Problems in Novelty Detection

THE AUSTRALIAN NATIONAL UNIVERSITY

Problem

Depending on how we choose C, the number of points selected as lying on the "wrong" side of the hyperplane $H := \{x|\langle \mathbf{w}, \mathbf{x}\rangle = 1\}$ will vary.

But we would like to **specify a certain fraction** ν beforehand.

Example

In an alarm device, we want to make sure that it goes off at most, say, once a month without reason. So only one of all the measurements on average should be considered unusual.

Idea

If we could adjust the threshold, i.e. $f(\mathbf{x}) \geq \rho$ rather than $f(\mathbf{x}) \geq 1$, we might be able to choose the fraction of points adaptively.

Adaptive Threshold

THE AUSTRALIAN NATIONAL UNIVERSITY

Goal

We want to solve the original optimization problem (the one in the $\mathbf{w}$ and ξ_i) as well as possible and at the same time achieve a large margin ρ such that

$$f(\mathbf{x}_i) \geq \rho - \xi_i \text{ for all } 1 \leq i \leq m$$

Idea

Simply subtract ρ (with an additional scaling factor) from the original objective function. We obtain

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}(\xi_i - \nu\rho) \\ \text{subject to} \quad & \langle \mathbf{w}, \mathbf{x}_i\rangle \geq \rho - \xi_i \\ & \xi_i \geq 0 \end{aligned}$$

Eliminating C

THE AUSTRALIAN NATIONAL UNIVERSITY

Problem

We now have two parameters C and ν to adjust. This is rather messy. Let's get rid of one of them.

Theorem

Denote by $\mathbf{w}, \rho$ the solution of

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{m}(\xi_i - \nu\rho) \\ \text{subject to} \quad & \langle \mathbf{w}, \mathbf{x}_i\rangle \geq \rho - \xi_i \text{ and } \xi_i \geq 0 \end{aligned}$$

and by $\mathbf{w}_C, \rho_C$ the solution

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}(\xi_i - \nu\rho) \\ \text{subject to} \quad & \langle \mathbf{w}, \mathbf{x}_i\rangle \geq \rho - \xi_i \text{ and } \xi_i \geq 0 \end{aligned}$$

Then $\mathbf{w}_C = C\mathbf{w}$ and $\rho_C = C\rho$.

Proof

THE AUSTRALIAN NATIONAL UNIVERSITY

Intuitive Reasoning

In the objective function the derivative with respect to $\mathbf{w}$ leaves the first term identical and the second term (specifying the kernel expansion) is multiplied by C. Moreover, the constraints can just be rescaled.

Formal Proof

We compute the Lagrangian.

$$L(\mathbf{w}, \rho, \xi) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}(\xi_i - \nu\rho) + \sum_{i=1}^{m}\alpha_i(\rho - \xi_i - \langle \mathbf{w}, \mathbf{x}_i\rangle) - \sum_{i=1}^{m}\eta_i\xi_i$$

This leads to the optimality conditions in $\mathbf{w}, \rho, \xi$ as follows

$$\partial_{\mathbf{w}}L = \mathbf{w} - \sum_{i=1}^{m}\alpha_i\mathbf{x}_i = 0, \quad \partial_{\rho}L = \sum_{i=1}^{m}(-C\nu + \alpha_i) = 0, \quad \partial_{\xi_i}L = C - \alpha_i - \eta_i$$

Now assume we have a solution for $C_0 = 1$. Then clearly rescaling all variables by C will lead to a feasible solution for $C \neq 1$. So, without loss of generality we always set $C = 1$.

Optimization Problem for Novelty Detection

THE AUSTRALIAN NATIONAL UNIVERSITY

Rewriting the Lagrangian

Sorting out all the dependencies in $\mathbf{w}, \rho$, and ξ_i we obtain

$$L(\mathbf{w}, \rho, \xi) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{m} \alpha_i \langle \mathbf{w}, \mathbf{x}_i \rangle - \rho \sum_{i=1}^{m} (\nu - \alpha_i) + \sum_{i=1}^{m} \xi_i (1 - \alpha_i - \eta_i)$$

with the saddle point conditions

$$\partial_{\mathbf{w}} L = \mathbf{w} - \sum_{i=1}^{m} \alpha_i \mathbf{x}_i = 0, \quad \partial_{\rho} L = \sum_{i=1}^{m} (-\nu + \alpha_i) = 0, \quad \partial_{\xi_i} L = 1 - \alpha_i - \eta_i$$

