CS725 Midsem

Closed notes, 30 Marks, 2 hours

Tuesday 20th 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

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

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

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

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 $A \to B$ is $B \to A$.

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

(**7 Marks**)

Solution Sketch:

This is the exact converse of the claim made in Tutorial 5, where we had to prove that the solution to the *Constrained Ridge Regression* problem was the same as that to the solution to *Penalized Ridge Regression* for some λ that is a function of ξ.

- Consider the *Penalized Ridge Regression* formulation

  $$\min(\| \Phi\mathbf{w} - \mathbf{y} \|^{\mathbf{2}} + \lambda_{\mathbf{Pen}} \| \mathbf{w} \|^{\mathbf{2}})$$

  setting gradient to $\mathbf{0}$ we get the solution

  $$\nabla_{\mathbf{w_{Pen}}}(f(\mathbf{w}) + \lambda_{\mathbf{Pen}}\mathbf{g}(\mathbf{w})) = \mathbf{0}$$

  Here, $f(\mathbf{w}) = (\Phi\mathbf{w} - \mathbf{y})^T(\Phi\mathbf{w} - \mathbf{y})$ and, $g(\mathbf{w}) = \|\mathbf{w}\|^2$.

- Solving we get,

$$\mathbf{w}_{Pen} = (\Phi^T\Phi + \lambda_{Pen}I)^{-1}\Phi^T\mathbf{y}$$

- Now[1] consider the *Constrained Ridge Regression* formulation in which we limit the weights of the coefficients by placing an upper bound $\xi = ||(\Phi^T\Phi + \lambda_{Pen}I)^{-1}\Phi^T\mathbf{y}||_\mathbf{2}^\mathbf{2}$, on size of the L2 norm of the weight vector, with λ set from the *Penalized Ridge Regression* formulation:

$$\text{argmin}_\mathbf{w}(\mathbf{\Phi w} - \mathbf{y})^T(\mathbf{\Phi w} - \mathbf{y})$$
$$\|\mathbf{w}\|_\mathbf{2}^\mathbf{2} \leq \xi$$

- As discussed in tutorial 5, the objective function, namely $f(\mathbf{w}) = (\mathbf{\Phi w} - \mathbf{y})^\mathbf{T}(\mathbf{\Phi w} - \mathbf{y})$ is strictly convex. The constraint function, $g(\mathbf{w}) = \|\mathbf{w}\|_\mathbf{2}^\mathbf{2} - \xi$, is also convex.

- To minimize the error function subject to constraint $|\mathbf{w}| \leq \xi$, we apply KKT conditions at the point of optimality $\mathbf{w_{Con}}$

$$\nabla_{\mathbf{w_{Con}}}(f(\mathbf{w}) + \hat{\lambda}\mathbf{g}(\mathbf{w})) = \mathbf{0}$$

(the first KKT condition). Here, $f(\mathbf{w}) = (\Phi\mathbf{w} - \mathbf{y})^T(\Phi\mathbf{w} - \mathbf{y})$ and, $g(\mathbf{w}) = \|\mathbf{w}\|^2 - \xi$.

- Solving we get,

$$\mathbf{w}_{Con} = (\Phi^T\Phi + \hat{\lambda}I)^{-1}\Phi^T\mathbf{y}$$

From the second KKT condition we get,

$$\|\mathbf{w}_{Con}\|^2 \leq \xi$$

From the third KKT condition,

$$\hat{\lambda} \geq 0$$

From the fourth condition

$$\hat{\lambda}\|\mathbf{w_{Con}}\|^\mathbf{2} = \hat{\lambda}\xi$$

- Values of $\mathbf{w_{Con}}$ and $\hat{\lambda}$ that satisfy all these equations would yield an optimal solution. That is, if

$$\|\mathbf{w}_{Con}\| = \|(\Phi^T\Phi)^{-1}\Phi^T\mathbf{y}\| \leq \xi$$

then $\hat{\lambda} = 0$ is the solution. Else, for some sufficiently large value, $\hat{\lambda}$ will be the solution to

$$\|\mathbf{w}_{Con}\| = \|(\Phi^T\Phi + \hat{\lambda}I)^{-1}\Phi^T\mathbf{y}\| = \xi$$

