learning to compare

using operator-valued large-margin

classifiers

andreas maurer

a binary classification task for pairs

$\mathcal{X}$ = input space, embedded in a Hilbertspace H by a suitable kernel:
$\mathcal{X} \subset H$ and $diam(\mathcal{X}) \leq 1$.

.

.

.

a binary classification task for pairs

$\mathcal{X}$ = input space, embedded in a Hilbertspace H by a suitable kernel:
$\mathcal{X} \subset H$ and $diam(\mathcal{X}) \leq 1$.

ρ = a probability measure on $\mathcal{X}^2 \times \{-1, 1\}$, the *pair oracle*
$\rho(x, x', r)$ is the probability to encounter the two inputs $x, x' \in \mathcal{X}$ being
- homonymous (same label) for $r = 1$ and
- heteronymous (different labels) for $r = -1$.

A pair classifier is a function on $\mathcal{X}^2$ to predict the third argument of ρ.

.

.

a binary classification task for pairs

$\mathcal{X}$ = input space, embedded in a Hilbertspace H by a suitable kernel:
$\mathcal{X} \subset H$ and $diam(\mathcal{X}) \leq 1$.

ρ = a probability measure on $\mathcal{X}^2 \times \{-1, 1\}$, the *pair oracle*
$\rho(x, x', r)$ is the probability to encounter the two inputs $x, x' \in \mathcal{X}$ being
• homonymous (same label) for $r = 1$ and
• heteronymous (different labels) for $r = -1$.
A pair classifier is a function on $\mathcal{X}^2$ to predict the third argument of ρ.

$S = \left(\left(x_1, x_1', r_1\right), ..., (x_m, x_m', r_m)\right) \in \left(\mathcal{X}^2 \times \{-1, 1\}\right)^m$
training sample, generated in m independent, identical trials of ρ, i.e. $S \sim \rho^m$.

.

a binary classification task for pairs

$\mathcal{X}$ = input space, embedded in a Hilbertspace H by a suitable kernel:
$\mathcal{X} \subset H$ and $diam\,(\mathcal{X}) \leq 1$.

$\rho =$ a probability measure on $\mathcal{X}^2 \times \{-1, 1\}$, the *pair oracle*
$\rho\,(x, x', r)$ is the probability to encounter the two inputs $x, x' \in \mathcal{X}$ being
• homonymous (same label) for $r = 1$ and
• heteronymous (different labels) for $r = -1$.
A pair classifier is a function on $\mathcal{X}^2$ to predict the third argument of ρ.

$S = \left(\left(x_1, x_1', r_1\right), ..., (x_m, x_m', r_m)\right) \in \left(\mathcal{X}^2 \times \{-1, 1\}\right)^m$
training sample, generated in m independent, identical trials of ρ, i.e. $S \sim \rho^m$.

Goal: Use S to find a pair classifier with low error probability.

pair classifiers induced by linear transformations

We will select our classifiers from the hypothesis space

$$\left\{ f_T : \left(x, x'\right) \mapsto sgn\left(1 - \left\|Tx - Tx'\right\|\right) : T \in \mathcal{L}\left(H\right) \right\}$$

.

.

pair classifiers induced by linear transformations

We will select our classifiers from the hypothesis space

$$\left\{f_T : \left(x, x'\right) \mapsto sgn\left(1 - \left\|Tx - Tx'\right\|\right) : T \in \mathcal{L}(H)\right\}$$

A choice of $T \in \mathcal{L}(H)$ then implies a choice of

- the pair classifier f_T,
- the pseudo-metric $d(x, x') = \|Tx - Tx'\|$
- the Mahalanobis distance $d^2(x, x') = \langle T^*T(x - x'), x - x'\rangle$and
- the positive semidefinite kernel $\kappa(x, x') = \langle T^*Tx, x'\rangle$

.

pair classifiers induced by linear transformations

We will select our classifiers from the hypothesis space

$$\left\{ f_T : \left(x, x' \right) \mapsto sgn \left(1 - \left\| Tx - Tx' \right\| \right) : T \in \mathcal{L}\left(H \right) \right\}$$

