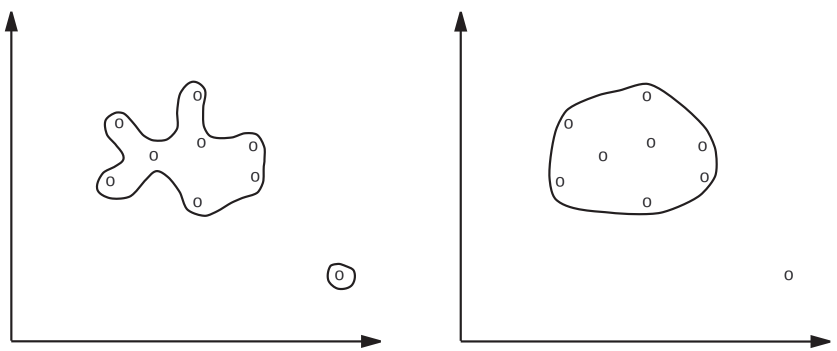
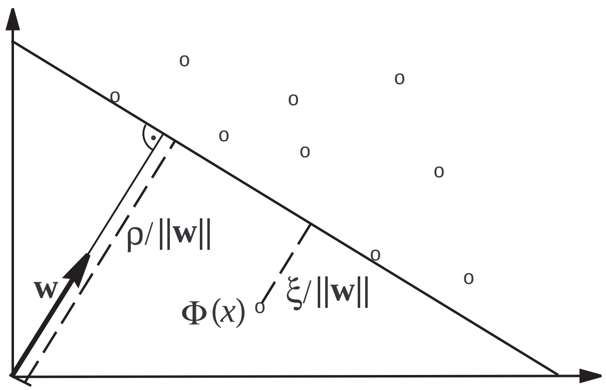
Unsupervised SVM Learning

$x_1, \ldots, x_m \in \mathcal{X}$ i.i.d. sample from P

- extreme view: unsupervised learning = density estimation
- easier problem: for $\alpha \in (0, 1]$, compute a region R such that
$$\mathrm{P}(R) \approx \alpha,$$
i.e., estimate *quantiles* of a distribution, not its density.
- becomes well-posed using a regularizer: find "smoothest" region that contains a certain fraction of the probability mass
- given only the training data, we will get a trade-off: try to enclose many training points (more than α) in a smooth region



ν-Soft Margin Separation



For $\nu \in (0, 1]$, compute

$$\min_{\mathbf{w}\in\mathcal{H},\boldsymbol{\xi}\in\mathbb{R}^m,\rho\in\mathbb{R}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{m}\sum_i \xi_i - \nu\rho$$

$$\text{subject to} \quad \langle \mathbf{w}, \Phi(x_i)\rangle \geq \rho - \xi_i, \quad \xi_i \geq 0 \quad \text{for all } i.$$

Result:

- the decision function $f(x) = \mathrm{sgn}(\langle \mathbf{w}, \Phi(x)\rangle - \rho)$ will be positive for "most" examples x_i contained in the training set
- $\|\mathbf{w}\|$ will be small, hence the separation from the origin large

Related approaches: enclose data in a sphere [49, 60]

Deriving the Dual Problem

Using multipliers $\alpha_i, \beta_i \geq 0$, we introduce a Lagrangian

$$L = \frac{\|\mathbf{w}\|^2}{2} + \frac{1}{\nu m}\sum_i \xi_i - \rho - \sum_i \alpha_i(\langle \mathbf{w}, \Phi(x_i)\rangle - \rho + \xi_i) - \sum_i \beta_i \xi_i,$$

and set the derivatives w.r.t. the primal variables $\mathbf{w}, \boldsymbol{\xi}, \rho$ equal to zero, yielding

$$\mathbf{w} = \sum_i \alpha_i \Phi(x_i), \tag{5}$$

$$\alpha_i = \frac{1}{\nu m} - \beta_i \leq \frac{1}{\nu m}, \tag{6}$$

$$\sum_i \alpha_i = 1. \tag{7}$$

Patterns with $\alpha_i > 0$ are Support Vectors.

