types of ML algorithms

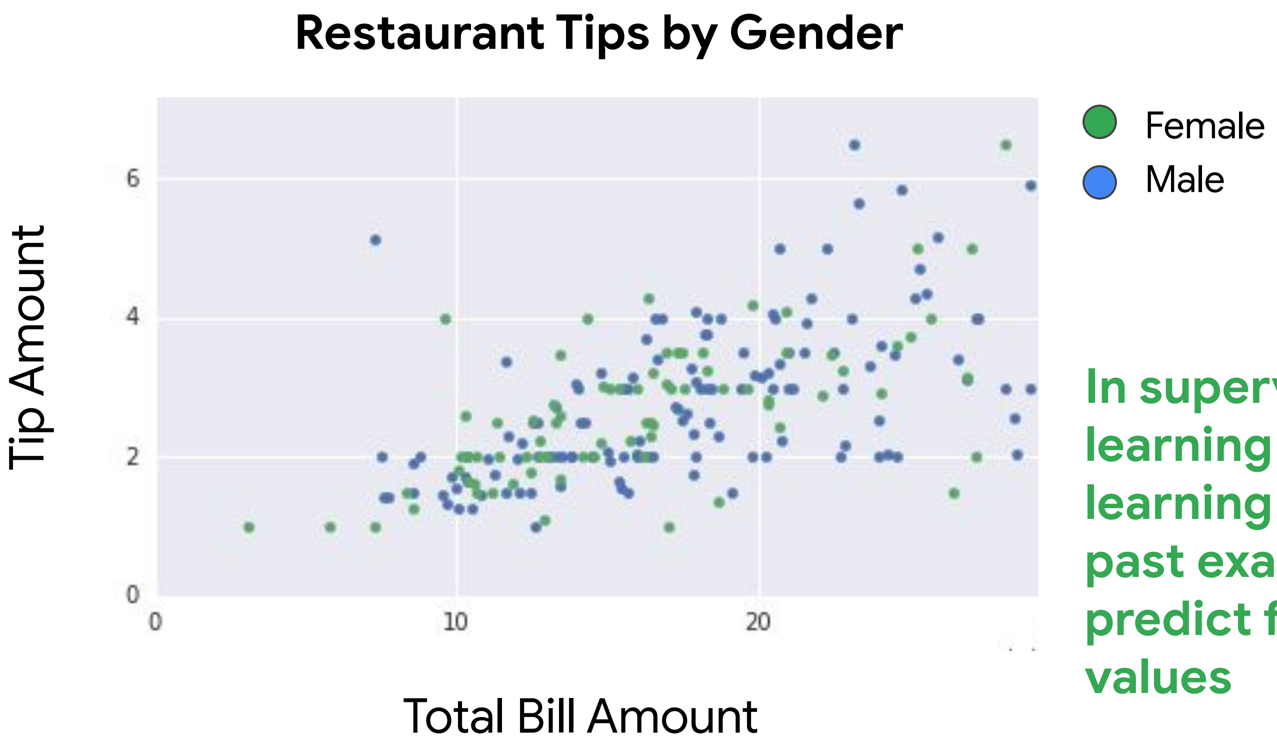
In unsupervised learning, data is not labeled

Income vs Job Tenure



Example Model: Clustering
Is this employee on the "fast-track" or not?

