CS725 Midsem

Closed notes, 30 Marks, 2 hours

Wednesday 7th September, 2016

Please answer **to the point** in the limited space provided for each question. You can do rough work in a separate sheet of paper provided to you. You can also assume any result stated or proved in the class (but NOT as part of the tutorials).

Problem 1. Relation between Penalized Ridge Regression (λ) and Constrained Ridge Regression (θ):

Show that the solution to the *Penalized Ridge Regression* problem

$$\mathbf{w}_{Pen} = \underset{\mathbf{w}}{\text{argmin}} \; ||\phi\mathbf{w} - \mathbf{y}||_2^2 + \lambda||\mathbf{w}||_2^2$$

is the same as that to the solution to the *Constrained Ridge Regression* problem

$$\mathbf{w}_{Con} = \underset{\mathbf{w}}{\text{argmin}} \; ||\phi\mathbf{w} - \mathbf{y}||_2^2$$
$$\textit{such that } ||\mathbf{w}||_2^2 \leq \xi$$

for some ξ that is a function of λ.

Hint1: This claim is the converse of the claim made in Tutorial 5, Problem 1. Recall that converse of $A \rightarrow B$ is $B \rightarrow A$.

Hint2: You can make convexity assumptions and use KKT conditions if required.

(**7 Marks**)

Problem 2. Consider the Lasso problem:

$$\mathbf{w}^* = \underset{\mathbf{w}}{\mathbf{argmin}} \, \|\phi\mathbf{w} - \mathbf{y}\|^{\mathbf{2}} \;\; \mathbf{s.t.} \;\; \|\mathbf{w}\|_{\mathbf{1}} \leq \eta, \tag{1}$$

where

$$\|\mathbf{w}\|_1 = \Big(\sum_{i=1}^{n} |w_i|\Big) \tag{2}$$

1. Since $\|\mathbf{w}\|_1$ is not differentiable, try reformulating this problem to another equivalent formulation so as to make functions in the constraint differentiable so that you are able to derive KKT conditions in the next step.

2. Now derive the Karush Kuhn Tucker conditions at primal variable $\mathbf{w}^*$ and the dual lagrange variables (which you will introduce). Are these conditions necessary/sufficient conditions for optimality?

   (**5 Marks**)

Problem 3. Posterior Distribution of w with very imprecise prior:

Let $y = \mathbf{w}^T\phi(\mathbf{x}) + \varepsilon$ and let dataset $\mathcal{D} = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_i, Y_i), \ldots, (\mathbf{X}_m, Y_m)\}$ was provided. Recall that the posterior distribution for $\mathbf{w}$ under a Gaussian prior was $\Pr(\mathbf{w} \mid \mathcal{D}) = \mathcal{N}(\mathbf{w} \mid \mu_m, \Sigma_m)$ where

$$\Sigma_m^{-1} = \lambda I + \Phi^T\Phi/\sigma^2$$

and

$$\mu_m = (\lambda\sigma^2 I + \Phi^T\Phi)^{-1}\Phi^T\mathbf{y}$$

How would you model a very imprecise Gaussian prior on $\Pr(\mathbf{w})$? Explain what happens to the parameters of the posterior $\Pr(\mathbf{w} \mid \mathcal{D})$ as this precision on the prior $\Pr(\mathbf{w})$ tends to 0. What is the connection between this expression and the data likelihood expression?

(**3 Marks**)

Problem 4. Interpreting the Primal and Dual Variables Solution to the Support Vector Regression Formulation:

Recall the Lagrange Function for the Support Vector Regression Problem:

$$L(\mathbf{w}, \alpha, \alpha^*, \mu, \mu^*) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i + \xi_i^*) + \sum_{i=1}^m \alpha_i\left(y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b - \epsilon - \xi_i\right)$$
$$+ \sum_{i=1}^m \alpha_i^*\left(b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*\right) - \sum_{i=1}^m \mu_i\xi_i - \sum_{i=1}^m \mu_i^*\xi_i^*$$

And the following KKT conditions for this SVR formulation:

$$\mathbf{w} - \sum_{i=1}^m(\alpha_i\phi(\mathbf{x}_i) - \alpha_i^*\phi(\mathbf{x}_i)) = 0 \ i.e., \ \mathbf{w} = \sum_{i=1}^m(\alpha_i - \alpha_i^*)\phi(\mathbf{x}_i)$$
$$C - \alpha_i - \mu_i = 0 \ i.e., \ \alpha_i + \mu_i = C, \ \alpha_i^* + \mu_i^* = C, \ \textstyle\sum_i(\alpha_i^* - \alpha_i) = 0$$
$$\alpha_i(y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b - \epsilon - \xi_i) = 0 \text{ AND } \mu_i\xi_i = 0 \text{ AND } \alpha_i^*(b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*) = 0$$
$$\text{AND } \mu_i^*\xi_i^* = 0$$

In the optimal weight vector $\mathbf{w} = \sum_{i=1}^m(\alpha_i - \alpha_i^*)\phi(\mathbf{x}_i)$, determine which types of points will contribute to $\mathbf{w}$ through a non-zero value of $(\alpha_i - \alpha_i^*)$. You can structure your answer along the following lines.

1. First prove that for any point $(\mathbf{x}_i, y_i)$, the product $\alpha_i\alpha_i^* = 0$.

   (**2 Marks**)

Why is this important for what we are trying to prove?

(**2 Marks**)

2. What will be the value of α and α^* for points that lie strictly outside the ϵ-insensitive tube? Justify your answer.

   (**2 Marks**)

3. What will be the value of α and α^* for points that lie strictly on the boundary of the ϵ-insensitive tube? And how about points lying within the ϵ-insensitive tube? Justify your answer.

   (**3 Marks**)

4. If all training data points lie strictly inside the ϵ-band of the SVR solution, what would the regression line be?

   (**3 Marks**)

Problem 5. Valid or Positive Definite Kernels:

We proved in class that the following kernel is valid or positive definite: $K(\mathbf{x}_1, \mathbf{x}_2) = (\langle \mathbf{x}_1, \mathbf{x}_2 \rangle)^d = \left(\sum_{i=1}^{n} x_{1i} x_{2i} \right)^d$, where $(\langle \mathbf{x}_1, \mathbf{x}_2 \rangle) = \sum_{i=1}^{n} x_{1i} x_{2i}$ is an inner product of vectors $\mathbf{x}_1, \mathbf{x}_2 \in \Re^n$ and $d \in Z^+$.

If required, assuming the above, prove that

$$K_{new}(\mathbf{x}_1, \mathbf{x}_2) = \left(\sum_{i=1}^{n} \sqrt{x_{1i}} \sqrt{x_{2i}} \right)^d$$

is also a positive definite kernel.

Hint: If needed, you can prove and then use the following more general claim:

If $K(\mathbf{x}_1, \mathbf{x}_2)$ is a positive definite kernel and $g(\mathbf{x}) : \Re^n \rightarrow \Re^n$ then $K_{new}(\mathbf{x}_1, \mathbf{x}_2) = K\left(g(\mathbf{x}_1), g(\mathbf{x}_2)\right)$ is also a positive definite kernel.

(**3 Marks**)