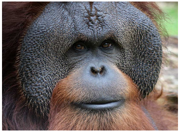
COS324: Introduction to Machine Learning

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

 $\sim$  $\not\sim$ 

- End of lecture describes ways to build multiclass classifiers from similarity operators

Problem Setting

- Training set of pairs of instances with similarity feedback

$$S = \left\{ (\mathbf{x}_i^1, \mathbf{x}_i^2, y_i) \right\}_{i=1}^{m}$$

- Feedback $y_i \in \{-1, +1\}$ similar $(y = +1)$ dissimilar $(y = -1)$

$$\mathbf{x}^1 \sim \mathbf{x}^1 \Leftrightarrow y = +1 \qquad \mathbf{x}^1 \not\sim \mathbf{x}^1 \Leftrightarrow y = -1$$

- Labels can be obtained directly from similarity feedback or as a by-product of multi-labeled date

$$(\mathbf{x}_i, y_i)\ ,\ (\mathbf{x}_j, y_j) \ \text{ where } \ y_i, y_j \in [k] \quad \Rightarrow \quad \left(\mathbf{x}_i, \mathbf{x}_j, (-1)^{\mathbb{1}[y_j \neq y_j]} \right)$$

- 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 $\mathbf{x}_i^j \in \mathcal{X}$ $(i \in [m]\ j \in [2])$
- Similarity function operates on pairs of elements

$$h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$

- If $\mathbf{x}_i^1 \sim \mathbf{x}_i^2$ we want

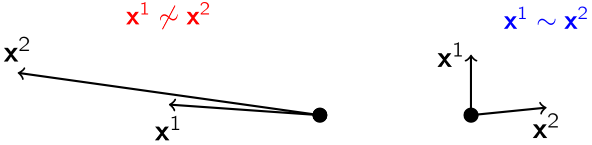
$$h(\mathbf{x}_i^1, \mathbf{x}_i^2) \gg 0$$

- If $\mathbf{x}_i^1 \nsim \mathbf{x}_i^2$ we want

$$h(\mathbf{x}_i^1, \mathbf{x}_i^2) \ll 0$$

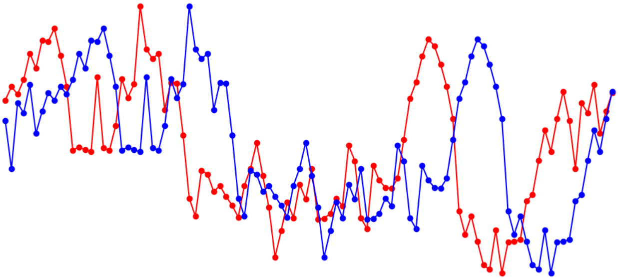
First Attempt

- 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



Shift Invariance

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



However, similarity score $h(\mathbf{u}, \mathbf{v}) \approx 0$

Bilinear Forms

Define

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

which amounts to

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

Example with $d = 4$

$\mathbf{u}$	u_1	u_2	u_3	u_4

$A_{1,1}$	$A_{1,2}$	$A_{1,3}$	$A_{1,4}$
$A_{2,1}$	$A_{2,2}$	$A_{2,3}$	$A_{2,4}$
$A_{3,1}$	$A_{3,2}$	$A_{3,3}$	$A_{3,4}$
$A_{4,1}$	$A_{4,2}$	$A_{4,3}$	$A_{4,4}$

v_1
v_2
v_3
v_4
$\mathbf{v}$

Surrogate Losses for Similarity Functions

- Example a pair $(\mathbf{x}^1, \mathbf{x}^2)$ and feedback $y \in \{-1, +1\}$
- Real-valued prediction $h(\mathbf{x}^1, \mathbf{x}^2) = \left(\mathbf{x}^1\right)^\top A\, \mathbf{x}^2$
- Hinge Loss with margin γ

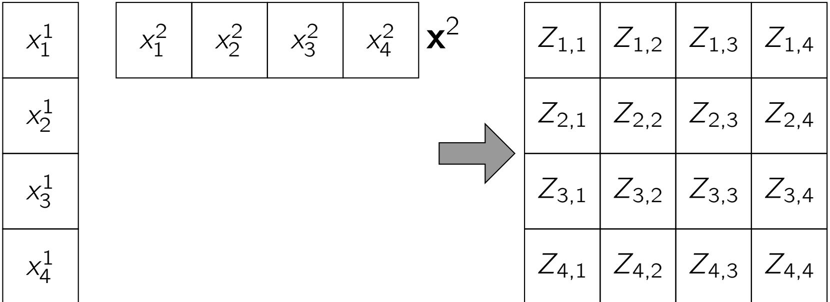
$$\left[\gamma - y \left(\mathbf{x}^1\right)^\top A\, \mathbf{x}^2\right]$$

- Logistic Loss with margin γ

$$\log\left(1 + \exp\left(\gamma - y \left(\mathbf{x}^1\right)^\top A\, \mathbf{x}^2\right)\right)$$

Alternative Formulation

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



- Remember, Z stands for Zorro  and not Zero

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

- Use SGD for ERM in order to find A

Skeleton of SGD for Similarity

Input: dataset of labeled pairs $S = \{\mathbf{x}_i^1, \mathbf{x}_i^2, y_i\}$

Transform: $\mathbf{x}_i^1, \mathbf{x}_i^2 \mapsto Z_i$ where $Z_i := \left(\mathbf{x}_i^1\right)\left(\mathbf{x}_i^2\right)^\top$

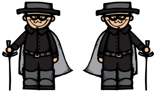
Loss: $f_i(A) = \ell(A \bullet Z_i)$ and $F(A) = \frac{1}{m}\sum_{i=1}^{m} f_i(A)$

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

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

Predict: for $(\tilde{\mathbf{x}}^1, \tilde{\mathbf{x}}^2)$ output $\operatorname{sign}\left(\left(\tilde{\mathbf{x}}^1\right)^\top A\, \tilde{\mathbf{x}}^2\right)$

Remarks

- It is often required that $h(\mathbf{x}^1, \mathbf{x}^2) = h(\mathbf{x}^2, \mathbf{x}^1)$
- Symmetric Zorro 

$$Z \stackrel{\text{def}}{=} \frac{1}{2} \left(\mathbf{x}^1\right) \left(\mathbf{x}^2\right)^\top + \frac{1}{2} \left(\mathbf{x}^2\right) \left(\mathbf{x}^1\right)^\top$$

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

$$\|\mathbf{x}\|_A^2 = \mathbf{x}^\top A\,\mathbf{x} \quad \Rightarrow \quad \|\mathbf{x}^1 - \mathbf{x}^2\|_A \stackrel{\text{def}}{=} \|\mathbf{v}\|_A \;\text{ where }\; \mathbf{v} = \mathbf{x}^1 - \mathbf{x}^2$$

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

$$A \in \{M \,:\, M \succcurlyeq 0\} \quad \text{where } M \succcurlyeq 0 \;\Leftrightarrow\; \forall \mathbf{v} : \mathbf{v}^\top M\,\mathbf{v} \geq 0$$

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