COS324: Introduction to Machine Learning
Lecture 12: Similarity Learning

Prof. Elad Hazan & Prof. Yoram Singer
Background

- So far we focused on multiclass problems where each example is associated with a label / category.
- We also discussed incorporation of misclassification cost that is not the same across classes.
- Settings where feedback is relative w.r.t pairs of instances.

End of lecture describes ways to build multiclass classifiers from similarity operators.
Problem Setting

- Training set of pairs of instances with similarity feedback
  \[ S = \{(x^1_i, x^2_i, y_i)\}_{i=1}^m \]

- Feedback \( y_i \in \{-1, +1\} \) similar \((y = +1)\) dissimilar \((y = -1)\)
  \[ x^1 \sim x^1 \iff y = +1 \quad x^1 \not\sim x^1 \iff y = -1 \]

- Labels can be obtained directly from similarity feedback or as a by-product of multi-labeled data
  
  \[(x_i, y_i), (x_j, y_j) \text{ where } y_i, y_j \in [k] \implies (x_i, x_j, (-1)^{1[y_j \neq y_j]})\]

- Similarity feedback “flattens” uneven class distribution

- Example: assume \( k = 3 \) and
  \[ |\{i : y_i = 1\}| = \frac{4m}{5} \quad |\{i : y_i = 2\}| = \frac{m}{10} \quad |\{i : y_i = 3\}| = \frac{m}{10} \]
  then number of similar pairs is \( \approx \frac{2m}{3} \)
ERM for Similarity Learning

- Instances $x_i^j \in X$ ($i \in [m] \ j \in [2]$)

- Similarity function operates on pairs of elements

  \[ h : X \times X \rightarrow \mathbb{R} \]

- If $x_i^1 \sim x_i^2$ we want

  \[ h(x_i^1, x_i^2) \gg 0 \]

- If $x_i^1 \not\sim x_i^2$ we want

  \[ h(x_i^1, x_i^2) \ll 0 \]
First Attempt

- Transform $\mathbf{x}^1, \mathbf{x}^2 \mapsto \Delta = \mathbf{x}^1 - \mathbf{x}^2$

- Set $h(\mathbf{x}^1, \mathbf{x}^2) = \sum_j w_j |\Delta[j]| + b$

- Works ‘ok’, just ‘ok’

- Left pair will be classified as **dissimilar** & right pair as **similar**

- $\mathbf{x}^1 \not\sim \mathbf{x}^2$

- $\mathbf{x}^1 \sim \mathbf{x}^2$
Shift Invariance

Two vectors $u, v$ such that $u[1 : d - m] \approx v[m + 1 : d]$

However, similarity score $h(u, v) \approx 0$
Bilinear Forms

Define

\[ h(u, v) = u^\top A v \quad \text{where} \quad A \in \mathbb{R}^{d \times d} \]

which amounts to

\[ h(u, v) = \sum_{i=1}^{d} \sum_{j=1}^{d} A_{ij} u_i v_j \]

Example with \( d = 4 \)

\[ u \]
\[ \begin{array}{cccc}
  u_1 & u_2 & u_3 & u_4 \\
\end{array} \]

\[ A \]
\[ \begin{array}{cccc}
  A_{1,1} & A_{1,2} & A_{1,3} & A_{1,4} \\
  A_{2,1} & A_{2,2} & A_{2,3} & A_{2,4} \\
  A_{3,1} & A_{3,2} & A_{3,3} & A_{3,4} \\
  A_{4,1} & A_{4,2} & A_{4,3} & A_{4,4} \\
\end{array} \]

\[ v \]
\[ \begin{array}{c}
  v_1 \\
  v_2 \\
  v_3 \\
  v_4 \\
\end{array} \]
Surrogate Losses for Similarity Functions

- **Example** a pair \((\mathbf{x}^1, \mathbf{x}^2)\) and feedback \(y \in \{-1, +1\}\)

- **Real-valued prediction** \(h(\mathbf{x}^1, \mathbf{x}^2) = (\mathbf{x}^1)^\top A \mathbf{x}^2\)

- **Hinge Loss** with margin \(\gamma\)

\[
\left[ \gamma - y (\mathbf{x}^1)^\top A \mathbf{x}^2 \right]
\]

- **Logistic Loss** with margin \(\gamma\)

\[
\log \left( 1 + \exp \left( \gamma - y (\mathbf{x}^1)^\top A \mathbf{x}^2 \right) \right)
\]
Alternative Formulation

- Define a $d \times d$ matrix $Z \overset{\text{def}}{=} (x^1)(x^2)^\top$

\[
\begin{array}{cccc}
 x_1^1 & x_2^1 & x_3^1 & x_4^1 \\
 x_1^2 & x_2^2 & x_3^2 & x_4^2 \\
 x_1^3 & x_2^3 & x_3^3 & x_4^3 \\
 x_1^4 & x_2^4 & x_3^4 & x_4^4 \\
\end{array}
\]

\[
\begin{array}{cccc}
 Z_{1,1} & Z_{1,2} & Z_{1,3} & Z_{1,4} \\
 Z_{2,1} & Z_{2,2} & Z_{2,3} & Z_{2,4} \\
 Z_{3,1} & Z_{3,2} & Z_{3,3} & Z_{3,4} \\
 Z_{4,1} & Z_{4,2} & Z_{4,3} & Z_{4,4} \\
\end{array}
\]

- Remember, $Z$ stands for Zorro and not Zero

- Define $A \cdot Z \overset{\text{def}}{=} \sum_{i,j} A_{i,j} Z_{i,j}$

- Use SGD for ERM in order to find $A$
Skeleton of SGD for Similarity

Input: dataset of labeled pairs $S = \{x_1^i, x_2^i, y_i\}$

Transform: $x_1^i, x_2^i \mapsto Z_i$ where $Z_i := (x_1^i) (x_2^i)^\top$

Loss: $f_i(A) = \ell(A \cdot Z_i)$ and $F(A) = \frac{1}{m} \sum_{i=1}^{m} f_i(A)$

Gradient: $\hat{\nabla}_A(F) = \frac{1}{|S'|} \sum_{i \in S'} \ell(A \cdot Z_i)Z_i$ where $\dot{\ell}(\mu) = \frac{d\ell}{d\mu}$

Train: call SGD with $F$, $S$, $\nabla_A F \Rightarrow \hat{A}$

Predict: for $(\tilde{x}_1, \tilde{x}_2)$ output $\text{sign}\left((\tilde{x}_1)^\top A \tilde{x}_2\right)$
Remarks

• It is often required that $h(x^1, x^2) = h(x^2, x^1)$

• Symmetric Zorro

\[
Z \overset{\text{def}}{=} \frac{1}{2} (x^1) (x^2)^\top + \frac{1}{2} (x^2) (x^1)^\top
\]

• Similarity matrix can be used to define a pseudo-metric

\[
\|x\|_A^2 = x^\top A x \quad \Rightarrow \quad \|x^1 - x^2\|_A \overset{\text{def}}{=} \|v\|_A \quad \text{where} \quad v = x^1 - x^2
\]

• However, need to constrain $A$ to be positive semi-definite (PSD)

\[
A \in \{ M : M \succeq 0 \} \quad \text{where} \quad M \succeq 0 \iff \forall v : v^\top M v \geq 0
\]

• Projecting a matrix onto the PSD cone is expensive: $O(d^3)$