Classification with Multi-Modal Data

Instructor: Guo-Jun Qi

CAP6676: Knowledge Representation
Multi-Modal Problem

• The same set of objects can be described in multiple modalities, e.g., text, visual, sensor and audio modalities

• Multi-Modal features are separated into K groups
  \[ X = (X^1, X^2, ..., X^K) \]

• We have both labeled and unlabeled data

• Task: learning with multi-modal data
  • Find target functions \( \{f_1, ..., f_K\} \) on K modalities so that the target functions agree on the labeling of these modalities.
Learning with Two Modalities

• Input
  • Features can be split into two spaces $X = X_1 \times X_2$
  • The two spaces are heterogeneous, i.e., different dimensions
  • They are redundant, but not completely correlated
    • In a video, we have two coupled visual and audio modalities
    • In satellite imaging, there are multiple related images in different spectrums

• Properties
  • Labeling Compatibility – The labels on different modalities should be identical for the same object.
  • Independence – Given the label of any example, the descriptions on different modalities should be independent of each other.
How does it work?

• Conditions
  • Compatible – Reduce the search space to where the two classifiers agree on unlabeled
  • Independence – If otherwise the descriptions of two modalities are dependent on each other, there is no need to learn two labeling functions since one of them can be derived directly from the other based on this dependency.
    • If $f_1: X_1 \rightarrow Y, g: X_2 \rightarrow X_1$, then $f_2: X_2 \rightarrow Y$ can be set to $f_2(x^2) = g(f_1(x^2))$

• Algorithms
  • Searching for compatible labeling functions
  • Co-regularization
  • Canonical correlation analysis (future lecture)
Searching for Compatible Labeling Functions

- **Intuitions**
  - Two individual labeling functions (i.e., classifiers) can be learned from the labeled examples of two modalities
  - The two labeling functions should agree on the unlabeled examples
    - This assumption can be used to enlarge the training set by annotating the unlabeled examples with identical labels.

- **Algorithms**
  - Co-training [BIMi98]
  - Co-EM [NiGh00]
  - Variants of Co-training [GoZh00]
Co-training
Co-training

Given:
- a set $L$ of labeled training examples
- a set $U$ of unlabeled examples

Create a pool $U'$.
Loop for $k$ iterations:

Use $L$ to train a classifier $h_1$ that considers only the $x_1$ portion of $x$
Use $L$ to train a classifier $h_2$ that considers only the $x_2$ portion of $x$
Allow $h_1$ to label $p$ positive and $n$ negative examples from $U'$
Allow $h_2$ to label $p$ positive and $n$ negative examples from $U'$
Add these self-labeled examples to $L$
Randomly choose $2p + 2n$ examples from $U$ to replenish $U'$

Train two classifiers from two views
Select the top unlabeled examples with the most confident predictions from the other classifier
Add these self-labeled examples to the training set
Webpage Categorization

View1: Page Text

View2: Hyperlink Text
Result

Figure 2: Error versus number of iterations for one run of co-training experiment.
Co-EM

- **Algorithm**
  - Labeled dataset L, Unlabeled dataset U, and let $U_1$ be empty and $U_2 = U$.
  - Iterate the following
    - Train a classifier $h_1$ from the feature set $X_1$ of L and $U_1$.
    - Probabilistically label all the unlabeled data in $U_2$ using $h_1$
      - Probabilistically labeling can be performed by drawing a label from the posterior $P(Y|X_1)$, e.g., a logistic regression classifier
    - Train a classifier $h_2$ from the feature set $X_2$ of L and $U_2$.
    - Let $U_1 = U$, probabilistically label all the unlabeled data in $U_1$ using $h_2$
  - Output $h_1$ and $h_2$

- **Co-EM vs. Co-Training**
  - soft vs. hard labeling
  - Selecting unlabeled data into training set: all vs. the most confident ones.
Canonical Correlation Analysis

• Intuitions
  • Reduce the feature space into a low-dimensional space containing discriminative information
  • With compatible assumptions, the discriminative information is contained in the directions that correlate the two modalities the most.
  • The goal is to maximize the correlation between the multimodal data in two projected spaces
Co-Regularization Framework

