Semi-Supervised Learning

5703: Machine Learning Practice
Semi-Supervised Learning

- Unsupervised Learning: only have samples in an input feature space
- Supervised Learning: samples are also labeled with class or a continuous value
- For many problems, observations are easy to collect, but the true labels are hard to come by
  - Expensive to measure these labels or have an expert provide them
  - One possibility: ignore samples that aren’t labeled and apply a supervised learning method
The Semi-Supervised Learning Problem

• We have:
  – Observations in an input feature space
  – Only a subset of the samples are labeled

• General approach: use the geometry of the full sample set to create models that better cover the feature space
Semi-Supervised Learning

Possibilities include:

• Infer “pseudo labels” for the unlabeled samples
  – Use a model constructed from the labeled samples to make guesses about the labels for the unlabeled samples (may be an iterative process)
  – Construct a model using the labeled and pseudo-labeled data

• Use unsupervised learning to project all samples into a lower dimensional and/or un-warped space
  – Then do supervised learning in the compressed space
Semi-Supervised Learning

Methods:
• Pseudo-labels: Label Propagation
• PCA compressed, followed by regression
Label Propagation

5703: Machine Learning Practice
Label Propagation

• Approach:
  – For labeled samples, identify “nearby” unlabeled samples
  – Copy the label to these new samples
  – Repeat

• What do we mean by nearby?
  – Could just take the k nearest neighbors
  – Could use Euclidean distance
  – With repeated steps, we can walk along the local manifold
Label Propagation

Algorithm

• Propagate labels
  – All samples have a true label or pseudo-label

• Use supervised learning method to learn a classifier using all of the data
Label Propagation

Variations:
• Samples keep their true labels, if available
• A subset of true labels are allowed to change
  – Allows us to “fix” incorrectly labeled samples
• Label Spreading: use affinity graph structure to propagate labels
  – Tends to provide smoother results
Example: Label Spreading

5703: Machine Learning Practice
Label Spreading

Live demo
Semi-Supervised Learning and Regression

5703: Machine Learning Practice
Semi-Supervised Learning

• These learning methods make a smoothness assumption:
  – Small changes in input feature position result in small changes in label
• With label propagation, this translates to small changes in the probability distribution
• Note that probability distribution values propagated along manifolds!
Semi-Supervised Learning for Regression

For regression: the manifold matters, here, too!

- Assume that the predicted output should vary smoothly along a manifold in the feature space
- And: we will make no commitment about how the value varies across regions with no samples
Drawing…
Semi-Supervised Learning and Regression

Step 1: Feature space embedding:

• Use both labeled and unlabeled data to discover a representation of the occupied manifolds in the feature space
• Capture these manifolds in terms of a neighborhood graph
• Compute geodesic distances between points
• Embed samples into a new space, translating geodesic distances into Euclidean distances (ISOmap!)
Semi-Supervised Learning and Regression

Step 2: Learn regression model using just the labeled data:
• First, project the samples into the lower dimensional space
• Then, learn the parameters of the function from compressed feature vectors to desired outputs