A choice of $T \in \mathcal{L}\left(H \right)$ then implies a choice of

- the pair classifier f_T,
- the pseudo-metric $d\left(x, x' \right) = \left\| Tx - Tx' \right\|$
- the Mahalanobis distance $d^2\left(x, x' \right) = \left\langle T^*T\left(x - x' \right), x - x' \right\rangle$and
- the positive semidefinite kernel $\kappa\left(x, x' \right) = \left\langle T^*Tx, x' \right\rangle$

The risk of the operator T is the error probability of the classifier f_T

$$R\left(T \right) = \Pr_{\left(x, x', r \right) \sim \rho} \left\{ f_T\left(x, x' \right) \neq r \right\} = \Pr_{\left(x, x', r \right) \sim \rho} \left\{ r \left(1 - \left\| Tx - Tx' \right\|^2 \right) \leq 0 \right\}$$

estimation and generalization

Let $f : \mathbb{R} \to \mathbb{R}$, $f \geq 1_{(-\infty,0]}$ with Lipschitz constant L.
For a training sample $S = \left(\left(x_1, x_1', r_1\right), ..., (x_m, x_m', r_m)\right)$
define the empirical risk estimate

$$\hat{R}_f(T, S) = \frac{1}{m} \sum_{i=1}^{m} f\left(r_i \left(1 - \left\|T\left(x_i - x_i'\right)\right\|^2\right)\right).$$

.

estimation and generalization

Let $f : \mathbb{R} \to \mathbb{R}$, $f \geq 1_{(-\infty,0]}$ with Lipschitz constant L.
For a training sample $S = \left(\left(x_1, x_1', r_1\right), ..., (x_m, x_m', r_m)\right)$
define the empirical risk estimate

$$\hat{R}_f(T, S) = \frac{1}{m} \sum_{i=1}^{m} f\left(r_i\left(1 - \left\|T\left(x_i - x_i'\right)\right\|^2\right)\right).$$

Theorem: $\forall \delta > 0$, with probability greater $1 - \delta$ in a sample $S \sim \rho^m$
$\forall T \in \mathcal{L}(H)$ with $\|T^*T\|_2 \geq 1$

$$R(T) \leq \hat{R}_f(T, S) + \frac{8L\|T^*T\|_2 + \sqrt{\ln\left(2\|T^*T\|_2/\delta\right)}}{\sqrt{m}}.$$

where $\|A\|_2 = Tr(A^*A)^{1/2}$ is the Hilbert-Schmidt- or Frobenius- norm of A.

regularized objectives

The theorem suggests to minimize the regularized objective

$$\Lambda_{f,\lambda}(T) := \frac{1}{m}\sum_{i=1}^{m} f\left(r_i\left(1 - \left\|T\left(x_i - x_i'\right)\right\|^2\right)\right) + \frac{\lambda \left\|T^*T\right\|_2}{\sqrt{m}}.$$

Since $\|T^*T\|_2 \leq \|T\|_2^2$ we can also use $\|T\|_2^2$ as a stronger regularizer (computationally more efficient, but slightly inferior in experiments).

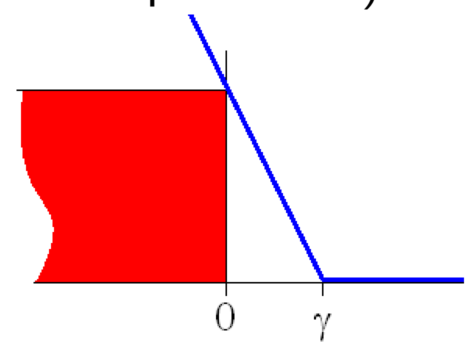
regularized objectives

The theorem suggests to minimize the regularized objective

$$\Lambda_{f,\lambda}(T) := \frac{1}{m}\sum_{i=1}^{m} f\left(r_i\left(1 - \left\|T\left(x_i - x_i'\right)\right\|^2\right)\right) + \frac{\lambda \|T^*T\|_2}{\sqrt{m}}.$$

Since $\|T^*T\|_2 \leq \|T\|_2^2$ we can also use $\|T\|_2^2$ as a stronger regularizer (computationally more efficient, but slightly inferior in experiments).

