# Midsem solutions

Tuesday 1st March, 2016

**Problem 1. Support Vector Regression:**

1. If all training data points lie strictly inside the $\epsilon$-band of the SVR, then for all $i$, $\xi_i = \xi_i^* = 0$ and using basic knowledge of SVR, we know that for all such points, $y_i - w^\top \phi(x_i) - b < \epsilon + \xi_i$ and $b + w^\top \phi(x_i) - y_i < \epsilon + \xi_i^*$. That is, $y_i - w^\top \phi(x_i) - b - \epsilon - \xi_i < 0$ and $b + w^\top \phi(x_i) - y_i - \epsilon - \xi_i^* < 0$.

   Since $\alpha_i(y_i - w^\top \phi(x_i) - b - \epsilon - \xi_i) = 0$ and $\alpha_i^*(b + w^\top \phi(x_i) - y_i - \epsilon - \xi_i^*) = 0$, we must have for all $i$, $\alpha_i = \alpha_i^* = 0$.

   $\Rightarrow w = \sum_{i=1}^n (\alpha_i - \alpha_i^*)\phi(x_i) = 0$

   Thus, the regression line will simply be $f(x) = b$, the bias term! In the case of a single dimensional $\phi(x)$, this will mean that $f(x)$ will be a simple horizontal line!

   Any two points are ok for the second part (based on your experience with the SVR applet at `https://www.csie.ntu.edu.tw/~cjlin/libsvm/` or any other SVR implementation, or your understanding of SVR

   (**4 Marks**)

2. (a) justification using strong duality: solving the SVR dual is equivalent to solving the primal owing to strong duality. Strong duality holds because KKT conditions are necessary and sufficient conditions owing to convexity of objective and of constraints (which are just linear). (b) while one could consider optimizing the dual using coordinate ascent (one coordinate at a time), the dual optimization problem has a linear constraint in $\sum_i (\alpha_i^* - \alpha_i) = 0$. Thus, one simply cannot do coordinate ascent with one coordinate at a time. (c) block coordinate ascent: Since either $\alpha_i = 0$ or $\alpha_i^* = 0$ for any $i$, SMO minimally does block coordinate ascent in two coordinates at a time $\alpha_i$ and $\alpha_j$ or $\alpha_i^*$ and $\alpha_j^*$ while holding all other $\alpha_k$ and $\alpha_k^*$ values to be constants from the previous iterations.

   (**3 Marks**)

3. This was stated on slide 16 of `https://www.cse.iitb.ac.in/~cs725/notes/lecture-slides/lecture-09-unannotated.pdf` of the class notes but not proved. Wish some of you attempted it and tried to rationalize this statement!

We will prove by contradiction. Suppose for any $i$, $\widehat{\xi}_i < 0$ were the optimal solution, then

$$y - \mathbf{w}^T\phi(x_i) - b \leq \epsilon - \widehat{\xi}_i < \epsilon - 0$$

We claim that

(a) keeping all other values of $\xi_j$ (for $j \neq i$) and $\xi_j^*$ (all $j$ including $i$) constant while setting $\xi_i = 0$ continues satisfying all the constraints

(b) because $\xi_i^2 = 0 < \left(\widehat{\xi}_i\right)^2$ and all other variable values are the same, use of $\xi_i$ yields a lower value of the objective than does $\widehat{\xi}_i$.

This contradicts our assumption that for some $i$, $\widehat{\xi}_i < 0$ was the optimal solution. Thus, at optimal solution, we must have that for all $i$, $\widehat{\xi}_i = 0$

We can similarly prove that $\widehat{\xi_i^*} = 0$ for all $i$.

(**4 Marks**)

**Problem 2.** • Its difference with Ridge regression is that here $b$ is not captured within $w$, and $b$ is not minimized as part of minimizing $\|w\|^2$ as ridge regression did. Its difference with Support Vector Regression is that there is no explicit $\epsilon$ band over which the penalty is relaxed.