Dual Problem

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^m} \quad \tfrac{1}{2}\textstyle\sum_{ij}\alpha_i\alpha_j k(x_i, x_j)$$

$$\text{subject to } 0 \le \alpha_i \le \tfrac{1}{\nu m}, \quad \textstyle\sum_i \alpha_i = 1.$$

The decision function is

$$f(x) = \operatorname{sgn}\left(\sum_i \alpha_i k(x_i, x) - \rho\right).$$

— a thresholded sparsified Parzen windows estimator

Support Vectors and Outliers

$$SV := \{i \mid \alpha_i > 0\}; \quad OL := \{i \mid \xi_i > 0\}$$

The KKT-Conditions imply:

- $\xi_i > 0 \Longrightarrow \alpha_i = 1/(\nu m)$, hence $OL \subset SV$
- $SV \setminus OL \subset \{i \mid \sum_j \alpha_j k(x_j, x_i) - \rho = 0\}$

The Meaning of ν

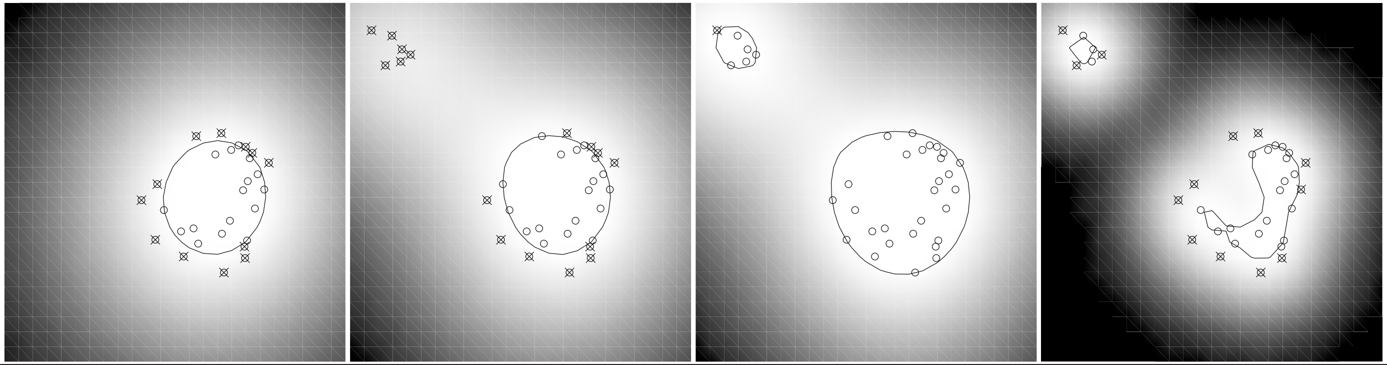
Proposition.
(i)

$$\frac{|OL|}{m} \leq \nu \leq \frac{|SV|}{m}$$

(ii) Suppose P *does not contain discrete components, and the kernel is analytic and non-constant. With probability 1, asymptotically,*

$$\frac{|OL|}{m} = \nu = \frac{|SV|}{m}.$$

Toy Examples using $k(x,y) = \exp(-\frac{\|x-y\|^2}{c})$



ν, width c	0.5, 0.5	0.5, 0.5	0.1, 0.5	0.5, 0.1
SVs/OLs	0.54, 0.43	0.59, 0.47	0.24, 0.03	0.65, 0.38

