Nightmare at Test Time: Robust Learning by Feature Deletion

-- Amir Globerson and Sam Roweis

ICML'06

Why Robust Learning?

- Non-stationary feature distribution for training and test data
- Small Samples/ Imbalanced Class distribution
- Adversarial classification (Spam filtering)
- Data with Uncertainty
- A specific situation: A feature presented at training data but disappear (change to 0) in test data

Intuition of Robust Learning

- 3 stocks *A, B, C* with the same risk
- If you are going to investigate $3000 on stocks.
- Strategy 1: $3000->A
- Strategy 2: $1000->A, $1000->B, $1000->C
- Which one to choose?
- *Do not assign any feature with too much weight. (Regularization term like |w|^2 ??)*

Game Theory (Min-Max)

- Consider an adversarial situation:
- Two Players:
- P1: Build Classifier
- P2: Delete features during testing

- What's P1's policy?
- --Maximize the worst performance

For each instance x_i, the worst case hinge loss is:

$$h^{wc}(\mathbf{w}, y_i\mathbf{x}_i) = \begin{array}{ll} \max & [1 - y_i\mathbf{w} \cdot (\mathbf{x}_i \circ (1 - \boldsymbol{\alpha}_i))]_+ \\ s.t. & \boldsymbol{\alpha}_i \in \{0, 1\} \\ & \sum_j \alpha_{ij} = K \end{array}$$

For the whole data set, w should be

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i h^{wc}(\mathbf{w}, y_i\mathbf{x}_i)$$

$$h^{wc}(\mathbf{w}, y_i\mathbf{x}_i) \quad = \quad \left[1 - y_i\mathbf{w}^T\mathbf{x}_i + s_i\right]_+ \quad ,$$

where we have defined

$$s_i = \max_{\boldsymbol{\alpha}_i \in \{0,1\}, \sum \alpha_{ij} = K} y_i\mathbf{w} \cdot (\mathbf{x}_i \circ \boldsymbol{\alpha}_i)$$

Solution:

choose those features with maximal $y_i w x_{ij}$

The solution won't change if we relax \ alpha to be [0, 1]

New formulation

Involves both w and \alpha

$$
\begin{aligned}
s_i = \quad & \max \quad y_i \left(\mathbf{w} \circ \mathbf{x}_i\right) \cdot \boldsymbol{\alpha}_i \\
& s.t. \quad 0 \leq \boldsymbol{\alpha}_i \leq 1 \\
& \qquad \textstyle\sum_j \alpha_{ij} = K
\end{aligned}
$$

Dual Form:

$$
\begin{aligned}
s_i = \quad & \min \quad K z_i + \textstyle\sum_j v_{ij} \\
& s.t. \quad z_i + \mathbf{v}_i \geq \left(y_i \mathbf{x}_i \circ \mathbf{w}\right) \\
& \qquad \mathbf{v}_i \geq 0
\end{aligned}
$$

$$
\begin{array}{ll}
\min & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \left[1 - y_i \mathbf{w}^T \mathbf{x}_i + t_i\right]_+ \\
\text{s.t.} & t_i \geq K z_i + \sum_j v_{ij} \\
 & \mathbf{v}_i \geq 0 \\
 & z_i + \mathbf{v}_i \geq (y_i \mathbf{x}_i \circ \mathbf{w})
\end{array}
$$

Disssusion

- New Problem? Spam Filtering?
- Robust Learning favors keeping all the redundant features, how to run feature selection under robust learning scenario?