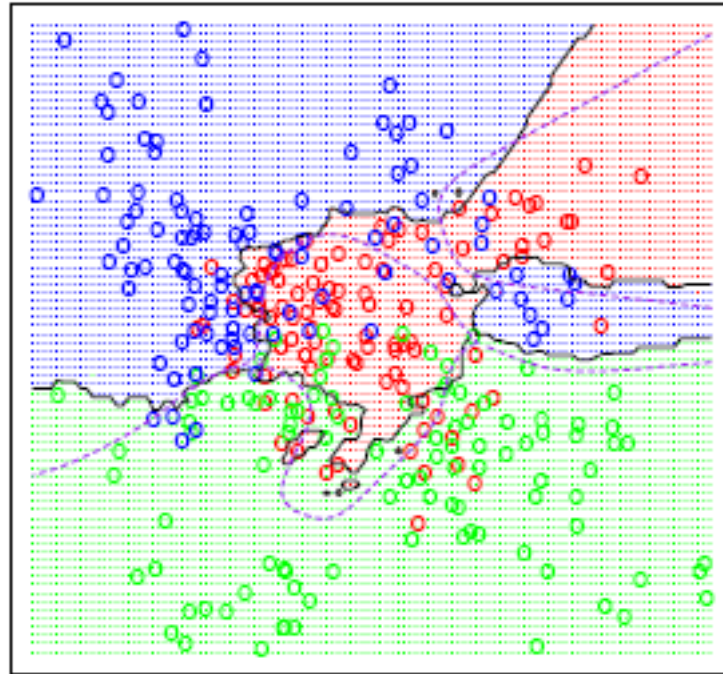


Distance Metric Learning for Large Margin Nearest Neighbor Classification

Daniel Arias Mutis

K - Nearest Neighbors



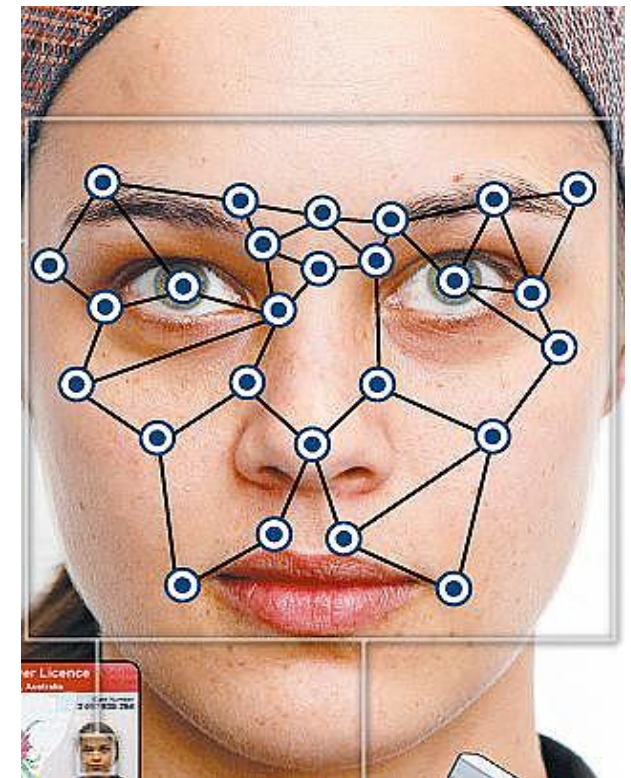
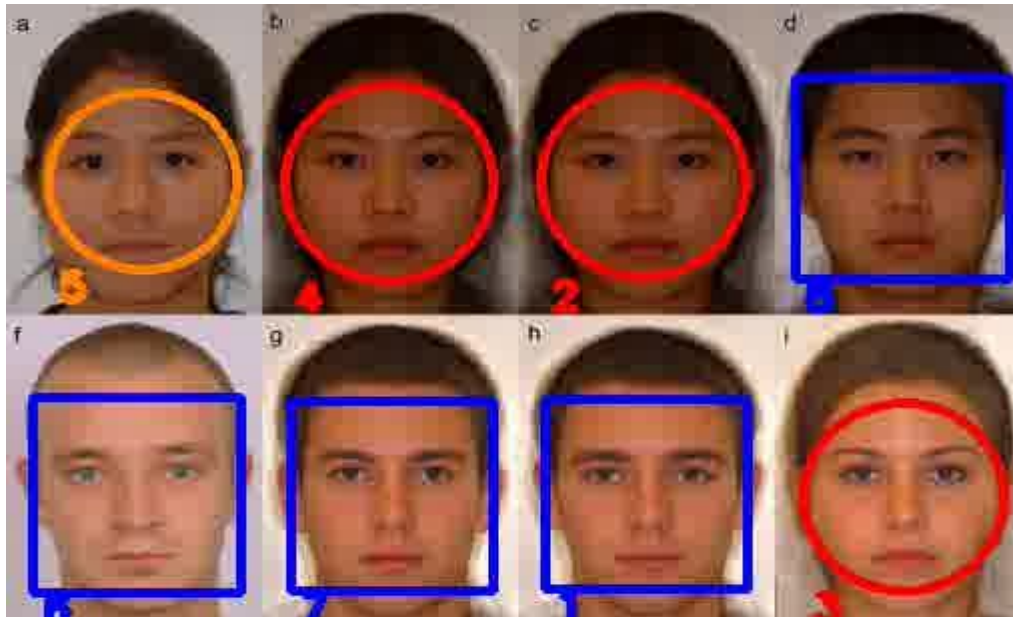
- classifies each unlabeled example by the majority label among its k-nearest neighbors training set.

