Nightmare at Test Time:
Robust Learning by Feature Deletion

Präsentieren:Yang yang

17.11.07

Outline

- Introduction
- FDROP:tractable quadratic program for training robust classifiers
- Experiments
  - Handwritten Digit Classification
  - Spam Filtering
- Summary

Introduction

- Testing comes ***after*** training
  - too much weight to any single input feature
    - with nonstationary feature distribution
    - with input sensor failure
- A common approach
  - Regularization which spreads the weight
  - Very generic and cannot iduce robustness

Introduction

- Solution
  - New algorithm
    - avoiding single feature over-weighting
    - Using quadratic programming
- The application of our methodes on
  - Handwritten digit recongnition
  - Spam filtering

Worst Case Deletion

Input:

- Labeled Sample (xi, yi) (i = 1,……., n),
- Feature vektor $\mathbf{x}_i \in \Re^d, y_i \in \{\pm 1\}$
- Number of features deleted from each sample point X: K

output:

- a linear classifier: $y(\mathbf{x}) = \mathrm{Sign}(\mathbf{w} \cdot \mathbf{x})$

Worst Case Deletion

Output:

- Performance Measure: Regularized hinge

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

- Hinge loss:

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

Hinge loss

- a convex upper bound on the zero

$$l_{zo}(\mathbf{w}, y, \mathbf{x}) \leq \sum_i \left[1 - y_i \mathbf{w} \cdot \mathbf{x}_i\right]_+$$

- Find **w** which minimizes the worst case hinge loss

$$\mathbf{w}^* \quad = \quad \arg\min_{\mathbf{w}} \sum_i h^{wc}(\mathbf{w}, y_i \mathbf{x}_i)$$

- Minimizing hinge loss→ minimizing on the training error

The FDROP

How can people Solving the minimax

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

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

Si is the maximum contribution of K features to the margin of sample **xi**

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

The FDROP

- The maximization problem for si has an LP

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

- Linear in all variables

$\underline{\text{FDROP:}}$

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

FDROP VS SVM

- FDROP is variant of SVM
  - linear classifier
  - the training objektiv is measured using a
  - regularized hinge loss

FDROP VS SVM

- FDROP is variant of SVM
  - differently error term compare to FDROP

<u>DSVM:</u>

$$\min \quad \frac{1}{2}\|\mathbf{w}\|^2 - \sum_i \alpha_i$$
$$\text{s.t.} \quad \mathbf{w} = \sum_i y_i \alpha_i \mathbf{x}_i$$
$$0 \leq \boldsymbol{\alpha} \leq C$$

<u>DFDROP:</u>

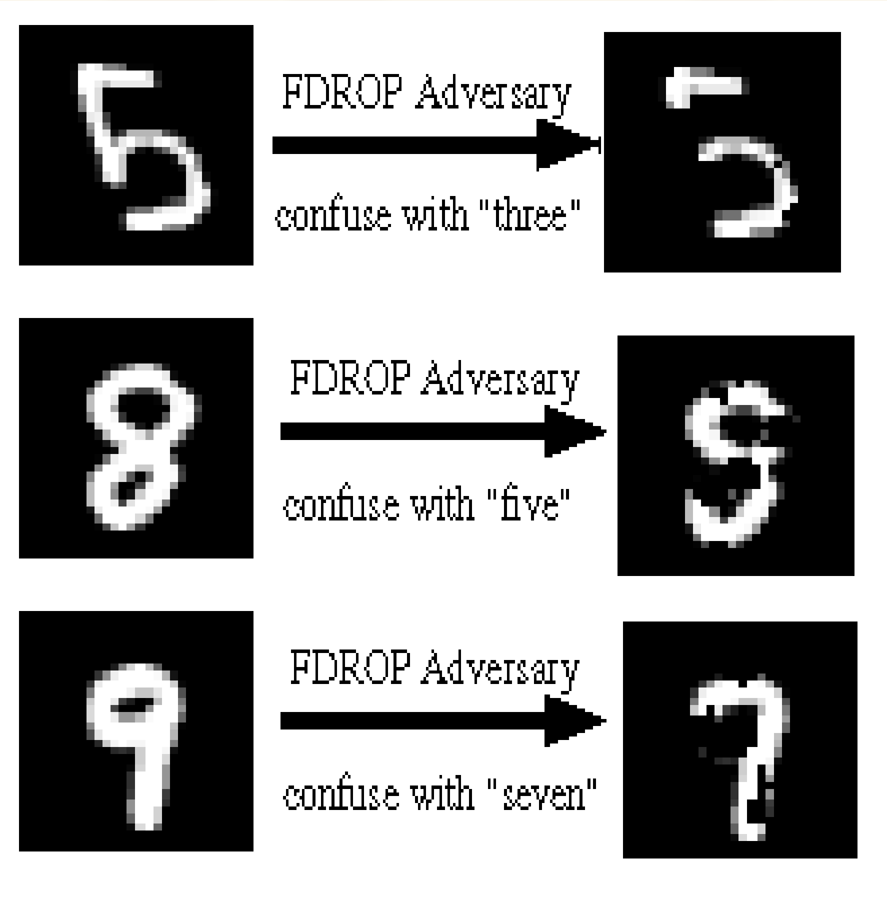
$$\min \quad \frac{1}{2}\|\mathbf{w}\|^2 - \sum_i \alpha_i$$
$$\text{s.t.} \quad \mathbf{w} = \sum_i y_i \alpha_i \mathbf{x}_i \circ (\mathbf{1} - \boldsymbol{\lambda}_\mathrm{i})$$
$$0 \leq \boldsymbol{\alpha} \leq C$$
$$0 \leq \boldsymbol{\lambda}_i \leq 1$$
$$\sum_j \lambda_{ij} = K$$

Handwritten Digit Classfication

- investigated the application of FDROP to classifying handwritten digits
- robustness to pixel deletion in these images
- Binary problems
- Small training sets of 50 samples per digit
- Chosen pairs which hard to distinguish
  - (5,3),(8,5),(9,7)....

