Lecture 21. Semi-supervised and Active Learning

COMP90051 Statistical Machine Learning

Semester 2, 2015
Lecturer: Andrey Kan

Content is based on slides provided by Ben Rubinstein

Types of learning revisited

- Supervised methods
  - Linear and logistic regression
  - Support vector machines
  - Neural networks, including convolutional neural networks
  - Probabilistic graphical models
- Unsupervised methods
  - K-means and hierarchical clustering
  - Dimensionality reduction
  - Community detection methods
  - Probabilistic graphical models
- This lecture
  - Semi-supervised learning
  - Active learning

Semi-Supervised Learning

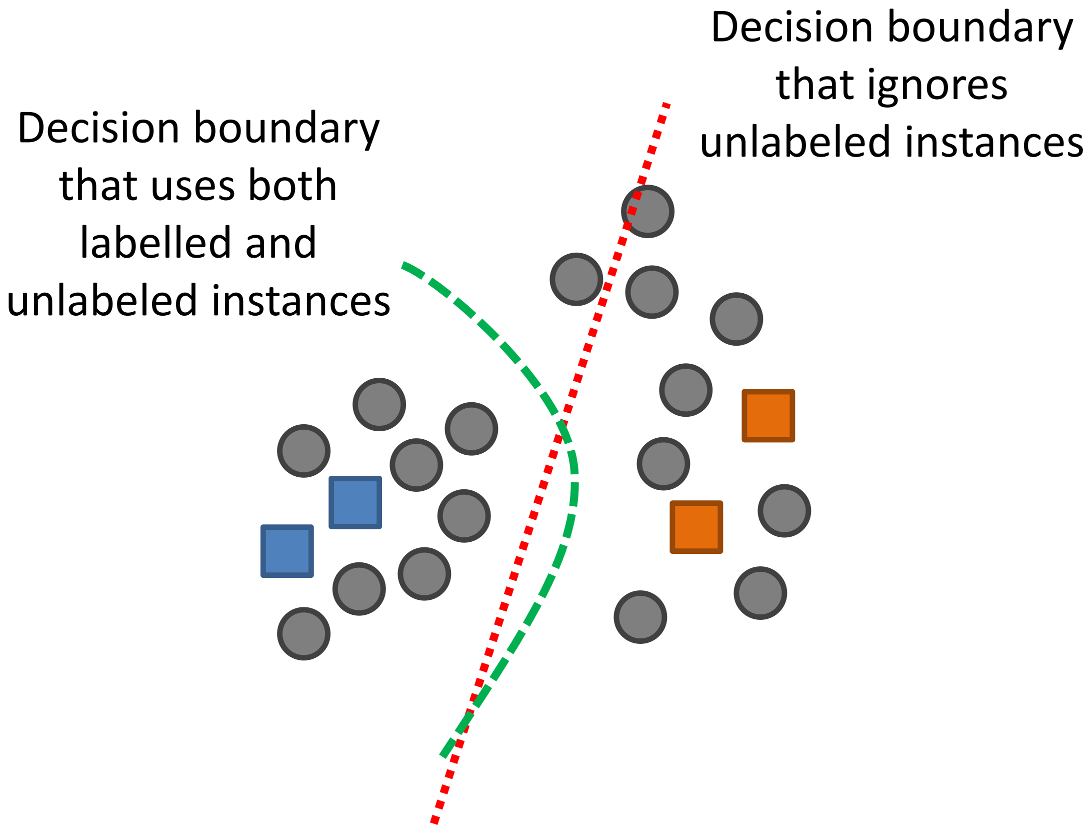
Training with instances, some of which are labelled

Motivating semi-supervised learning

- What if we had a small amount of labelled training data, and lots of unlabelled training data?
- What if we had a small amount of labelled training data and a limited budget to label more training data?
- What if we had same vs. different “constraints” instead of labels?
- Data is (often) cheap and abundant; labelling tends to be expensive
  * Example: Switchboard corpus -- 400 hours to label hour of speech data