Variants of the Dual Problem, II

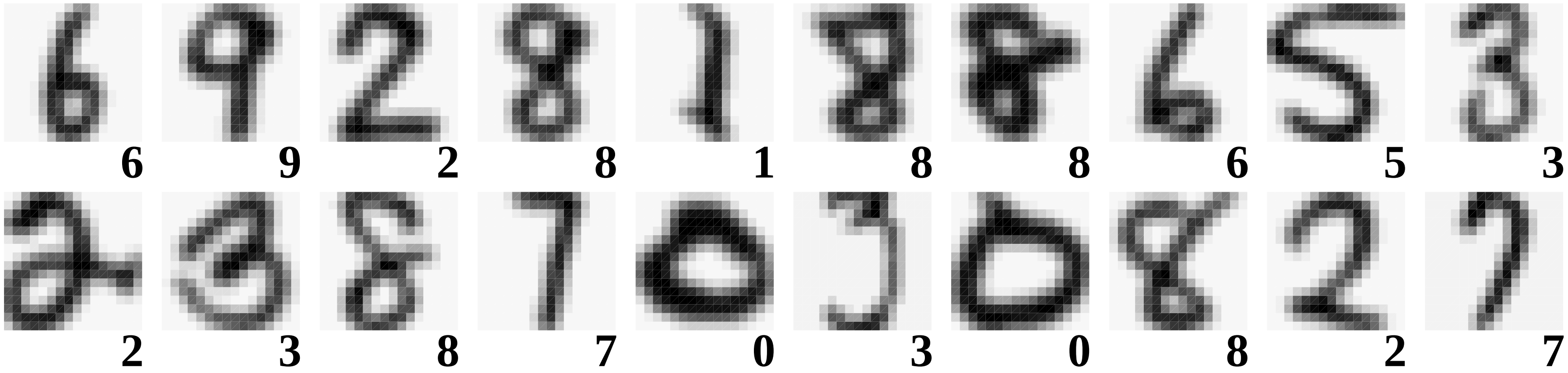
- can modify the approach to separate from some point other than the origin
- – the mean of some (scarce) negative data points
  – the mean of the training set ("quantile" PCA)

Resistance

Proposition 5 *Local movements of outliers parallel to* $\mathbf{w}$ *do not change the hyperplane.*

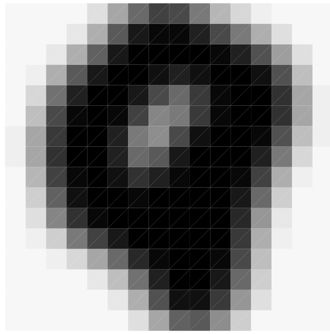
USPS Handwritten Digit Outlier Detection

Typical examples (random selection):

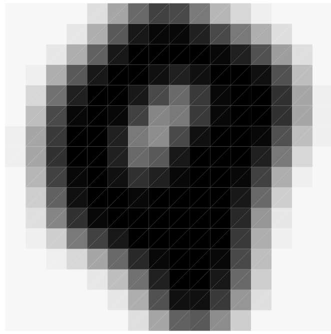


Present experiment: perform outlier detection on the 2007-element USPS test set (using $\nu = 5\%$)
(training time: < 1 min.)

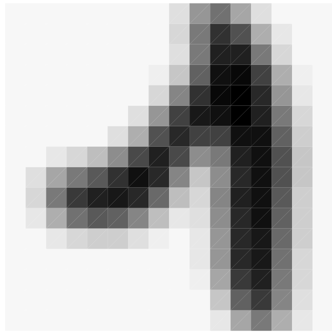
Next slides: the outliers, ranked by their "badness"