For f we take the hinge loss f_γ with margin γ :
f_γ has Lipschitz constant $1/\gamma$ and is convex.



Since $\|T(x - x')\|^2 = \langle T^*T(x - x'), x - x'\rangle$ is linear in T^*T, the objective $\Lambda_{f_\gamma,\lambda}(T)$ is a convex function of T^*T.

optimization problem

Find $T \in \mathcal{L}(H)$ to minimize

$$\Lambda_{f_\gamma,\lambda}(T) = \Omega(T^*T) = \frac{1}{m}\sum_{i=1}^{m} f_\gamma\left(r_i\left(1 - \left\|T\left(x_i - x_i'\right)\right\|^2\right)\right) + \frac{\lambda}{\sqrt{m}}\left\|T^*T\right\|_2.$$

$\Lambda_{f_\gamma,\lambda}$ is *not convex* in T, but Ω is convex in T^*T.

.

optimization problem

Find $T \in \mathcal{L}(H)$ to minimize

$$\Lambda_{f_\gamma,\lambda}(T) = \Omega(T^*T) = \frac{1}{m}\sum_{i=1}^{m} f_\gamma\left(r_i\left(1 - \left\|T\left(x_i - x_i'\right)\right\|^2\right)\right) + \frac{\lambda}{\sqrt{m}}\left\|T^*T\right\|_2.$$

$\Lambda_{f_\gamma,\lambda}$ is *not convex* in T, but Ω is convex in T^*T.

First possibility: Solve convex optimization problem for Ω on set of positive semidefinite operators by alternating projections (as in Xing et al.)
Then take square root operator to get T.

.

optimization problem

Find $T \in \mathcal{L}(H)$ to minimize

$$\Lambda_{f_\gamma,\lambda}(T) = \Omega(T^*T) = \frac{1}{m}\sum_{i=1}^{m} f_\gamma\left(r_i\left(1 - \left\|T\left(x_i - x_i'\right)\right\|^2\right)\right) + \frac{\lambda}{\sqrt{m}}\left\|T^*T\right\|_2 .$$

$\Lambda_{f_\gamma,\lambda}$ is *not convex* in T, but Ω is convex in T^*T.

First possibility: Solve convex optimization problem for Ω on set of positive semidefinite operators by alternating projections (as in Xing et al.)
Then take square root operator to get T.

Second possibility (my choice**):** Do gradient-descent of $\Lambda_{f_\gamma,\lambda}$ in T
No problems with local minima:
If T is a stable local minimizer of $\Lambda_{f_\gamma,\lambda}$,
then T^*T is a stable local minimizer of Ω.

algorithm

Given sample S, regularization parameter λ, margin γ, learning rate θ
initialize $\lambda' = \lambda/\sqrt{m}$ (where $m = |S|$)
initialize $T = (v_1, ..., v_m)$ (where the v_i are row-vectors)
repeat

  Compute $\|T^*T\|_2 = \left(\sum_{ij} \left\langle v_i, v_j \right\rangle^2\right)^{1/2}$
  For $i = 1, ..., d$ compute $w_i = 2\,\|T^*T\|_2^{-1} \sum_j \left\langle v_i, v_j \right\rangle v_i$
  Fetch (x, x', r) from sample S
  For $i = 1, ..., d$ compute $a_i \leftarrow \langle v_i, x - x' \rangle$
  Compute $b \leftarrow \sum_{i=1}^{d} a_i^2$
  If $r\,(1 - b) < \gamma$
    then for $i := 1, ..., d$ do $v_i \leftarrow v_i - \theta\left(\frac{r}{\gamma} a_i\,(x - x') + \lambda' w_i\right)$
    else for $i := 1, ..., d$ do $v_i \leftarrow v_i - \theta\lambda' w_i$

until convergence

experiments

with invariant character-recognition, spatial rotations (COIL100) and face recognition (ATT).