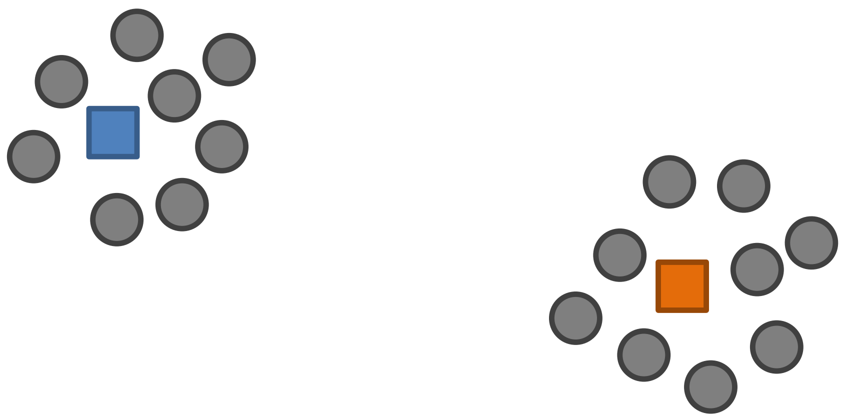
Semi-supervised learning: Example 1

- Semi-supervised learning = learning from both labelled and unlabelled data



Semi-supervised learning: Example 2

- Semi-supervised learning = learning from both labelled and unlabelled data



Labelled points provide us with cluster labels

Semi-supervised learning

- Semi-supervised learning = learning from both labelled and unlabelled data
- Semi-supervised classification
  * Training data consists of l labelled instances $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{l}$ and u unlabelled instances $\{\boldsymbol{x}_j\}_{j=l+1}^{l+u}$, often $u \gg l$
  * **Goal**: learn a better classifier f than is possible from labelled data alone
- Constrained clustering
  * A set of unlabelled instances $\{\boldsymbol{x}_j\}_{j=1}^{u}$ and a set of constraints between some pairs of instances
  * Usually in the form "must-link" and "cannot-link"
  * **Goal**: better clustering than from unlabelled data alone

Self training

- Perhaps the simplest example of semi-supervised learning is **self-training** (aka bootstrapping)

1. Initialise: $L = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{l}$, $U = \{\boldsymbol{x}_j\}_{j=l+1}^{l+u}$
2. Repeat:
   a) Train f from L using supervised learning
   b) Apply f to each instance in U (prediction)
   c) Identify a subset $U' \subseteq U$ where $f(\boldsymbol{x}_j)$ is "confident"
   d) Construct set $U'' = \left\{\left(\boldsymbol{x}_j, f\left(\boldsymbol{x}_j\right)\right) \middle| \boldsymbol{x}_j \in U'\right\}$
   e) Update $U \leftarrow U \setminus U'$, $L \leftarrow L \cup U''$
3. Until L is unchanged from one iteration to the next

Self training example

- Background: k-nearest neighbour classifier (supervised) predicts class of a new instances as majority class of the k closest training instances
- Propagating 1-nearest neighbour

1. Initialise: $L = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{l}$, $U = \{\boldsymbol{x}_j\}_{j=l+1}^{l+u}$
2. Repeat:
   a) Select $\{\boldsymbol{x}, \boldsymbol{x}'\} = \arg\min_{\boldsymbol{x}\in L, \boldsymbol{x}'\in U} d(\boldsymbol{x}, \boldsymbol{x}')$
   b) Update $U \leftarrow U \setminus \boldsymbol{x}'$, $\quad L \leftarrow L \cup \{(\boldsymbol{x}', f(\boldsymbol{x}'))\}$
3. Until $U = \emptyset$

Co-training: background

- Assume each instance has two views: $\boldsymbol{x} = [\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}]$
- Example: instance = pair (image + caption)
  - Instance features = [image based features, text based features]

"Carlton Gardens in autumn"

"The inner city is home to an extensive network of lively laneways and arcades."

Images: Xiquinho Silva and Rae Allen, Wikimedia Commons (CC2); Captions: Wikipedia

Co-training: algorithm

1. Initialise: $L^{(1)} = \left\{\left(\boldsymbol{x}_i^{(1)}, y_i\right)\right\}_{i=1}^{l}$, $L^{(2)} = \left\{\left(\boldsymbol{x}_i^{(2)}, y_i\right)\right\}_{i=1}^{l}$, $U = \left\{\boldsymbol{x}_j\right\}_{j=l+1}^{l+u}$
2. Repeat:
   a) Train $f^{(1)}$ on $L^{(1)}$; train $f^{(2)}$ on $L^{(2)}$
   b) Apply $f^{(1)}$ and $f^{(2)}$ separately to U
   c) Identify a subset $U^{(1)} \subseteq U$ where $f^{(1)}$ is “confident”
   d) Identify a subset $U^{(2)} \subseteq U$ where $f^{(2)}$ is “confident”
   e) Label instances from $U^{(2)}$ according to $f^{(2)}$, and add them to $L^{(1)}$
   f) Label instances from $U^{(1)}$ according to $f^{(1)}$, and add them to $L^{(2)}$
   g) Update $U \leftarrow U \setminus \left(U^{(1)} \cup U^{(2)}\right)$
3. Until U is unchanged from one iteration to the next