(**2 Marks**)

**Solution:**

- The objective function is convex in $w$ and $b$ as well as differentiable (you will need to show why). In the absence of constraints, setting the gradient of the objective to 0 and solving should give us the global minima.
- Thus, $\nabla_{w,b}L(w^*, b^*) = 0$ is a necessary and sufficient condition for optimality
- w.r.t $w$, we have:
  $w + C\sum_i(\phi(x_i)\phi^T(x_i))w + C\sum_i(y_i - b)\phi(x_i) = 0$
- w.r.t $b$, we have:
  $nb + \sum_i(\phi^\top(x_i)w - y_i) = 0$
- Unlike SVR formulation which had linear inequalities here we have only linear equalities, which can be solved
- Thus, we obtain the closed form solution:

$$w = (\Phi^\top\Phi + \frac{1}{C}\begin{bmatrix} 0 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix})^{-1}\Phi^\top y$$

- LS-SVM gives us a closed-form expression for $w$. Owing to convexity (as discussed before), this solution is the global minimum (optimum).
- **Difference wrt Ridge regression:** Overhead of estimating $b$!

$$b = \frac{-1}{n} \sum_k (\phi^\top(x_k)w - y_k) = 0$$

- **Difference wrt SVR:** The value of $b$ is computed as an average over multiple points. In the case of SVR, the value $b$ is computed using any one of some selected points (support vectors) only which can lead to numerical in stability
- As in the case of ridge regression, we use the following identity[1] $(P^{-1}+B^TR^{-1}B)^{-1}B^TR^{-1} = PB^T(BPB^T+R)^{-1}$. We can show that the decision function becomes $f(x) = \sum_{i=1}^m \alpha_i K(x_k, x_i) - \frac{1}{n}\sum_k ((\sum_{i=1}^m \alpha_i K(x, x_i)) - y_k)$
  where $\alpha_i = \left(([K(x_i, x_j)] + \frac{1}{C}I)^{-1}y\right)_i$

(**5 Marks**)

**Problem 3.** Show that the following kernel is positive semi-definite: $K(x_1, x_2) = (< x_1, x_2 > +c)^d$, where $< x_1, x_2 >$ is an inner product of vectors $x_1$ and $x_2$ and $d \in Z^+$.

**Prove that $(x_1^\top x_2)^d$ is a Mercer kernel ($d \in \mathbb{Z}^+$, $d \geq 1$)**

- We want to prove that
  $\int_{x_1}\int_{x_2}(x_1^\top x_2)^d g(x_1)g(x_2)\,dx_1dx_2 \geq 0$,
  for all square integrable functions $g(x)$
- Here, $x_1$ and $x_2$ are vectors s.t $x_1, x_2 \in \Re^t$
- Thus, $\int_{x_1}\int_{x_2}(x_1^\top x_2)^d g(x_1)g(x_2)\,dx_1dx_2$

$$= \int_{x_{11}} .. \int_{x_{1t}} \int_{x_{21}} .. \int_{x_{2t}} \left[ \sum_{n_1..n_t} \frac{d!}{n_1!..n_t!} \prod_{j=1}^{t} (x_{1j}x_{2j})^{n_j} \right] g(x_1)g(x_2)\,dx_{11}..dx_{1t}dx_{21}..dx_{2t}$$

$$\text{s.t. } \sum_{i=1}^{t} n_i = d$$

$$= \sum_{n_1 \ldots n_t} \frac{d!}{n_1! \ldots n_t!} \int_{x_1}\int_{x_2} \prod_{j=1}^{t} (x_{1j}x_{2j})^{n_j}\, g(x_1)g(x_2)\,dx_1dx_2$$

$$= \sum_{n_1 \ldots n_t} \frac{d!}{n_1! \ldots n_t!} \int_{x_1}\int_{x_2} (x_{11}^{n_1}x_{12}^{n_2}\ldots x_{1t}^{n_t})g(x_1)\,(x_{21}^{n_1}x_{22}^{n_2}\ldots x_{2t}^{n_t})g(x_2)\,dx_1dx_2$$