Eliminating Primal Variables

The green and blue terms vanish and we obtain (after a sign change)

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \\ \text{subject to} \quad & \sum_{i=1}^{m} \alpha_i = m\nu \text{ and } \alpha_i \in [0, 1] \end{aligned}$$

Proof of ν-Property

THE AUSTRALIAN NATIONAL UNIVERSITY

Theorem

- At least a fraction of ν points will lie on the "wrong" side of the hyperplane, i.e. $\langle \mathbf{w}, \mathbf{x}_i \rangle \leq \rho$.
- At least a fraction of $1 - \nu$ points will satisfy $\langle \mathbf{w}, \mathbf{x}_i \rangle \geq \rho$.

Proof

All we have to do is check the constraints $\sum_{i=1}^{m} \alpha_i = m\nu$ and $\alpha_i \in [0, 1]$.

- Since only Support Vectors, i.e. points with $f(\mathbf{x}_i) \leq \rho$ have nonzero α_i and simultaneously $\alpha_i \in [0, 1]$, we need at least $\lceil \nu m \rceil$ points to satisfy $\sum_i \alpha_i = \nu m$. This proves the first part.
- Since only for those points where $f(\mathbf{x}_i) < \rho$ the Lagrange multipliers $\alpha_i = 1$, we cannot have more than $\lfloor \nu m \rfloor$ of them. This shows the second part.

Classification

THE AUSTRALIAN NATIONAL UNIVERSITY

Idea

We make the width of the margin which we so far set to 1 a variable of the optimization problem, i.e.

$$y_i f(\mathbf{x}_i) \geq \rho \text{ instead of } y_i f(\mathbf{x}_i) \geq 1$$

Primal Objective Function

Now we have to make ρ part of the optimization problem. Since we want a *large* margin we have to *subtract* it from the original objective function and we obtain

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|^2 + \sum_{i=1}^{m} (\xi_i - \nu\rho) \\ \text{subject to} \quad & y_i(\langle w, x_i \rangle + b) - \rho + \xi_i \geq 0 \text{ for all } 1 \leq i \leq m \end{aligned}$$

As in novelty detection we may eliminate C (see problem sheet).

Dual Problem

THE AUSTRALIAN NATIONAL UNIVERSITY

Lagrange Function

$$L(\mathbf{w}, b, \xi, \rho, \alpha, \eta) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{m} (\xi_i - \nu\rho) + \sum_{i=1}^{m} \alpha_i \left(\rho - \xi_i - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)\right) - \sum_{i=1}^{m} \eta_i \xi_i$$

Saddle Point Conditions

$$\begin{aligned} \partial_{\mathbf{w}} L &= \mathbf{w} - \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i = 0 &&\iff \mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i \\ \partial_b L &= \sum_{i=1}^{m} -\alpha_i y_i = 0 &&\iff \sum_{i=1}^{m} \alpha_i y_i = 0 \\ \partial_{\xi_i} L &= 1 - \alpha_i - \eta_i = 0 &&\iff \alpha_i \in [0, 1] \text{ with } \eta_i = 1 - \alpha_i \\ \partial_{\rho} L &= \sum_{i=1}^{m} (\alpha_i - \nu) = 0 &&\iff \sum_{i=1}^{m} \alpha_i = m\nu \end{aligned}$$

Dual Problem

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \\ \text{subject to} \quad & \alpha_i \in [0, 1], \ \textstyle\sum_{i=1}^{m} \alpha_i y_i = 0 \text{ and } \sum_{i=1}^{m} \alpha_i = \nu m \end{aligned}$$

Interpretation

THE AUSTRALIAN NATIONAL UNIVERSITY

ν-Property

As in novelty detection, the ν-property holds, i.e. we have at least νm points on or beyond the margin and at most $(1-\nu)m$ points on the "right" side of the margin. This follows directly from the summation constraint on α_i.

Advantage

Now we can specify the training error beforehand (or at least the margin error). This should be in the order of the error we can already expect (e.g. human error rate on OCR, previous results that other algorithms got, etc.).