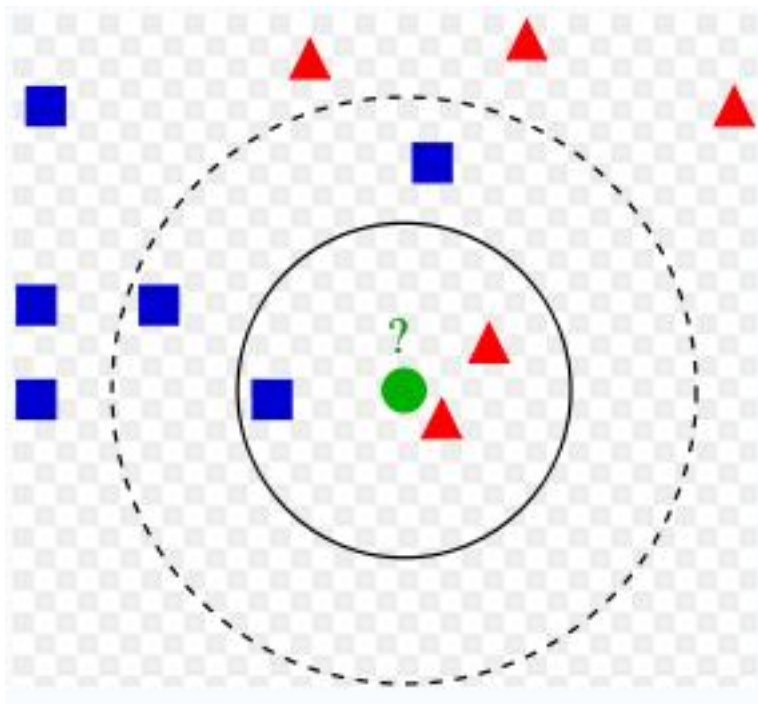
K - Nearest Neighbors

simple

competitive results

Previous knowledge (distance should be adapted to any particular case)





Do we like all our neighbors?



Large Margin Nearest Neighbor

*"The metric is optimized with the goal that **k-nearest neighbors** always belong to the same class while examples from different classes are separated by a large margin"*

improve: all similarly labeled inputs must be tightly clustered

Model

Training set $\{(\vec{x}_i, y_i)\}_{i=1}^n$ inputs $\vec{x}_i \in R^d$ with class labels y_i

the Matrix $y_{ij} \in \{0, 1\}$ indicate if y_i and y_j match

Learn the linear transformation $L : R^d \rightarrow R^d$, which will be used to compute:

$$D(\vec{x}_i, \vec{x}_j) = \|L(\vec{x}_i - \vec{x}_j)\|^2$$

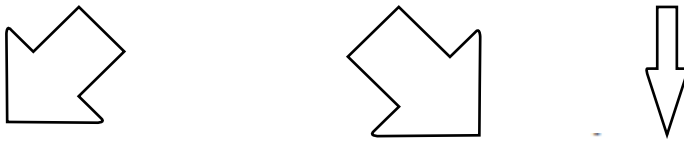
We specify k target neighbors (neighbors with the same label than \vec{x}_i)

Matrix $\eta_{ij} \in \{0, 1\}$ indicate whether \vec{x}_j is a target neighbor of \vec{x}_i