[1] Recall we used it for Ridge regression on slide 12 of `https://www.cse.iitb.ac.in/~cs725/notes/lecture-slides/lecture-12-unannotated.pdf`

$$= \sum_{n_1 \ldots n_t} \frac{d!}{n_1! \ldots n_t!} \left( \int_{x_1} (x_{11}^{n_1} \ldots x_{1t}^{n_t}) g(x_1)\, dx_1 \right) \left( \int_{x_2} (x_{21}^{n_1} \ldots x_{2t}^{n_t}) g(x_2)\, dx_2 \right)$$

$$\underline{\text{(integral of decomposable product as product of integrals)}}$$

$$\text{s.t.} \sum_{i}^{t} n_i = d$$

- Realize that both the integrals are basically the same, with different variable names
- Thus, the equation becomes:

$$\sum_{n_1 \ldots n_t} \frac{d!}{n_1! \ldots n_t!} \left( \int_{x_1} (x_{11}^{n_1} \ldots x_{1t}^{n_t}) g(x_1)\, dx_1 \right)^2 \geq 0$$

  $\underline{\text{(the square is non-negative for reals)}}$

- Thus, we have shown that $(x_1^\top x_2)^d$ is a Mercer kernel.

**What about $\sum_{d=1}^{r} \alpha_d (x_1^\top x_2)^d$ s.t. $\alpha_d \geq 0$?**

- $K(x_1, x_2) = \sum_{d=1}^{r} \alpha_d (x_1^\top x_2)^d$
- Is $\int_{x_1} \int_{x_2} \left( \sum_{d=1}^{r} \alpha_d (x_1^\top x_2)^d \right) g(x_1) g(x_2)\, dx_1 dx_2 \geq 0$?
- We have

$$\int_{x_1} \int_{x_2} \left( \sum_{d=1}^{r} \alpha_d (x_1^\top x_2)^d \right) g(x_1) g(x_2)\, dx_1 dx_2$$

$$= \sum_{d=1}^{r} \alpha_d \int_{x_1} \int_{x_2} (x_1^\top x_2)^d g(x_1) g(x_2)\, dx_1 dx_2$$

- We have already proved that $\int_{x_1} \int_{x_2} (x_1^\top x_2)^d g(x_1) g(x_2)\, dx_1 dx_2 \geq 0$
- Also, $\alpha_d \geq 0$, $\forall d$
- Thus,

$$\sum_{d=1}^{r} \alpha_d \int_{x_1} \int_{x_2} (x_1^\top x_2)^d g(x_1) g(x_2)\, dx_1 dx_2 \geq 0$$

- By which, $K(x_1, x_2) = \sum_{d=1}^{r} \alpha_d (x_1^\top x_2)^d$ is a Mercer kernel.

(**5 Marks**)

**Problem 4.** This problem is directly related to problem 5 of tutorial 3 in which the weighing factor $r_i^{x'}$ of each training data point $(\mathbf{x}_i, y_i)$ is now also a function of the query or test data point $(\mathbf{x}', ?)$, so that we write it as $r_i^{x'} = K(\mathbf{x}', \mathbf{x}_i)$ for $i = 1, \ldots, m$. Let $r_{m+1}^{x'} = 1$ and let $R$ be an $(m+1) \times (m+1)$ diagonal matrix of $r_1^{x'}, r_2^{x'}, \ldots, r_{m+1}^{x'}$.

$$R = \begin{bmatrix} r_1^{x'} & 0 & ... & 0 & \\ 0 & r_2^{x'} & ... & 0 & \\ ... & ... & ... & ... & 1 \\ 0 & 0 & 0 & ... & r_{m+1}^{x'} \end{bmatrix}$$