- **Indeed, the value of $\hat{\lambda} = \lambda_{Pen}$ is the solution to this above equation since ξ itself was chosen such that**

$$\xi = ||(\Phi^T\Phi + \lambda_{Pen}I)^{-1}\Phi^T\mathbf{y}||_\mathbf{2}^\mathbf{2}$$

[1] The discussion hereafter is exactly similar to the discussion of solution to tutorial 5, problem 1. But we have no need to discuss the Bound on λ in the regularized least square solution:

Problem 2. Consider the Lasso problem:

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

where

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

1. Since $\left\|\mathbf{w}\right\|_1$ 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**)

Solution:

- $$\mathbf{w}^* = \underset{\mathbf{w}}{\mathbf{argmin}} \left\|\phi\mathbf{w} - \mathbf{y}\right\|^{\mathbf{2}} \ \mathbf{s.t.}\ \left\|\mathbf{w}\right\|_{\mathbf{1}} \leq \eta, \tag{3}$$

  where

  $$\left\|\mathbf{w}\right\|_1 = \Big(\sum_{i=1}^{n} |w_i|\Big) \tag{4}$$

- Since $\left\|\mathbf{w}\right\|_1$ is not differentiable, one can express (4) as a set of constraints

  $$\sum_{i=1}^{n} \xi_i \leq \eta,\ w_i \leq \xi_i,\ -w_i \leq \xi_i$$

- The resulting problem is a linearly constrained Quadratic optimization problem (LCQP):

  $$\mathbf{w}^* = \underset{\mathbf{w}}{\mathbf{argmin}} \left\|\phi\mathbf{w} - \mathbf{y}\right\|^{\mathbf{2}} \ \mathbf{s.t.}\ \sum_{\mathbf{i=1}}^{\mathbf{n}} \xi_{\mathbf{i}} \leq \eta,\ \mathbf{w_i} \leq \xi_{\mathbf{i}},\ -\mathbf{w_i} \leq \xi_{\mathbf{i}} \tag{5}$$

- Lagrangian is

  $$\left\|\phi\mathbf{w} - \mathbf{y}\right\|^2 \ + \beta(\sum_{i=1}^{n} \xi_i - \eta) + \sum_{i=1}^{n} (\theta_i(\ w_i - \xi_i) + \lambda_i(-w_i - \xi_i))$$

- KKT conditions: Setting gradient wrt $\mathbf{w}$ to $\mathbf{0}$:

$$2(\phi^T\phi)\mathbf{w} - \mathbf{2\phi^T y} + (\theta - \lambda) = \mathbf{0}$$

  Setting gradient wrt ξ_i to 0:

$$\beta - \theta_i - \lambda_i = 0$$

$$\beta(\sum_{i=1}^{n} \xi_i - \eta) = 0$$

$$\forall\ i,\ \theta_i(\mathbf{w_i} - \mathbf{\xi_i}) = \mathbf{0 \text{ and } \lambda_i(-w_i - \xi_i) = 0}$$

- We have also shown the equivalence of Lasso formulations in (4) and (6):

$$\mathbf{w}^* = \underset{\mathbf{w}}{\mathbf{argmin}} \left\|\phi\mathbf{w} - \mathbf{y}\right\|^{\mathbf{2}} + \lambda \left\|\mathbf{w}\right\|_{\mathbf{1}} \tag{6}$$

Problem 3. Posterior Distribution of $\mathbf{w}$ 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**)

Solution:

The key is to realize (from discussions in the class) that corresponding to the posterior distribution $\Pr(\mathbf{w} \mid \mathcal{D}) = \mathcal{N}(\mathbf{w} \mid \mu_m, \Sigma_m)$ **was the prior** $\Pr(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid 0, \frac{1}{\lambda}I)$. We discussed how λ (reciprocal of variance) corresponds to precision of the belief of 0 mean for each of the individual w_i's and therefore actually reflects precision of the prior. As $\lambda \to 0$, the prior will tend to have 0 precision or ∞ spread (variance), meaning that the prior is very imprecise. As $\lambda \to 0$ $\Pr(\mathbf{w} \mid \mathcal{D}) \to \mathcal{N}(\mathbf{w} \mid \mu_m^0, \Sigma_m^0)$ where

$$(\Sigma_m^0)^{-1} = \Phi^T\Phi/\sigma^2$$

and

$$\mu_m^0 = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{y}$$

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

   Solution: By design, we know that if $\xi_i > 0$ then $\xi_i^* = 0$ and vice versa.