Cost Function

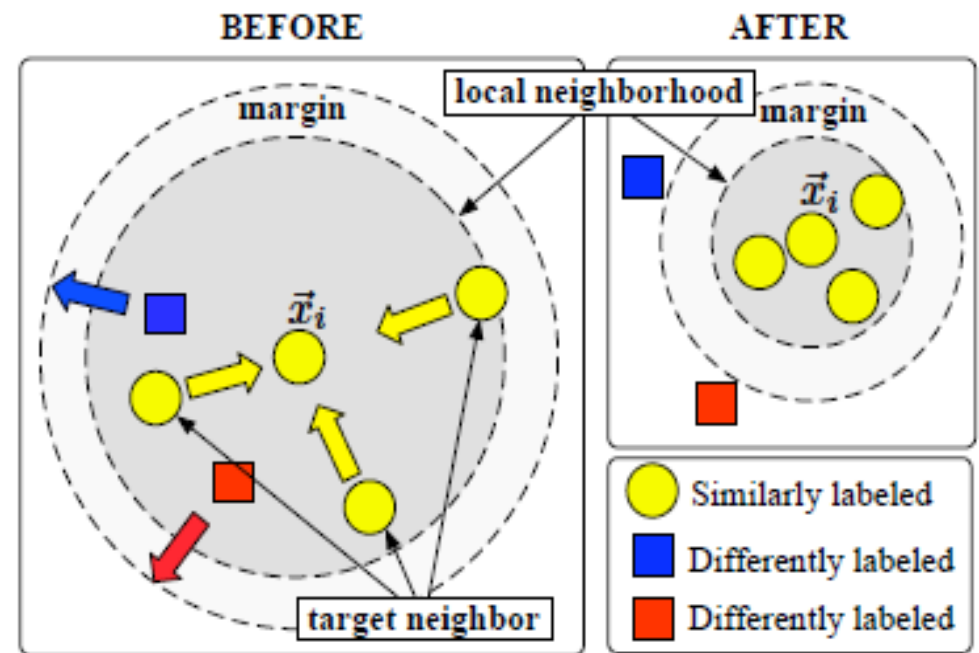
Target neighbors

match



$$\varepsilon(\mathbf{L}) = \sum_{ij} \eta_{ij} \|\mathbf{L}(\vec{x}_i - \vec{x}_j)\|^2 + c \sum_{ijl} \eta_{ij} (1 - y_{il}) [1 + \|\mathbf{L}(\vec{x}_i - \vec{x}_j)\|^2 - \|\mathbf{L}(\vec{x}_i - \vec{x}_l)\|^2]_+,$$

first term penalize large distance between inputs with the same label.
Second term (standard Hinge loss) penalize small distances between inputs with different labels



Convex optimization

semidefinite programming, (constrain the matrix whose elements are linear in the unknown variables to be positive semidefinite)

$$\mathcal{D}(\vec{x}_i, \vec{x}_j) = (\vec{x}_i - \vec{x}_j)^\top \mathbf{M} (\vec{x}_i - \vec{x}_j),$$

$\mathbf{M} = \mathbf{L}^\top \mathbf{L}$ parametrizes the Mahalanobis distance

Convex optimization

Slack variables for all variables with different labels i, j such that $y_{ij} \neq 0$.

Semidefinite program

$$\begin{aligned} & \text{Minimize } \sum_{i,j} \eta_{ij} (\vec{x}_i - \vec{x}_j)^\top \mathbf{M} (\vec{x}_i - \vec{x}_j) + c \sum_{i,j} \eta_{ij} (1 - y_{ij}) \xi_{ijl} \text{ subject to:} \\ & \quad (1) (\vec{x}_i - \vec{x}_l)^\top \mathbf{M} (\vec{x}_i - \vec{x}_l) - (\vec{x}_i - \vec{x}_j)^\top \mathbf{M} (\vec{x}_i - \vec{x}_j) \geq 1 - \xi_{ijl} \\ & \quad (2) \xi_{ijl} \geq 0 \\ & \quad (3) \mathbf{M} \succeq 0. \end{aligned}$$

Support Vector Machine

They are related: maximize borders.

SVM scales the time of training at least linearly with the number of classes.

KLMNN do not show an explicit dependence on the number of classes.

Experiments






From UCI machine learning repository

- Wine
- Iris
- Balance
- Isolet: letter name recognition.

Face Recognition

MNIST handwritten recognition

Test Image:	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8	8	9	9
Nearest neighbor after training:	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8	8	9	9
Nearest neighbor before training:	2	2	2	1	0	8	9	7	1	6	6	0	7	9	1	3	5	4	1

Test Image:							
Among 3 nearest neighbors after but not before training:							
Among 3 nearest neighbors before but not after training:							

Results

PCA are used to reduce dimensionality. of image, speech and text.

	Iris	Wine	Faces	Bal	Isolet	News	MNIST
examples (train)	106	126	280	445	6238	16000	60000
examples (test)	44	52	120	90	1559	2828	10000
classes	3	3	40	3	26	20	10
input dimensions	4	13	1178	4	617	30000	784
features after PCA	4	13	30	4	172	200	164
constraints	5278	7266	78828	76440	37 Mil	164 Mil	3.3 Bil
active constraints	113	1396	7665	3099	45747	732359	243596
CPU time (per run)	2s	8s	7s	13s	11m	1.5h	4h
runs	100	100	100	100	1	10	1

Table 1: Properties of data sets and experimental parameters for LMNN classification.