Further, let

$$\Phi = \begin{bmatrix} \phi_1(x_1) & ... & \phi_p(x_1) & 1 \\ ... & ... & ... & 1 \\ \phi_1(x_m) & ... & \phi_p(x_m) & 1 \end{bmatrix}$$

and

$$\widehat{\mathbf{w}} = \begin{bmatrix} w_1 \\ ... \\ w_p \\ b \end{bmatrix}$$

and

$$\mathbf{y} = \begin{bmatrix} y_1 \\ ... \\ y_m \end{bmatrix}$$

The sum-square error function then becomes

$$\frac{1}{2} \sum_{i=1}^{m} r_i (y_i - (\widehat{\mathbf{w}}^T \phi(x_i) + b))^2 = \frac{1}{2} ||\sqrt{R}\mathbf{y} - \sqrt{R}\Phi\widehat{\mathbf{w}}||_2^2$$

where $\sqrt{R}$ is a diagonal matrix such that each diagonal element of $\sqrt{R}$ is the square root of the corresponding element of $R$. This is a convex function being minimized (prove this using techniques similar to what we employed for least squares linear regression) and therefore has a global minimum at $\widehat{\mathbf{w}}_*^{x'}$ where the gradient must become 0. (again work out the steps using techniques similar to what we employed for least squares linear regression). The expression for the solution $\widehat{\mathbf{w}}^*$ that minimizes this error function is therefore

$$\widehat{\mathbf{w}}_*^{x'} = (\Phi^T R \Phi)^{-1} \Phi^T R \mathbf{y}$$

(**4 Marks**)

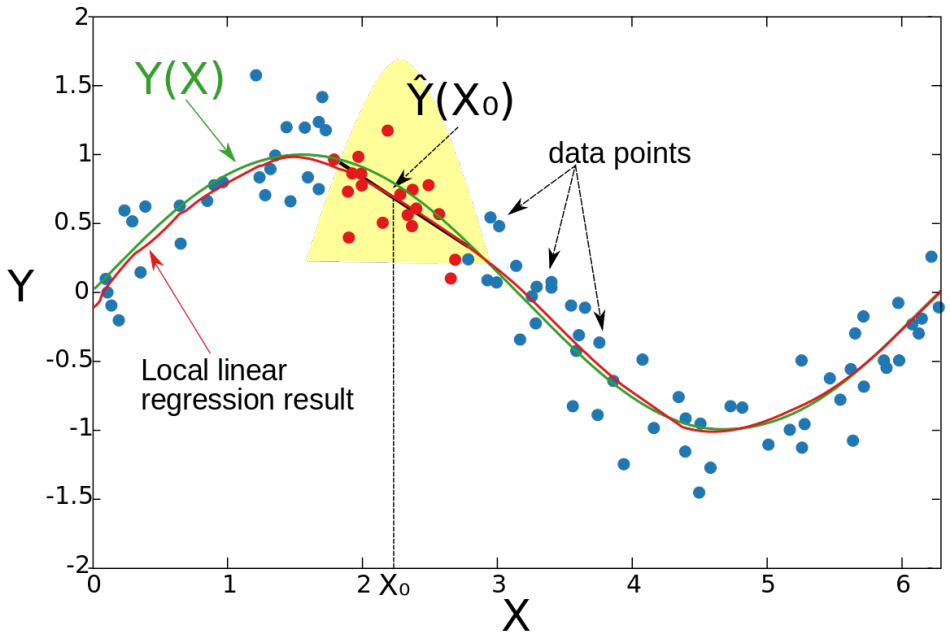
Let us refer to this model as local linear regression (Section 6.1.1 of Tibshi's book).

As compared to linear regression, local linear regression gives more importance to points in $\mathcal{D}$ that are closer/similar to $\mathbf{x}'$ and less importance to points that are less similar. Thus,

this method can be important if the regression curve is supposed to take different shapes or different parameters in different parts of the space. For example, in two different regions, the ideal regression curve might be linear in each but with different parameters. In this sense, local linear regression comes close to k-nearest neighbor. But unlike k-nearest neighbor, local linear regression gives you a smooth solution since contribution for regression at a point comes from all data points (in proportion to their closeness) and not just the k closest points.

(**1.5 Marks**)

Taking clue from the discussion above, one can try and plot this regression curve.




(**1.5 Marks**)