   Let $\alpha_i > 0$ and $\alpha_i^* > 0$. We will show that this leads to a contradiction. First of all, by virtue of Complimentary slackness, this would mean $y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b - \epsilon - \xi_i = 0$ AND $b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^* = 0$. Adding up the two equalities gives us: $\xi_i + \xi_i^* = -2\epsilon$. Since only one of ξ_i and ξ_i^* can be non-zero, this implies that the non-zero component is negative, which is a contradiction since $\xi_i, \xi_i^* \geq 0$. Thus, atleast one of α_i and α_i^* must be 0.

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

   (**2 Marks**)

   Solution:

   The significance of $\alpha_i\alpha_i^* = 0$ is that without this equality, we could have had $\alpha_i > 0$ and $\alpha_i^* > 0$ yet canceling out each other's effects when $\alpha_i = \alpha_i^*$. But with $\alpha_i\alpha_i^* = 0$, we can be assured that $\alpha_i - \alpha_i^* \neq 0$ when one of the two α's are non-zero. That is, $\alpha_i - \alpha_i^* = \max\{\alpha_i, \alpha_i^*\}$

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**)

   Solution:

   If $\alpha_i = C$, then $\mu_i = 0$ and $y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b - \epsilon = \xi_i \geq 0$. That is, $\alpha_i = C$ corresponds to points lying above (or beyond) the upper $\epsilon-$band. Similarly, $\alpha_i^* = C$ corresponds to points lying below (or beyond) the lower $\epsilon-$band. Thus, $\alpha_i = C$ and $\alpha_i^* = C$ correspond to points lying either outside or on the $\epsilon-$tube.

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**)

   Solution:

   For any point on the upper margin, $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon = 0$ and $\xi_i = 0 \implies \mu_i \geq 0 \implies \alpha_i \in [0, C]$. That is, $\alpha_i \in [0, C]$ corresponds to points lying on the $\epsilon-$band above the regression curve. Similarly, $\alpha_i^* \in [0, C]$ corresponds to points lying on the $\epsilon-$band below the regression curve.

   Further, if $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i < 0$, then $\alpha_i = 0$, $\mu_i = C$ and $\xi_i = 0$ that is, $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon < 0$ which means that for points within the ϵ band, $\xi_i = 0$ and $\alpha_i = 0$. Similarly, one can argue for $b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i - \epsilon < 0$ leading to $\alpha_i^* = 0$. That is, points lying within the $\epsilon-$tube do not contribute to the weight vector.

   Thus the overall implication is that in the weight vector, $\mathbf{w} = (\alpha_i - \alpha_i^*)\phi(\mathbf{x}_i)$, the contribution $\alpha_i - \alpha_i^*$ is non-zero only for points lying on or within the $\epsilon-$tube.

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

   (**3 Marks**)

   Solution:

   If all training data points lie strictly inside the ϵ-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!

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**)

Solution:

Since we know that $K(\mathbf{x}_1, \mathbf{x}_2)$ is a positive definite kernel, there must exist $\phi : \Re^n \rightarrow \mathcal{H}$ such that $K(\mathbf{x}_1, \mathbf{x}_2) = (\langle \phi(\mathbf{x}_1), \phi(\mathbf{x}_2) \rangle)$

We are given that $g(\mathbf{x}) : \Re^n \rightarrow \Re^n$. Consider $\phi_g : \Re^n \rightarrow \mathcal{H}$ such that $\phi_g(\mathbf{x}) = \phi(g(\mathbf{x}))$. Then, $K_{new}(\mathbf{x}_1, \mathbf{x}_2) = K\left(g(\mathbf{x}_1), g(\mathbf{x}_2)\right) = (\langle \phi_g(\mathbf{x}_1), \phi_g(\mathbf{x}_2) \rangle)$. That is, $K_{new}(\mathbf{x}_1, \mathbf{x}_2)$ is a positive definite kernel.