CS725/CS403 Midsem

Closed notes, 30 Marks, 2 hours

Tuesday 1st March, 2016

Problem 1. Support Vector Regression:

1. If all training data points lie strictly inside the ϵ-band of the SVR solution, what would the regression line be? You can build upon basic understanding of Support Vector Regression and the KKT conditions for Support Vector Regression which are reproduced below for your convenience:

   - Differentiating the Lagrangian w.r.t. w,
     $w - \alpha_i \phi(x_i) + \alpha_i^* \phi(x_i) = 0$
     i.e. $w = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \phi(x_i)$
   - Differentiating the Lagrangian w.r.t. ξ_i,
     $C - \alpha_i - \mu_i = 0$
     i.e. $\alpha_i + \mu_i = C$
   - Differentiating the Lagrangian w.r.t ξ_i^*,
     $\alpha_i^* + \mu_i^* = C$
   - Differentiating the Lagrangian w.r.t b,
     $\sum_i (\alpha_i^* - \alpha_i) = 0$
   - Complimentary slackness:
     $\alpha_i (y_i - w^\top \phi(x_i) - b - \epsilon - \xi_i) = 0$
     $\mu_i \xi_i = 0$
     $\alpha_i^* (b + w^\top \phi(x_i) - y_i - \epsilon - \xi_i^*) = 0$
     $\mu_i^* \xi_i^* = 0$

   If you have happened to play around with the SVR applet at `https://www.csie.ntu.edu.tw/~cjlin/libsvm/` (or any other SVR implementation), state any other 2 interesting observations about SVR that you remember. You could also state these observations based on your understanding and expectation of the behavior of SVR.

   (**4 Marks**)

2. What are the chief ideas behind the Sequential Minimal Optimization (SMO) Algorithm for Support Vector Regression? You need NOT write the detailed steps of SMO. However, your answer must include the following points: (a) justification using strong duality (b) linear constraint in dual (c) block coordinate ascent

   (**3 Marks**)

3. Suppose in the SVR formulation, we replace the error component of the objective

$$C\sum_{i=1}^{n}(\xi_i+\xi_i^*)$$

(s.t. $\xi_i \geq 0$, $\xi_i^* \geq 0$) with

$$C\sum_{i=1}^{n}\left((\xi_i)^2+(\xi_i^*)^2\right)$$

That is, we replaced the L_1 norm error with L_2 (or squared) norm error. Explain why you would no longer need the set of constraints $\xi_i \geq 0$, $\xi_i^* \geq 0$ in the optimization problem with L_2 norm error.

(**4 Marks**)

Problem 2. Consider the following regression formulation:

$$\min_{\mathbf{w},b}\frac{1}{2}\|\mathbf{w}\|^2+\frac{C}{2}\sum_{i=1}^{n}\left(y_i-(\mathbf{w}^\top\phi(x_i)+b)\right)^2 \tag{1}$$

- Contrast this formulation against the formulations of ridge regression and support vector regression discussed in class.

  (**2 Marks**)

- Does this formulation have a closed form **optimal** solution? Prove. Now is there an equivalent completely kernelized expression for the function $f(x)=\mathbf{w}^T\phi(x)+b$ in terms of the solution to the original problem in (1)?

  Contrast this solution with that of Ridge regression and the standard support vector regression.

  (**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 $d\in Z^+$.
(**5 Marks**)

Problem 4. In tutorial 3 (problem 5), we had discussed weighted regression. In this problem, we will deal with weighted regression, with the weights obtained using some kernel $K(.,.)$. Given a training set of points $\mathcal{D}=\{(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_i,y_i),\ldots,(\mathbf{x}_n,y_n)\}$, we predict a regression function $f(x')=(\mathbf{w}^\top\phi(x')+b)$ for each test (or query point) x' as follows:

$$(\mathbf{w}',b')=\operatorname*{argmin}_{\mathbf{w},b}\sum_{i=1}^{n}K(x',x_i)\left(y_i-(\mathbf{w}^\top\phi(x_i)+b)\right)^2$$

1. If there is a closed form expression for $(\mathbf{w}', b')$ and therefore for $f(x')$ in terms of the known quantities, derive it.

   (**4 Marks**)

2. How does this model compare with linear regression and $k-$nearest neighbor regression? What are the relative advantages and disadvantages of this model?

   (**1.5 Marks**)

3. In the one dimensional case (that is when $\phi(x) \in \Re$), graphically try and interpret what this regression model would look like, say when $K(.,.)$ is the linear kernel[1].

   (**1.5 Marks**)

[1]Hint: What would the regression function look like at each training data point?