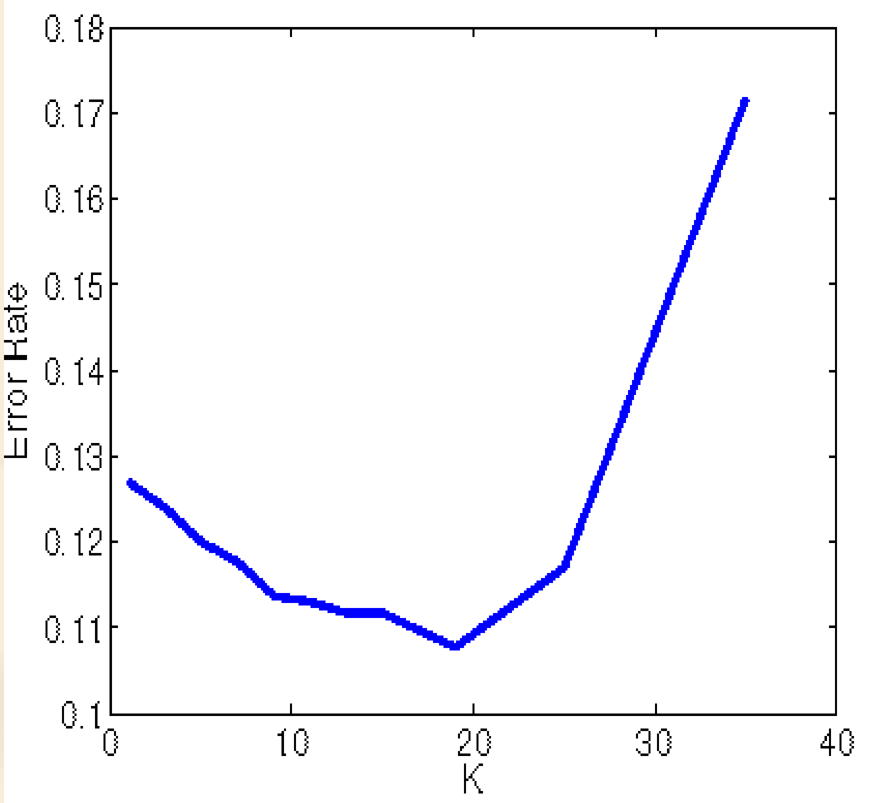
Handwritten Digit classification

- visual representation of the feature deletion process
- K destructive featuredeleted(K=50)
- maximize the resemblance between the given digit and the digit in the other class



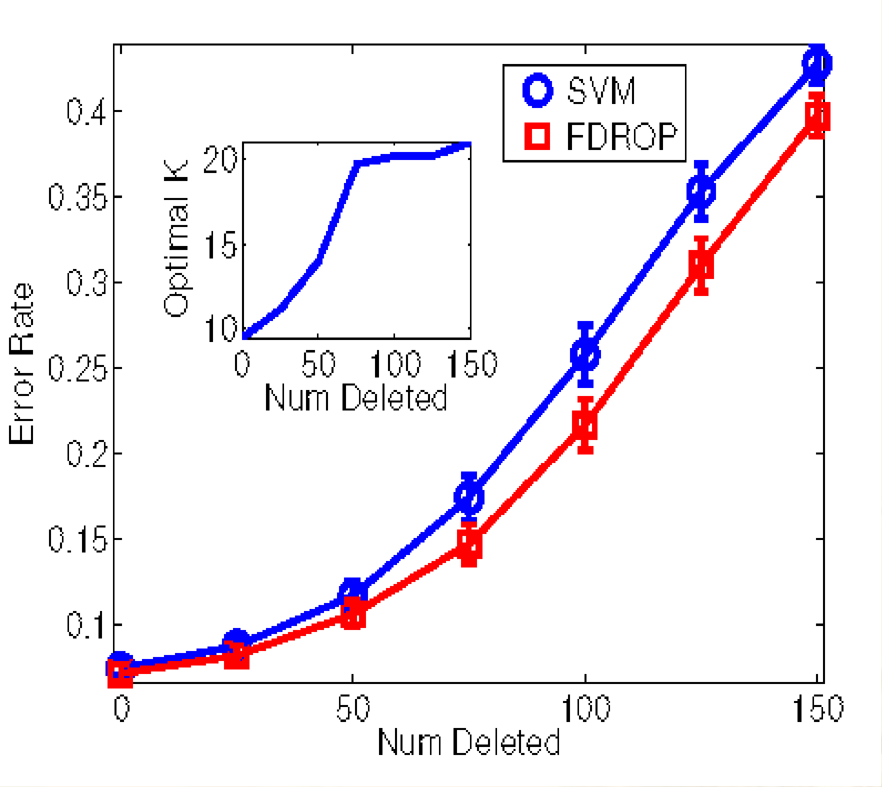
Handwritten Digit Classification

- Classification error for the digit pair (4; 7)
- K=50
- dependence on K
- e.g
  - Book and exam
  - howmuch book read in order to better point



Handwritten Digit Classification

- the dependence of classification error on the number of deleted features
- FDROP suffers less degradation in error when compared to SVM
- optimal K grows monotonously
- features dropped randomly



Summary

- Presented a new classification algorithm that is robust to worst case feature deletion
  - FDROP
  - Hinge loss
- Handwritten Digit classifiation

Discussion

Thanks for you attention