Supervised learning implies the data is already labeled

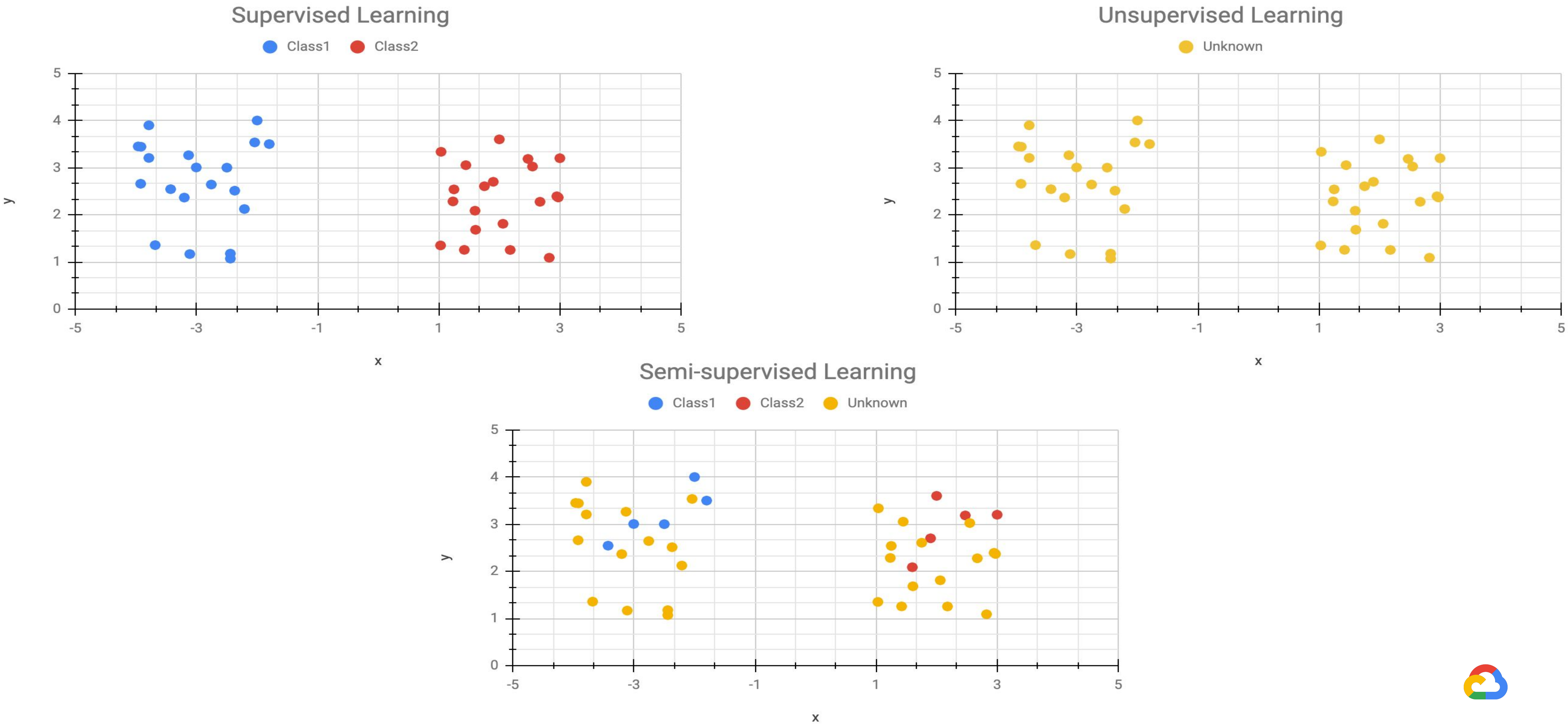


In supervised learning we are learning from past examples to predict future values

Aren't there just supervised and unsupervised learning?

No, there is another.

