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: $3000 \to A$
- Strategy 2: $1000 \to A$, $1000 \to B$, $1000 \to C$
- 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:

\[
h_{wc}(w, y_i x_i) = \max \left[ 1 - y_i w \cdot (x_i \circ (1 - \alpha_i)) \right]_+ \\
s.t. \quad \alpha_i \in \{0, 1\} \\
\sum_j \alpha_{ij} = K
\]

For the whole data set, \( w \) should be

\[
w^* = \arg\min_w \frac{1}{2} \|w\|^2 + C \sum_i h_{wc}(w, y_i x_i)
\]
\[ h^{wc}(\mathbf{w}, y_i \mathbf{x}_i) = \left[ 1 - y_i \mathbf{w}^T \mathbf{x}_i + s_i \right]_+ , \]

where we have defined

\[ s_i = \max_{\alpha_i \in \{0, 1\}, \sum \alpha_{ij} = K} y_i \mathbf{w} \cdot (\mathbf{x}_i \circ \alpha_i) \]

**Solution:**

choose those features with maximal \( y_i \mathbf{w} \mathbf{x}_{i,j} \)

The solution won’t change if we relax \( \alpha \) to be \([0, 1]\)
New formulation

\[ s_i = \max_{\alpha_i} y_i \left( w \circ x_i \right) \cdot \alpha_i \]
\[ s.t. \quad 0 \leq \alpha_i \leq 1 \]
\[ \sum_j \alpha_{ij} = K \]

Dual Form:

\[ s_i = \min \quad K z_i + \sum_j v_{ij} \]
\[ s.t. \quad z_i + v_i \geq (y_i x_i \circ w) \]
\[ v_i \geq 0 \]
$$\min \quad \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_i \left[ 1 - y_i \mathbf{w}^T \mathbf{x}_i + t_i \right]_+$$

s.t.
$$t_i \geq K z_i + \sum_j v_{ij}$$
$$\mathbf{v}_i \geq 0$$
$$z_i + \mathbf{v}_i \geq (y_i \mathbf{x}_i \circ \mathbf{w})$$
Discussion

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