Results

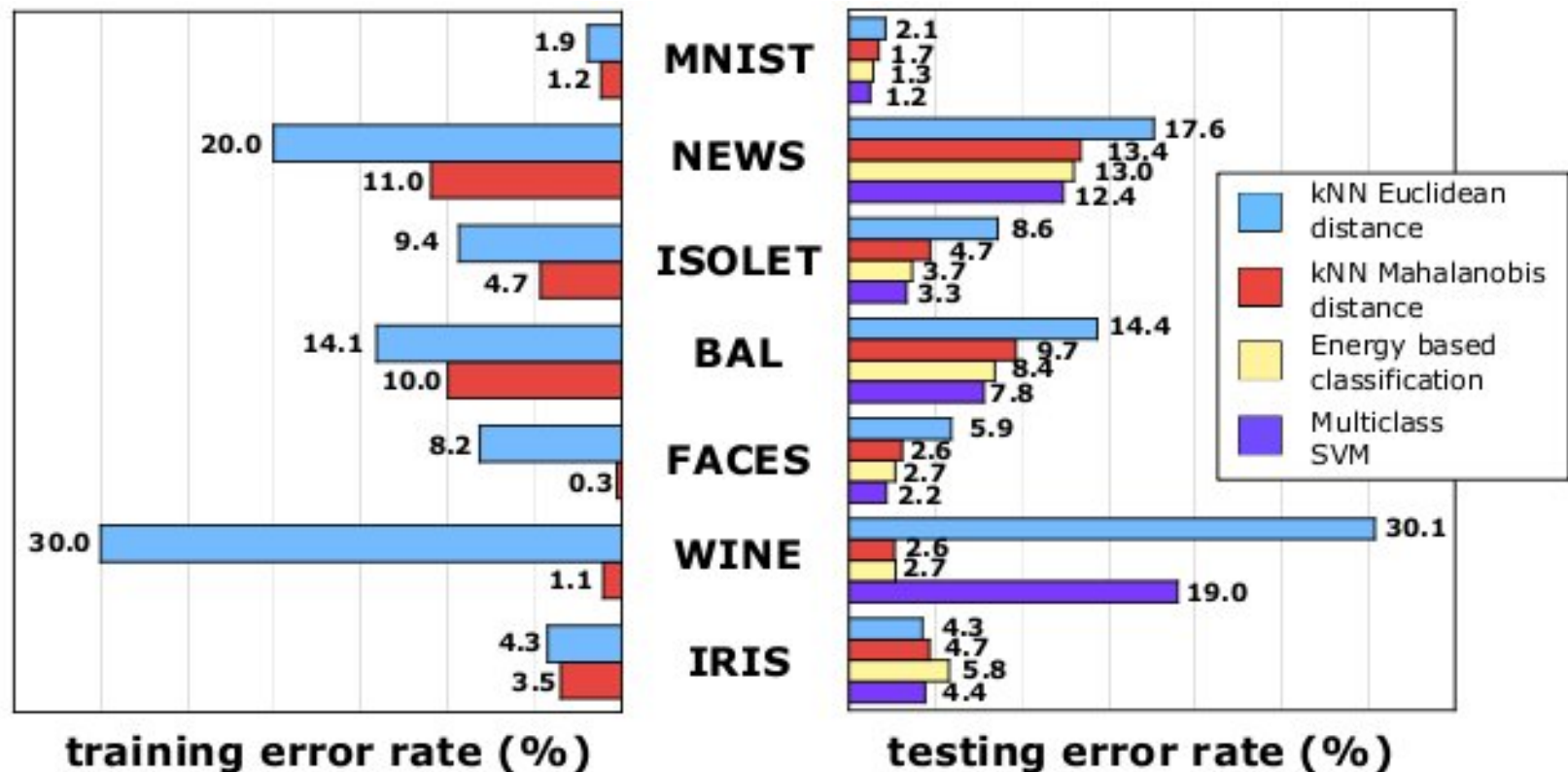


Figure 2: Training and test error rates for kNN classification using Euclidean versus Mahalanobis distances. The latter yields lower test error rates on all but the smallest data set (presumably due to over-training). Energy-based classification (see text) generally leads to further improvement. The results approach those of state-of-the-art multiclass SVMs.

Weinberger K Q, Blitzer J, and Saul L K. Advances in Neural Information Processing Systems, vol. 18, pp. 1473-1480, MIT Press, Cambridge, MA, 2006