Semi-supervised learning is in between

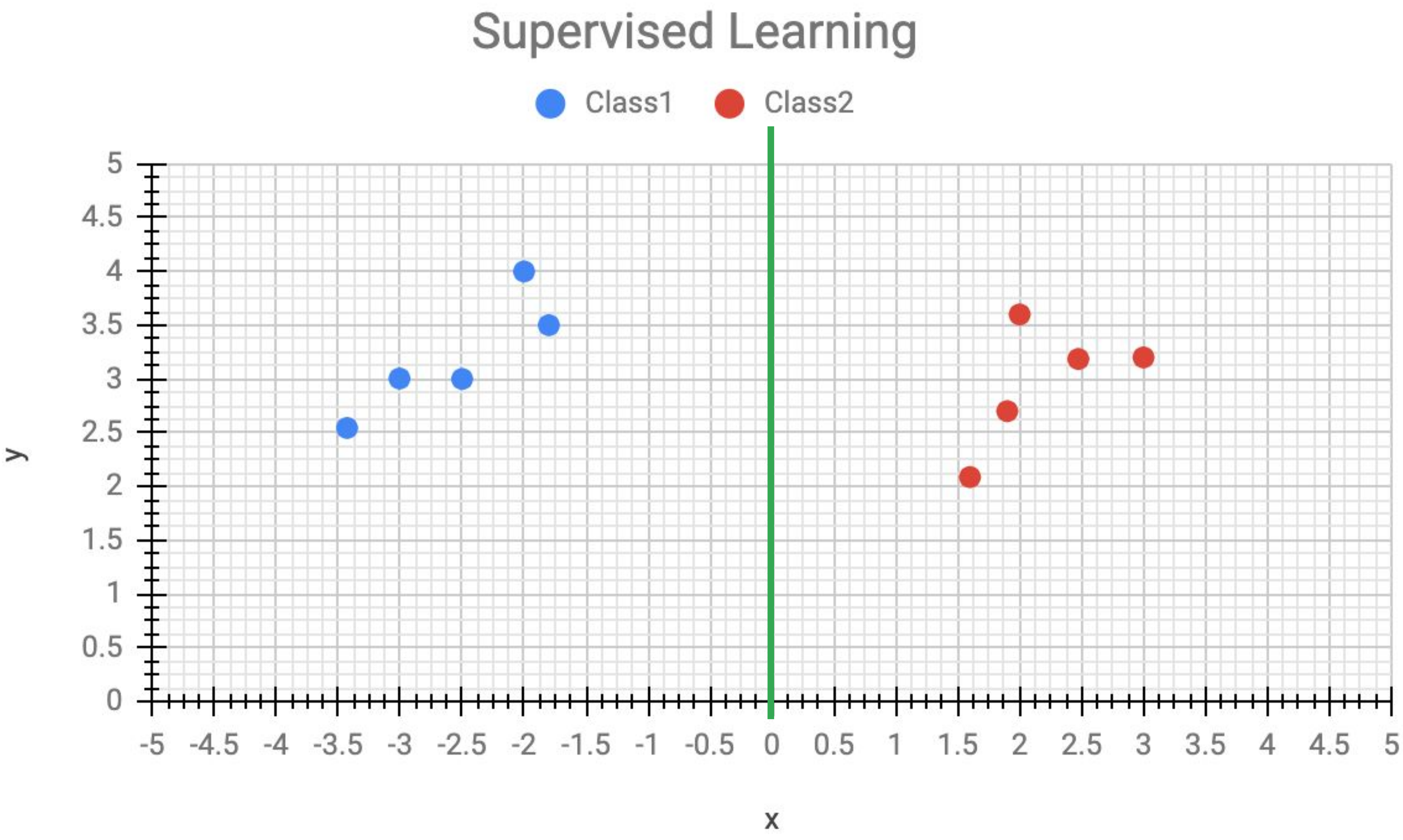


Supervised Learning

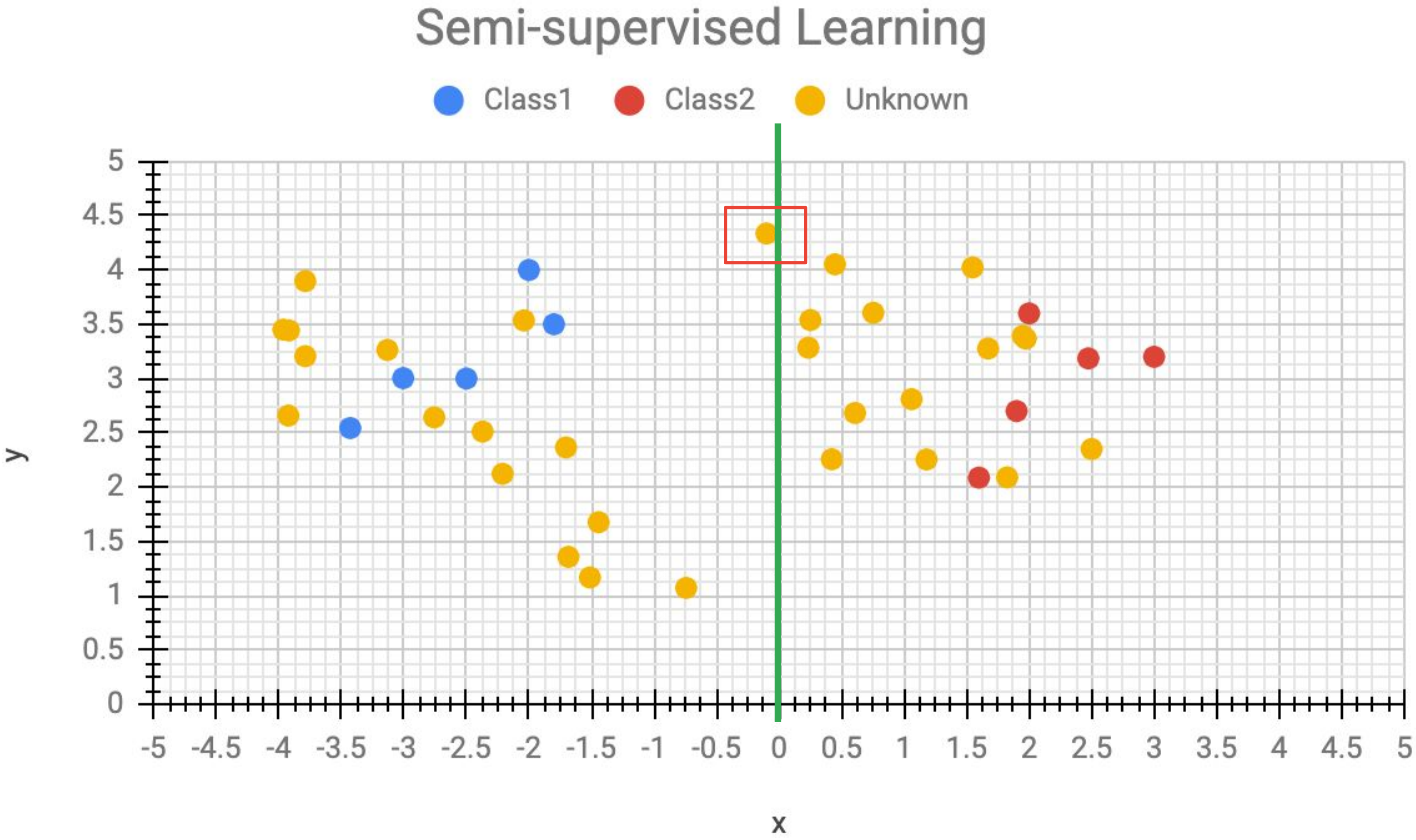
Unsupervised Learning

Semi-supervised Learning

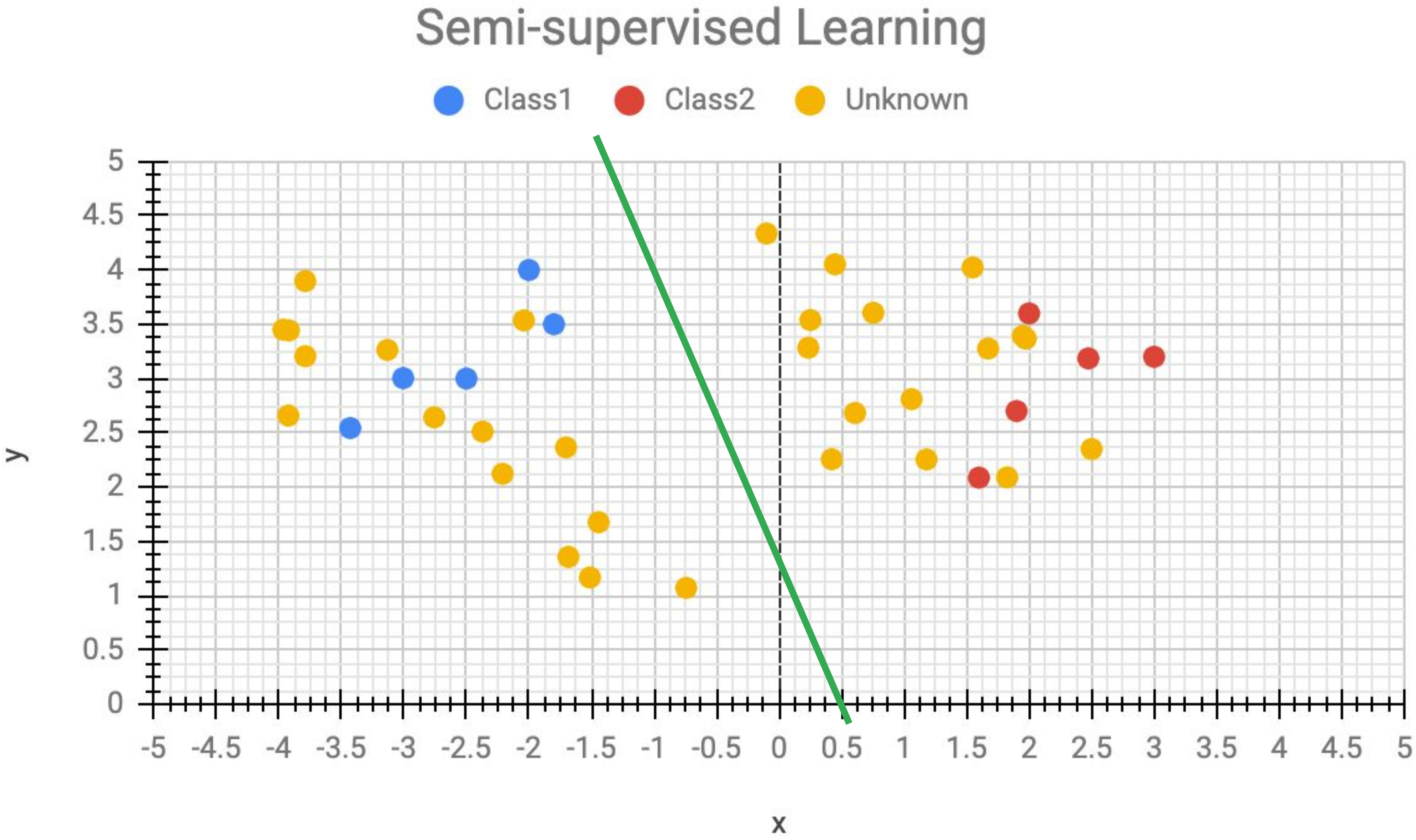
Supervised Learning with only a few examples



Classification boundary not able to generalize well



Much better generalization using unlabeled data



So how to use unlabeled data to help the labeled data generalize better?

