Tutorials 3 and 4

Tuesday 7th February, 2017

Problem 1. Equivalence between Ridge Regression and Bayesian Linear Regression (with fixed σ^2 and λ):

Consider the Bayesian Linear Regression Model

$$y = \mathbf{w}^T\phi(\mathbf{x}) + \varepsilon \textit{ and } \varepsilon \sim \mathcal{N}(0, \sigma^2)$$
$$\mathbf{w} \sim \mathcal{N}(0, \alpha I) \textit{ and } \mathbf{w} \mid \mathcal{D} \sim \mathcal{N}(\mu_m, \Sigma_m)$$
$$\mu_m = (\lambda\sigma^2 I + \phi^T\phi)^{-1}\phi^T\mathbf{y} \textit{ and } \Sigma_m^{-1} = \lambda I + \phi^T\phi/\sigma^2$$

Show that $\mathbf{w}_{MAP} = \underset{\mathbf{w}}{\text{argmax}} \ \Pr(\mathbf{w} \mid \mathcal{D})$ is the same as that of *Regularized Ridge Regression*.

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

In other words, The Bayes and MAP estimates for Linear Regression coincide with that of *Regularized Ridge Regression*.

Solution Sketch: Taking the negative log of the log likelihood we see that maximizing the log of the posterior distribution is equivalent to minimizing the ridge regression objective.

$$\Pr(\mathbf{w} \mid \mathcal{D}) = \mathcal{N}(\mathbf{w} \mid \mu_m, \Sigma_m) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_m|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{w}-\mu_m)^T\Sigma_m^{-1}(\mathbf{w}-\mu_m)}$$

$$-\log\Pr(\mathbf{w}) = \cancel{\frac{n}{2}\log(2\pi)} + \cancel{\frac{1}{2}\log|\Sigma_m|} + \frac{1}{2}(\mathbf{w}-\mu_m)^T\Sigma_m^{-1}(\mathbf{w}-\mu_m)$$

$\cancel{\mu_m^T \mu_m}$

$$\mathbf{w}_{MAP} = \underset{\mathbf{w}}{\text{argmax}} - \log\Pr(\mathbf{w}|\mathcal{D}) = \underset{\mathbf{w}}{\text{argmax}}\frac{1}{2}\mathbf{w}^T\Sigma_m^{-1}\mathbf{w} - \mathbf{w}^T\Sigma_m^{-1}\mu_m$$

$w^T\phi^T\phi w - 2w^T\phi^T y$

All this manipulation (adding or canceling redundant terms) OK for argmax BUT NOT for max

that is,

$$\mathbf{w}_{MAP} = \underset{\mathbf{w}}{\text{argmax}}\frac{1}{2}\mathbf{w}^T\left(\lambda I + \phi^T\phi/\sigma^2\right)\mathbf{w} - \mathbf{w}^T\left(\lambda I \cancel{+ \phi^T\phi/\sigma^2)\left((\lambda\sigma^2 I} \cancel{+ \phi^T\phi)^{-1}}\phi^T\mathbf{y}\right)$$

σ^2

$||\phi w - y||^2 + \lambda||w||^2$

and after expanding and canceling out redundant terms, and later, after completing squares:

$$\mathbf{w}_{MAP} = \underset{\mathbf{w}}{\text{argmax}}\frac{1}{2\sigma^2}\mathbf{w}^T\left(\phi^T\phi\mathbf{w} - 2\phi^T\mathbf{y}\right) + \lambda\mathbf{w}^T\mathbf{w} = \underset{\mathbf{w}}{\text{argmax}}\frac{1}{2}||\phi\mathbf{w} - \mathbf{y}||^2 + \sigma^2\lambda||\mathbf{w}||^2 = \mathbf{w}_{Ridge}$$

$y^T y$ added without loss

Problem 2. Ridge Regression and Error Minimization:

1. *Prove the following Claim:* LS

   The sum of squares error on training data using the weights obtained after minimizing ridge regression objective is greater than or equal to the sum of squares error on training data using the weights obtained after minimizing the ordinary least squares (OLS) objective.

   More specifically, if ϕ and $\mathbf{y}$ are defined on the training set $\mathcal{D} = \{(\mathbf{x}_1, y_1) \ldots (\mathbf{x}_m, y_m)\}$ as

$$\phi = \begin{bmatrix} \phi_1(\mathbf{x}_1) & \phi_2(\mathbf{x}_1) & \ldots\ldots & \phi_n(\mathbf{x}_1) \\ . & & & \\ . & & & \\ \phi_1(\mathbf{x}_m) & \phi_2(\mathbf{x}_m) & \ldots\ldots & \phi_n(\mathbf{x}_m) \end{bmatrix} \tag{1}$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ . \\ y_m \end{bmatrix} \tag{2}$$

   and if

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

   and

$$\mathbf{w}_{OLS} = \underset{\mathbf{w}}{\text{argmin}}\ ||\phi\mathbf{w} - \mathbf{y}||_2^2$$