• Intuitions
  • Train two classifiers from the two modalities simultaneously
  • Add a regularization term to enforce that the two classifiers agree on the predictions of the unlabeled data

\[
\min R(f_1; L_1) + R(f_2; L_2) + R(f_1, f_2; U_1, U_2)
\]

• \(R(f_1; L_1)\) and \(R(f_2; L_2)\) are the training loss of two classifiers on the two modalities of labeled data
  • E.g., \(R(f_1; L_1) = -\sum_{(x,y)\in L} \log(1 + \exp(yf_1(x)))\) logistic loss with \(f_1(x) = w^T x\)
  • \(R(f_1, f_2; U_1, U_2)\) is the disagreement between two classifiers on unlabeled data of two modalities
Loss function

• Penalizing disagreement

- Exponential: \[ \sum_{x \in U} \exp(-\hat{y}_2 f_1(x)) + \exp(-\hat{y}_1 f_2(x)) \]

- Least Square: \[ \sum_{x \in U} (f_1(x) - f_2(x))^2 \]

- Bhattacharyya distance: \[ E_U(B(p_1, p_2)) \]
  \[ B(p_1, p_2) = -\log \sum \sqrt{p_1(y)p_2(y)} \]
Bhattacharyya distance

Exponential loss

Least square
Loss function

• When two classifiers do not agree
  • Loss grows exponentially, quadratically and linearly

• When two classifiers agree
  • Little penalty, or penalizing the margin (increasing the margin of agreement)
Training Co-RLS

• Input:
  • two labeled datasets on different modalities $L_1$ and $L_2$
  • Unlabeled datasets on two modalities $U_1$ and $U_2$

• Output:
  • two classifiers $f_1(x) = w_1^T x$ and $f_2(x) = w_2^T x$

• Iteration:

\[ \Delta w_1 = \frac{\partial}{\partial w_1} R(f_1; L_1) + R(f_2; L_2) + R(f_1, f_2; U_1, U_2) \]

\[ \Delta w_2 = \frac{\partial}{\partial w_2} R(f_1; L_1) + R(f_2; L_2) + R(f_1, f_2; U_1, U_2) \]

\[ w_1 \leftarrow w_1 - \alpha \Delta w_1 = \frac{\partial}{\partial w_2} R(f_1; L_1) + R(f_2; L_2) + R(f_1, f_2; U_1, U_2) \]

\[ w_2 \leftarrow w_2 - \alpha \Delta w_2 = \frac{\partial}{\partial w_2} R(f_1; L_1) + R(f_2; L_2) + R(f_1, f_2; U_1, U_2) \]
Training Co-RLS

• Input:
  • two labeled datasets on different modalities $L_1$ and $L_2$
  • Unlabeled datasets on two modalities $U_1$ and $U_2$

• Output:
  • two classifiers $f_1(x) = w_1^T x$ and $f_2(x) = w_2^T x$

• Stochastic iteration:
  \[
  \Delta w_1 = \frac{\partial}{\partial w_1} R(f_1; (x_n^1, y_n^1)) \\
  \Delta w_2 = \frac{\partial}{\partial w_2} R(f_1; (x_n^2, y_n^2)) \\
  \Delta w_1 = \frac{\partial}{\partial w_1} R(f_1, f_2; x_n^1, x_m^2) \\
  \Delta w_2 = \frac{\partial}{\partial w_2} R(f_1, f_2; x_n^1, x_m^2)
  \]

\[
  w_1 \leftarrow w_1 - \alpha \Delta w_1 \\
  w_2 \leftarrow w_2 - \alpha \Delta w_2
  \]
Summary

• Learning with multimodal data
  • Each data has several different views of modalities
  • Making use of related modalities to improve the compatibility between classifiers

• Co-training – mutually label unlabeled data, and train each classifier with the labeled examples by the other classifier

• Co-EM – soft labeling and use all the unlabeled data for training

• Co Regularization – enforcing compatibility by minimizing the disagreement between classifiers on different modalities