ESTIMATING THE SUPPORT OF A HIGH-DIMENSIONAL DISTRIBUTION: SCHOLKOPF ET AL

VISHVAS VASUKI

1. Depth with which this was read, progress

Did not read and understand sections 2, 4, parts of 5, parts of 7 and appendix A well.

2. Problems

Given a sample drawn from a distribution drawn from a high dimensional space, how would you estimate its support?

3. Motivating real life scenarios

Novelty detection. They try out the example on the USPS hand-written digit database and seem to succeed in identifying outliers

4. The Model

4.1. **Terms and Variables used.**

4.2. **The specification of the problem.**

Find set S such that $Pr(x \notin S) < p \in (0, 1]$. Can be solved by probability density estimation techniques, but actually simpler.

5. Results, methods and ideas

5.1. **Using soft margin hyperplane in feature space corresponding to a kernel.** Given N examples $\{s_i\}$; project to some feature space associated with kernel $k(x, y) = \phi(x)^T\phi(y)$; want to find hyperplane $w^T\phi(x) - \rho$ such that all points in the support fall on one side of the hyperplane, outliers fall on the other side: support identifier $f = sgn(w^T x - \rho)$; so, allowing a soft margin, want to solve $\max_{\rho,w} \frac{\rho}{\|w\|} + C\sum\xi_i$ such that $w^T\phi(x_i) + \xi_i \geq \rho, \xi_i \geq 0$; $\equiv$ obj fn: $\min_{w,\xi,\rho} \|w\|^2/2 + \frac{1}{\nu N}\sum\xi_i - \rho$, for some coefficient $0 \leq \nu \leq 1$.

Thence get Lagrangian: $L(w, \xi, \rho, \alpha, \beta) = \|w\|^2/2 + \frac{1}{\nu N}\sum\xi_i - \rho - \sum\alpha_i(w^T\phi(x_i) + \xi_i - \rho) - \sum\beta_i\xi_i$ with $\alpha, \beta \geq 0$.

Set derivatives wrt primal vars w, ξ, ρ to 0 to get: $w = \sum_i \alpha_i\phi(x_i); \alpha_i = \frac{1}{\nu N} - \beta_i \leq \frac{1}{\nu N}, \sum_i \alpha_i = 1$. Thence, the support identifier becomes
$f = sgn(\sum_i \alpha_i k(x_i, x) - \rho)$; dual optimization problem becomes
$min_\alpha - 2^{-1}\sum_{i,j} a_i a_j k(x_i, x_j)$ subject to $0 \leq \alpha_i \leq (\nu N)^{-1}, \sum_i \alpha_i = 1$. Solving this, discover w; then recover ρ using $\rho = w^T\phi(x_i)$ for x_i with $\alpha_i \neq 0; \beta_i \neq 0$ (support vector with $\beta_i > 0$): $\exists x_i$ as $\sum\alpha_i = 1; \alpha_i \geq 0$.

5.1.1. *Choosing kernel, tuning parameters.* $\nu \propto$ softness of the margin, number of support vectors, thence the runtime, sensitivity to appearence of novelty.

With Gaussian kernel, any data set is seperable as everything is mapped to same quadrant in feature space.

5.1.2. *Comparison with thresholded Kernel Density estimator.* If $\nu = 1, \alpha_i = 1/N$, support identifier $f = sgn(\sum_i \alpha_i k(x_i, x) - \rho)$ same as one using a Kernel (Parzen) Density estimator. What happens when $\nu < 1$?

5.1.3. *Comparison with using soft margin balls in the feature space.* Here one solves: $\min_{R,\xi,c} R^2 + \frac{1}{\nu N} \sum_i \xi_i$ subject to $\left\|\phi(x_i - c)\right\|^2 - \xi_i \leq R^2, \xi_i \geq 0$.

After using the Lagrangian, finding the critical points and substituting, this leads to the dual $\min_\alpha \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j) - \sum_i \alpha_i k(x_i, x_i)$ subject to $0 \leq \alpha_i \leq \frac{1}{\nu N}, \sum \alpha_i = 1$, and the solution $c = \sum \alpha_i \phi(x_i)$ corresponding to support identifier $f = sgn(R^2 - \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j) + 2 \sum_i k(x_i, x) - k(x, x))$ [**Check**].

For homogenous kernels, $k(x, x)$ is a constant and the dual minimization problem and the support identifier is equivalent to the minimization problem derived from the hyperplane formulation. So, all mapped patterns lie in a sphere in feature space; finding the smallest sphere containing them is equivalent to finding the segment of the sphere containing the data points, which reduces to finding the separating hyperplane.

5.1.4. *Connection to binary classification.* Hyperplane $(w, \rho = 0)$
separates $\{(x_i, 1)\}$ from $(-x_i, -1)$ with margin $\rho / \left\|w\right\|$ and vice-versa.

6. Assumptions

7. What could have been tried to yield equivalent results

8. Open problems

[**OP**]:How to decide width of Gaussian kernel to use? Can you use information about the abnormal class in choosing the kernel?

9. Interesting facts and results from elsewhere

10. Comments on writing style

Well written. Some graphs could have used better explanation.

11. Questions