Co-training: Assumptions

- There is a feature split $\boldsymbol{x} = [\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}]$ which leads to independent classifiers
  - That is, $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$ are conditionally independent given the label
- $\boldsymbol{x}^{(1)}$ or $\boldsymbol{x}^{(2)}$ alone is sufficient to train a good classifier

Active Learning

Iteratively request labels and use the model (trained thus far) to do so

Active learning

- *Active learning* builds off the hypothesis that a classifier can achieve higher accuracy with fewer training instances if it is allowed to have some say in the selection of training instances
- The underlying assumption is that labelling is a finite resource, which should be expended in a way which optimises machine learning effectiveness
- Active learners pose *queries* (unlabelled instances) for labelling by an *oracle* (e.g., a human annotator)

Active learning: Sampling

- There are three main sampling approaches in active learning

1. Membership query synthesis: the active learner synthesises queries for labelling
   * E.g., proposes a particular combination of chemicals to use in a yeast growth medium
2. Stream-based selective sampling: for each instance from the stream, the model decides to query or discard
3. Pool-based sampling: the active learner selects from a fixed set of unlabelled instances what it wants to be labelled

Active learning: Query strategies

- One simple query strategy is to query those instances the classifier is least confident of the classification for:

$$\boldsymbol{x}^* = \underset{\boldsymbol{x} \in U}{\arg\max}\big(1 - P(\hat{y}|\boldsymbol{x})\big)$$

  - Where $\hat{y} = \underset{y \in S}{\arg\max}\, P(\hat{y}|\boldsymbol{x})$
  - and S is a set of possible labels

- Alternatively, it may be appropriate to perform "margin sampling":

$$\boldsymbol{x}_M^* = \underset{\boldsymbol{x} \in U}{\arg\min}\big(P(\hat{y}_1|\boldsymbol{x}) - P(\hat{y}_2|\boldsymbol{x})\big)$$

  - Where $\hat{y}_1$ and $\hat{y}_1$ are the first and second most-probable label predictions for $\boldsymbol{x}$

Active learning: Query strategies

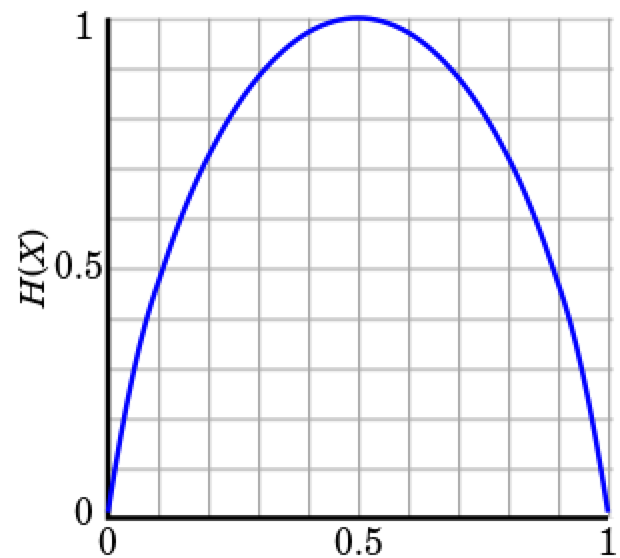
- Or better still, to use entropy as an uncertainty measure:

$$x_H^* = \arg\max_{\boldsymbol{x}\in U} - \sum_{y_i \in S} P(y_i|\boldsymbol{x}) \log_2 P(y_i|\boldsymbol{x})$$

- For binary classification with labels 0 and 1, entropy is:

$$H(p) = -p \log_2 p - (1-p) \log_2 (1-p)$$

  - Where $p = P(y = 1|\boldsymbol{x})$



$P(y = 1|\boldsymbol{x})$

Example: binary classification with labels 0 and 1

Image: Brona and Alessio Damato, Wikimedia Commons (CC3)

Query-by-committee

- A more complex strategy involving multiple classifiers is *query-by-committee* (QBC), where a suite of classifiers is trained over a fixed training set L, and the instance where there is the highest disagreement is selected for querying
- QBC assumes that it is possible to generate a suite of set of heterogeneous base classifiers, much like …
- Determination of relative disagreement can again occur via entropy, or alternatively via one-vs-rest relative entropy

Active Learning: Practicalities

- Active learning is used increasingly widely, but must be handled with some care:
  - empirically shown to be a robust strategy, but a theoretical justification has been slower to prove
  - querying is inherently biased towards a particular class set and learning approach(es), which may limit the general utility of the resulting dataset
  - results to suggest that active learning is more highly reliant on "clean" labelling

Summary

- What are unsupervised and semi-supervised learning
- What is self-training, and how does it operate?
- What is co-training, and how does it operate? What assumptions is it based on?
- What is active learning?
- What are the main sampling strategies in active learning?
- Outline a selection of query strategies in active learning.