1. training T from one task/group of tasks
2. training nearest-neighbour test-classifiers with a *single* example/class on a test task, using both the input metric and the metric induced by T.
3. recording the error rates of the test classifiers

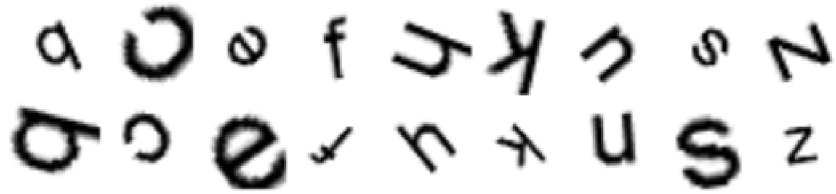
The pixel vectors x are embedded in the space H with the Gaussian rbf-kernel:

$$\kappa(x_1, x_2) = 2^{-1} \exp\left(-4 \left\| \frac{x_1}{\|x_1\|} - \frac{x_2}{\|x_2\|} \right\|^2\right).$$

The parameters $\gamma = 1$ and $\lambda = 0.05$ are used throughout.

rotation- and scale-invariant character recognition

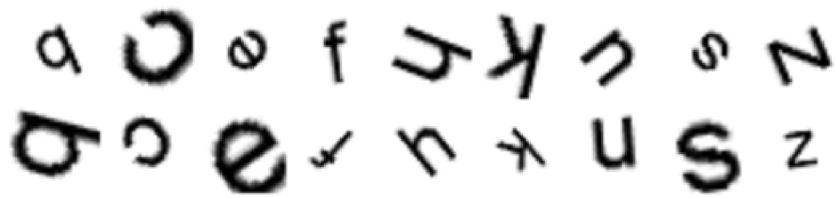
Typical pattern used to train the preprocessor (4000 examples from 20 classes)



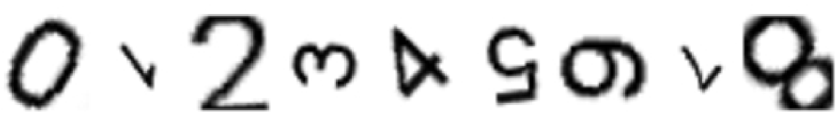
.

rotation- and scale-invariant character recognition

Typical pattern used to train the preprocessor (4000 examples from 20 classes)



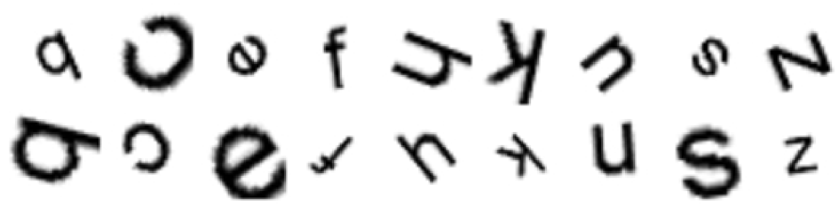
Nine digits used to train a single-nearest-neighbour classifier



.

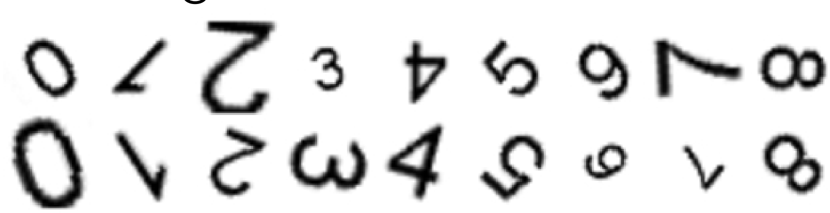
rotation- and scale-invariant character recognition

Typical pattern used to train the preprocessor (4000 examples from 20 classes)