   $LS(w_{OLS}) \leq LS(w_{Ridge})$

   then you should prove that

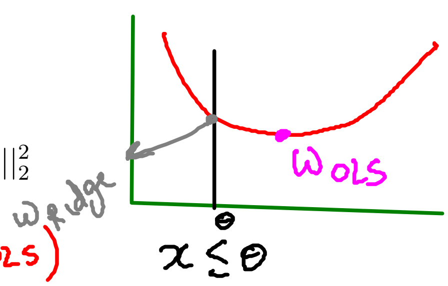
$$||\phi\mathbf{w}_{Ridge} - \mathbf{y}||_2^2 \geq ||\phi\mathbf{w}_{OLS} - \mathbf{y}||_2^2$$

   Solution: If

$$\mathbf{w}_{OLS} = \underset{\mathbf{w}}{\text{argmin}}\ ||\phi\mathbf{w} - \mathbf{y}||_2^2$$

   then by definition of argmin,

$$||\phi\mathbf{w}_{Ridge} - \mathbf{y}||_2^2 \geq ||\phi\mathbf{w}_{OLS} - \mathbf{y}||_2^2$$



   Also, one can reformulate $\forall w\ LS(w) \geq LS(w_{OLS})$

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

   as

$$\mathbf{w}_{Ridge} = \underset{\mathbf{w}}{\text{argmin}}\ ||\phi\mathbf{w} - \mathbf{y}||_2^2$$

$$\textit{such that } ||\mathbf{w}||_2^2 \leq \theta$$

   for some θ corresponding to a value of λ. The solution to a constrained minimization problem will always be greater than or equal to its unconstrained counterpart.

Alrc

2. If it is the case that ridge regression leads to greater error than ordinary least squares regression, then why should one be interested in ridge regression at all?

   Solution: This is still acceptable since ridge regression incorporates prior (as per Bayesian interpretation). The idea is ultimately to do well on unseen (test) data as well. Therefore, high training error might be acceptable if test error can be lowered.

Problem 3. Gradient descent is a very helpful algorithm. But it is not guaranteed to converge to global minima always. Give an example of a continuous function and initial point for which gradient descent converges to a value which is not global minima?

Problem 4. Step Length Considerations

$$Df = \begin{bmatrix} 2x_1 - 4 + 2x_2 \\ 2x_1 + 4x_2 + 2 \end{bmatrix} \xrightarrow{(4,-4)} \begin{bmatrix} -4 \\ -6 \end{bmatrix}$$

1. Consider the function

$$f(x) = x_1^2 - 4x_1 + 2x_1x_2 + 2x_2^2 + 2x_2 + 14$$

$t = \underset{t}{\operatorname{argmin}} f\left(\begin{bmatrix} 4 \\ -4 \end{bmatrix} - t\begin{bmatrix} -4 \\ -6 \end{bmatrix}\right) = \underset{t}{\operatorname{argmin}}\ at^2 + bt + c = \frac{-b}{2a}$

$x_1 = 4 + 4t \quad x_2 = -4 + 6t$

   This function has a minimum at $x = (5, 3)$. Suppose you are at a point $(4, -4)^T$ after few iterations, using the **exact line search algorithm** discussed in the class, find the point for the next iteration.

2. Now consider solving the Least Squares Linear Regression problem using the gradient descent algorithm. And let us say $w^{(0)} = 0$ and that the step length $t^{(k)}$ is computed using exact line search for each value of k. In how many steps will the gradient descent algorithm converge? What would be your answer if we had a different initialization for $w^{(0)}$?

   Solution:

$$t^{(k)} = \underset{t}{\operatorname{argmin}}\ \mathcal{E}\left(\underbrace{\mathbf{w^{(k)}}}_{w^{(0)}=0} + \mathbf{2t}\left(\phi^{\mathbf{T}}\mathbf{y} - \phi^{\mathbf{T}}\phi\underbrace{\mathbf{w^{(k)}}}_{=0} - \lambda\underbrace{\mathbf{w^{(k)}}}_{=0}\right)\right) \tag{3}$$

Soln:

$$t^{(0)} = \underset{t}{\operatorname{argmin}} \|\phi(2t\phi^T y) - y\|^2 + \lambda\|2t\phi^T y\|^2$$

$$= \underset{t}{\operatorname{argmin}}\ \underbrace{4t^2\|\phi\phi^T y\|^2 + \lambda 4t^2\|\phi^T y\|^2}_{a} - \underbrace{4ty^T\phi^T\phi y}_{b} + \cancel{\|y\|^2}$$

$$= \frac{-b}{2a}$$