This is much easier than using the C parameter.

Additional Constraint

The price we pay is an additional summation constraint. Furthermore, to obtain b, ρ we have to use two observations and solve a linear system.

Practical Trick — we can get b, ρ directly from a quadratic optimizer as the dual variables to the constraints (dual dual = primal).

Regression

THE AUSTRALIAN NATIONAL UNIVERSITY

Problem

The precision ε has to be specified beforehand. This is rather tricky, since we usually do not know the noise within the y_i.

Idea

We make the width of the ε-insensitive tube a variable of the optimization problem and try to minimize ε along with the other variables.

Primal Objective Function

Since we want a *small* tube we have to *add* ε to the original objective function

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^* + \nu\varepsilon) \\ \text{subject to} \quad & (\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq y_i - \varepsilon - \xi_i \\ & (\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \leq y_i + \varepsilon + \xi_i \\ & \xi_i, \xi_i \geq 0 \text{ for all } 1 \leq i \leq m \end{aligned}$$

Unllike in novelty detection we cannot eliminate C: y_i breaks the scaling freedom.

Lagrange Function + Saddle Point

THE AUSTRALIAN NATIONAL UNIVERSITY

Lagrange Function

$$\begin{aligned} L(\mathbf{w}, b, \varepsilon, \xi, \xi^*, \alpha, \alpha^*, \eta, \eta^*) = & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^* + \nu\varepsilon) - \sum_{i=1}^{m}(\eta_i\xi_i + \eta_i^*\xi_i^*) \\ & + \sum_{i=1}^{m}\alpha_i^*\left((\langle \mathbf{w}, \mathbf{x}_i\rangle + b) - y_i - \varepsilon - \xi_i^*\right) \\ & + \sum_{i=1}^{m}\alpha_i\left(y_i - \varepsilon - \xi_i - (\langle \mathbf{w}, \mathbf{x}_i\rangle + b)\right) \end{aligned}$$

Saddle Point Conditions

The partial derivatives of L with respect to $\mathbf{w}, b, \varepsilon, \xi, \xi^*$ have to vanish. This leads to the usual results, plus

$$\partial_\varepsilon L = Cm\nu - \sum_{i=1}^{m}\alpha_i^* + \alpha_i = 0.$$

This translates into one more equality constraint of the dual optimization problem.

Dual Problem

THE AUSTRALIAN NATIONAL UNIVERSITY

Quadratic Program

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\sum_{i,j=1}^{m}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)k(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)y_i \\ \text{subject to} \quad & \sum_{i=1}^{m}(\alpha_i + \alpha_i^*) = C\nu m \\ & \sum_{i=1}^{m}(\alpha_i - \alpha_i^*) = 0 \\ & \alpha_i, \alpha_i^* \in [0, C] \end{aligned}$$

Interpretation

We have one term less in the objective function and one more summation constraint in the optimization problem.

Again, the ν-property holds, i.e. only a fraction of at most νm points lies **outside** the ε-insensitive tube, and a fraction of at most $(1-\nu)m$ points inside.

Using a QP Optimizer

Typical QP Optimizer

It solves a quadratic program of the following type

$$\begin{aligned} \text{minimize} \quad & c^\top x + \frac{1}{2} x^\top Q x \\ \text{subject to} \quad & l_i \leq x_i \leq u_i \text{ and } Ax + b = 0 \end{aligned}$$

Rewriting the SV Optimization Problem

We explain the case of classification. There we had

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \\ \text{subject to} \quad & \alpha_i \in [0,1],\ \textstyle\sum_{i=1}^m \alpha_i y_i = 0 \text{ and } \sum_{i=1}^m \alpha_i = \nu m \end{aligned}$$

We identify the terms: set $Q_{ij} := y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$ with $Q \in \mathbb{R}^{m \times m}$, $c_i = 0$ with $c \in \mathbb{R}^m$, $l_i = 0, u_i = 1$ with $l, u \in \mathbb{R}^m$. For the constraints use

$$A_{i1} = y_i \text{ and } A_{i2} = 1 \text{ where } A \in \mathbb{R}^{m \times 2} \text{ and } b = (0, -\nu m)$$