Unsupervised SVM Learning

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

- extreme view: unsupervised learning $= \text{density estimation}$
- easier problem: for $\alpha \in (0, 1]$, compute a region $R$ such that
  $$P(R) \approx \alpha,$$
  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 $\alpha$) in a smooth region
\( \nu\)-Soft Margin Separation

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

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

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

Result:

- the decision function \( f(x) = \text{sgn}(\langle w, \Phi(x) \rangle - \rho) \) will be positive for “most” examples \( x_i \) contained in the training set
- \( \|w\| \) 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

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

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

$$w = \sum_i \alpha_i \Phi(x_i),$$

(5)

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

(6)

$$\sum_i \alpha_i = 1.$$  

(7)

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

B. Schölkopf, Canberra, February 2002
Dual Problem

$$\min_{\alpha \in \mathbb{R}^m} \frac{1}{2} \sum_{ij} \alpha_i \alpha_j k(x_i, x_j)$$
subject to $0 \leq \alpha_i \leq \frac{1}{\nu m}$, $\sum_i \alpha_i = 1$.

The decision function is

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

— a thresholded sparsified Parzen windows estimator

B. Schölkopf, Canberra, February 2002
Support Vectors and Outliers

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

The KKT-Conditions imply:

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

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 \( k(x, y) = \exp\left(-\frac{\|x-y\|^2}{c}\right) \)

<table>
<thead>
<tr>
<th>( \nu ), width ( c )</th>
<th>0.5, 0.5</th>
<th>0.5, 0.5</th>
<th>0.1, 0.5</th>
<th>0.5, 0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVs/OLs</td>
<td>0.54, 0.43</td>
<td>0.59, 0.47</td>
<td>0.24, 0.03</td>
<td>0.65, 0.38</td>
</tr>
</tbody>
</table>

B. Schölkopf, Canberra, February 2002
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”

B. Schölkopf, Canberra, February 2002
$-458$
$-458 \ 0$
$-282$
$-216$
$-216 \ 2$
$-200 \, 3$
$-186$
7

-186 9
−162
\[ -143 \]
-93
\[-78\]
Other Applications

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