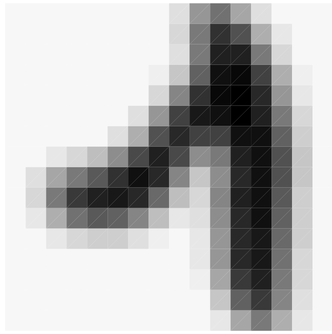
-513



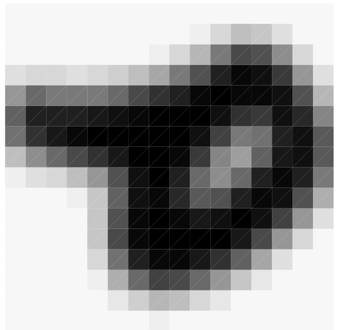
-513 **9**



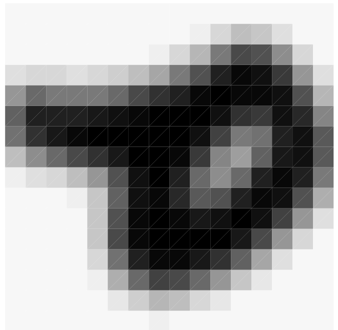
-507



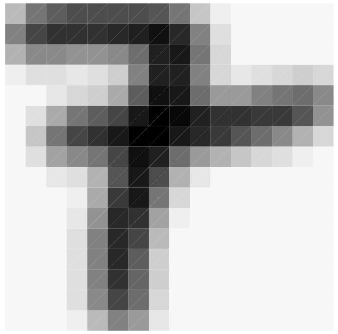
−507 **1**



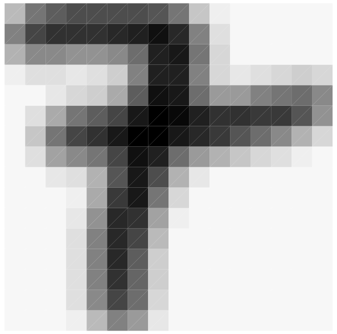
-458



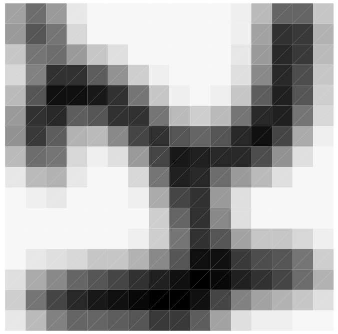
-458 **0**



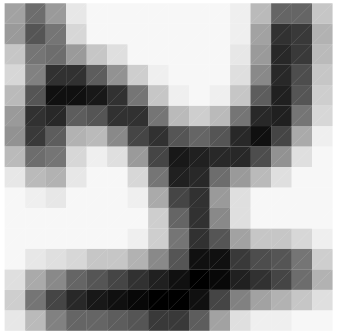
-282



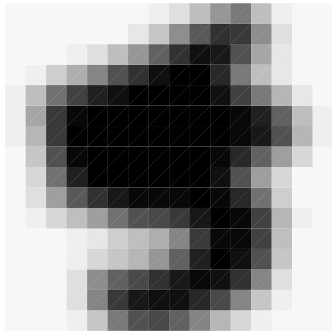
-282 **7**



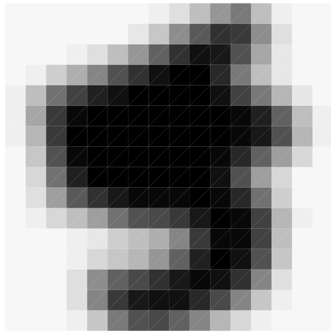
-216



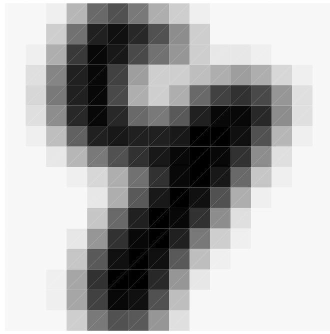
-216 **2**



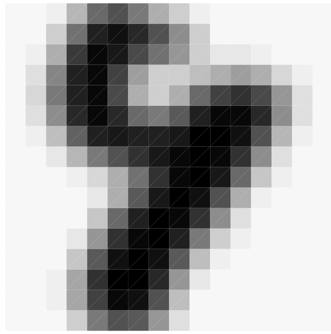
-200



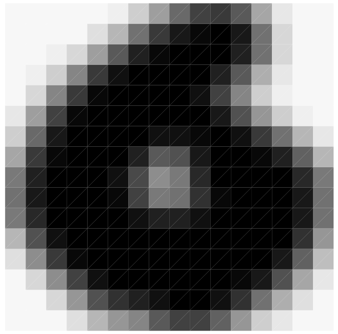