Nine digits used to train a single-nearest-neighbour classifier

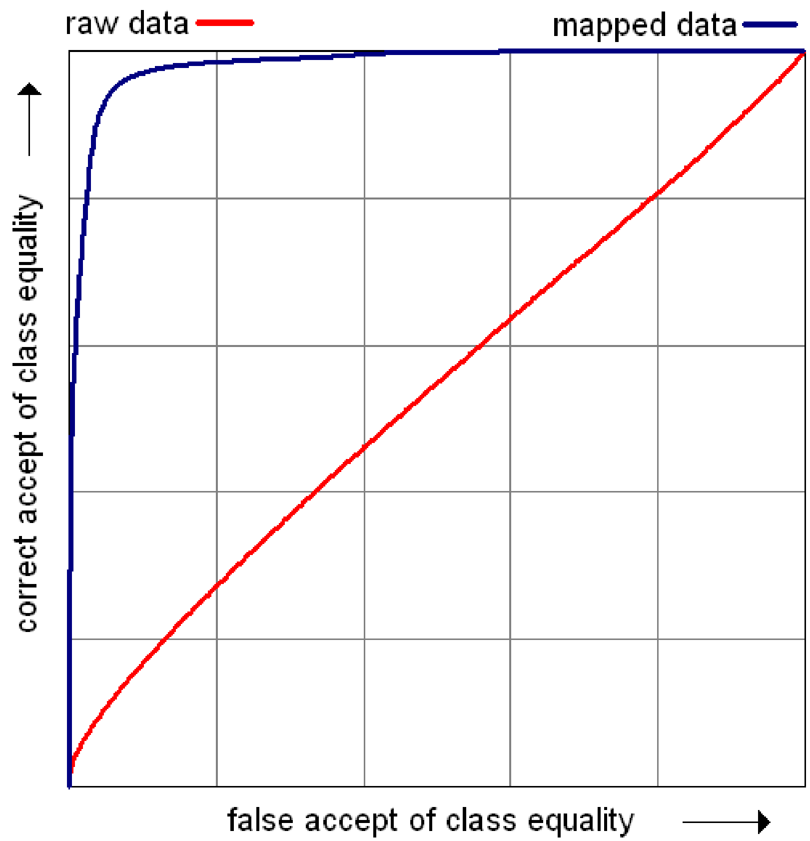
Some digits used to test the classifier:



results for rotation/scale-invariant OCR

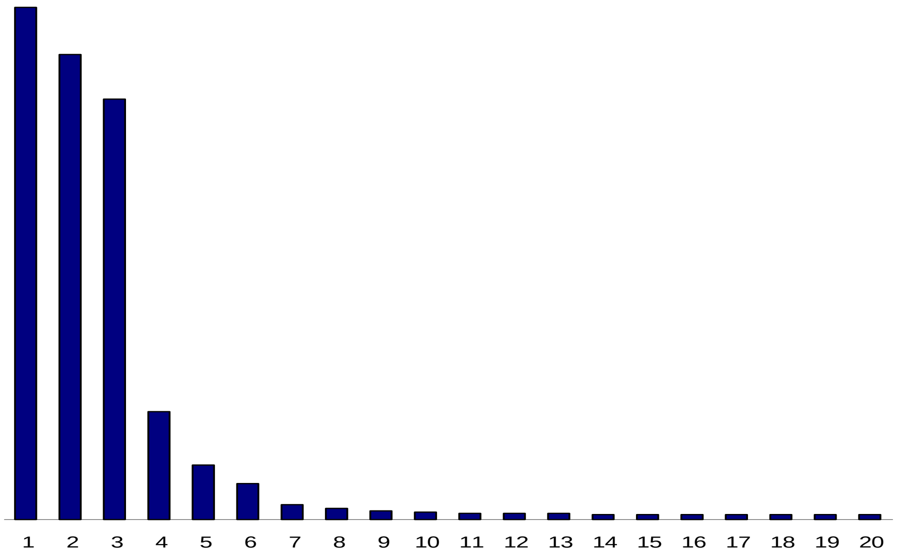
ROC-Area input	**0.539**
ROC-Area T	**0.982**
1-NN Error input	**0.822**
1-NN Error T	**0.093**

γ	1
σ	4
λ	0.005
Sample size	4000
Iterations	1000k



norms and singular-value-spectrum of T

$$\|T\|_1 = \quad 61.5$$
$$\|T\|_2 = \quad 27.7$$
$$\|T\|_\infty = \quad 17.3$$



Thank you!