There are many algorithms, but first some assumptions about the structure of the data distribution.

- Continuity
  - The closer points are, the higher probability they share a label.
- Cluster
  - There are discrete clusters and points in the same cluster have a higher probability they share a label.
    - A label can be associated with multiple clusters.
- Manifold
  - The data lie approximately on a manifold of much lower dimension than the input space.

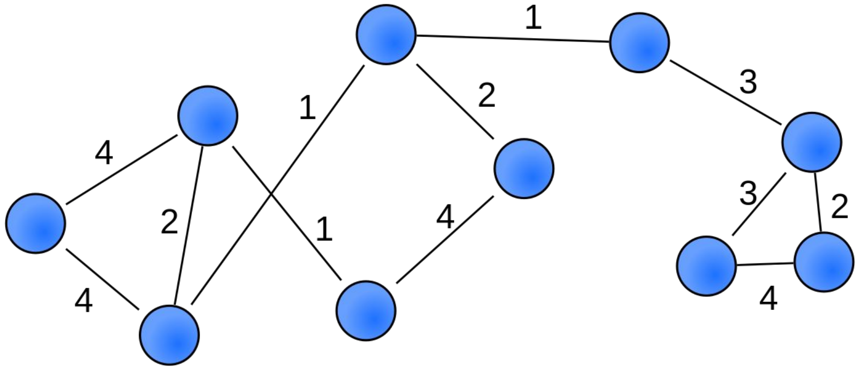
Self-training

- Most simple and easy approach.
- Repeat training on ever growing "labeled" dataset.
- Hard or soft labels can be used.

1. Train a model on labeled dataset.
2. Calculate predictions for entire unlabeled dataset.
3. Apply labels to predictions that have high confidence.
4. Pop those records from unlabeled dataset and push to labeled dataset.
5. Repeat until some convergence criteria is met.