-200 **3**



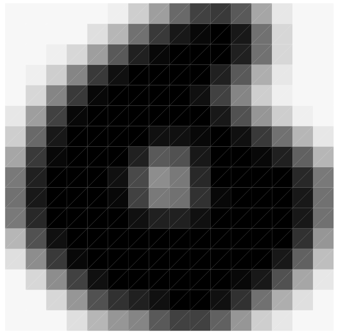
-186



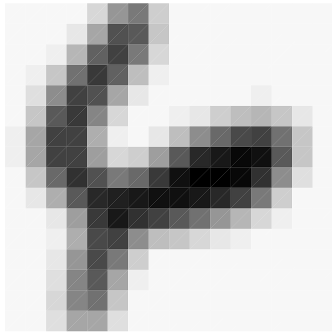
-186 **9**



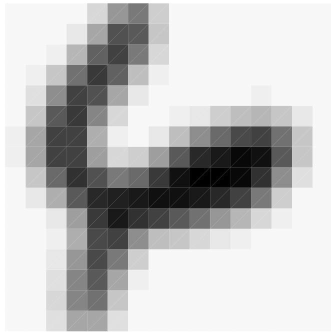
-162



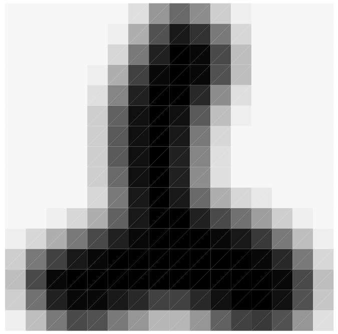
−162 **0**



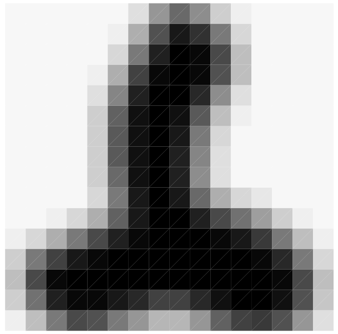
-143



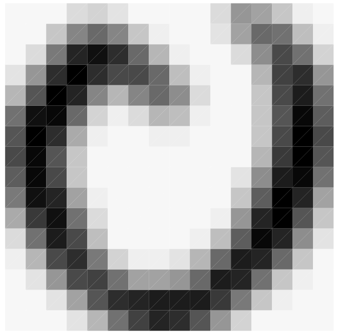
−143 **6**



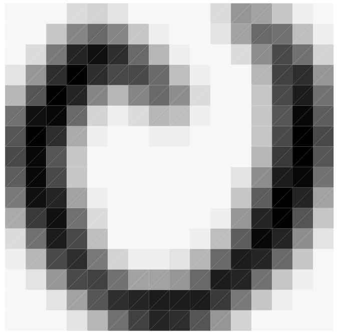
-128



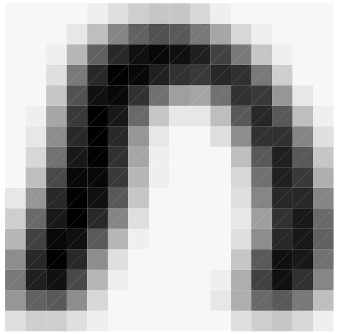
-128 **6**



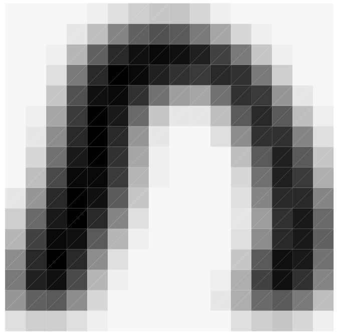
-93



-93 **5**



-78



-78 **0**

Other Applications

- Jet engine condition monitoring [29]
- Network intrusion detection (Wankadia et al., 2001)