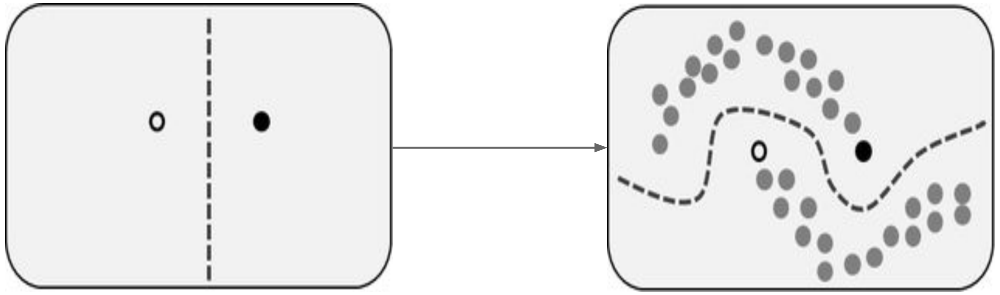
Graph-based models

1. Generate an undirected graph with a node for each point in both the labeled and unlabeled data.
   a. This can be through k nearest neighbors or some distance metric d.
2. The edges between two nodes will have a weight based on the normed difference and distance like exp(-||x - y||^2 / d).
3. Labeled points propagate their labels scaled by the connecting edges to all of the unlabeled nodes.



Manifold regularization

- Going one step further, we can use the graph as a proxy for the true manifold.
- We essentially add constraints based on the data's structure to limit the functions learned from the data.
- Add term to Tikhonov regularization to enforce function's smoothness relative to manifold and ambient input space.



$$\underset{f\in\mathcal{H}}{\operatorname{argmin}}\left(\boxed{\frac{1}{l}\sum_{i=1}^{l}V(f(x_i),y_i)+\lambda_A\|f\|_{\mathcal{H}}^2}+\boxed{\lambda_I\int_{\mathcal{M}}\|\nabla_{\mathcal{M}}f(x)\|^2\,dp(x)}\right)$$

Bayes' rule

$$p(\mathbf{w}|y, X) = \frac{p(y|X, \mathbf{w}) * p(\mathbf{w})}{p(y|X)}$$

$$posterior = \frac{likelihood * prior}{marginal\ likelihood}$$

Bayes' rule

- Likelihood indicates the compatibility of the evidence with the given hypothesis.
- Prior is the estimate of the probability of the hypothesis H before the data E, the current evidence, is observed.
- Marginal likelihood or the evidence, corresponds to new data that were not used in computing the prior probability.
- Posterior is the probability of a hypothesis given the observed evidence.

$$posterior = \frac{likelihood * prior}{marginal\ likelihood}$$

Generative models

- Want to find the distribution of data points belonging to each class p(x | y).
- From labeled data we know probability of a given point x having label y is p(y | x).
- From Bayes' rule, this is equal to p(x | y)p(y)/p(x).
- Can use information of our unlabeled data p(x) to better estimate p(x | y).
- Parameterize joint distribution to get conditional distributions p(x, y | w) = p(y | w)p(x | y, w).
- Can be done with GANs, Gaussian processes, etc.

$$\underset{\Theta}{\operatorname{argmax}} \left(\boxed{\log p(\{x_i, y_i\}_{i=1}^{l} | \theta)} + \boxed{\lambda \log p(\{x_i\}_{i=l+1}^{l+u} | \theta)} \right)$$

Low-density separation

- Will get better performance having smoother decision boundaries go through low-density regions.
- Transductive SVMs (TSVMs) are a very powerful tool for semi-supervised problems.
- Rather than finding maximal margin over just labeled data, assign labels to unlabeled data in such a way to maximize the margin over all data.
- Two different hinge loss terms, each for labeled and unlabeled.
- Can be costly, a lot of research searching for optimizations.

$$f^* = \operatorname*{argmin}_{f} \left(\sum_{i=1}^{l} \boxed{(1 - y_i f(x_i))_+} + \lambda_1 ||h||_{\mathcal{H}}^2 + \lambda_2 \sum_{i=l+1}^{l+u} \boxed{(1 - |f(x_i)|)_+} \right)$$

Cluster-then-label techniques

- Uses SSL's cluster assumption.
- Find point clusters of high density regions.
- Assign labels to those clusters.
- Do supervised training to find decision boundaries between clusters (low density regions).

Let's look at some code!

Thank you!

Questions?

cloud.google.com