Stat365/665 (Spring 2015) Data Mining and Machine Learning **Lecture: 1**

STATS 665 Homework 2

Lecturer: Leon Lixing Yu *Scribe: Leon Lixing Yu*

1 Problem 1

1.1 closed form of $\hat{\beta}$

$\hat{\beta}$ is given by:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2}\|X\beta - y\|_2^2 + \lambda\|\beta\|_2^2$$

In order to find *argmin*, we need to take the derivative of the loss function and make it equal to 0. In this assignment, I will use the notation, $x_j^{(i)}$, to denote that for the given data $(x^{(i)}, y^{(i)})_{i=1}^n$, $x_j^{(i)}$ is the jth element of $x^{(i)}$, where $x^{(i)} \in \mathbb{R}^d, y \in \mathbb{R}$, $X \in \mathbb{R}^{d\times n}, Y \in \mathbb{R}^n$. Therefore, we have:

$$\frac{d}{d\beta_j} l(\beta|X,Y) = (\sum_{i=1}^{n}((\sum_{j=1}^{d} x_j^{(i)}\beta_j) - y^{(i)})^2 + \lambda\sum_{j=1}^{d}\beta_j^2)' = 0$$

$$\Rightarrow 2X^T(X\beta - Y) + 2\lambda(I_d)\beta = 0$$

$$\Rightarrow \hat{\beta} = (X^TX + \lambda(I_d))^{-1}X^TY$$

1.2 Find a simple expression for $\|\hat{\beta} - \beta^*\|$

FIXME

1.3 Find a closed form of $\hat{f}$

Notation Note: Following the previous convention, I am using $(x^{(i)}, y^{(i)})_{i=1}^n$ instead of $(x_i, y_i)_{i=1}^n$ to represent the dataset. In this question, I use $f(x^{(i)})$ and $< f, \phi(x^{(i)}) >$ interchangably. We know that:

$$f \in \mathcal{H} \Rightarrow \quad f(x^{(i)}) \in \mathcal{H} \Rightarrow \quad < f, \phi(x^{(i)}) > \in \mathcal{H}$$

With the notation given, we have equation:

$$\hat{f} = \underset{f}{\operatorname{argmin}} \sum_{i=1}^{n}(y^{(i)} - < f, \phi(x^{(i)}) >_{\mathcal{H}})^2 + \lambda\|f\|_{\mathcal{H}}^2$$

Using representer theorm, we have:

$$f = \sum_{i=1}^{n}\alpha^{(i)}\phi(x^{(i)}) = \sum_{i=1}^{n}\alpha^{(i)}k(x^{(i)}, \cdot)$$

The equation above means f is a linear combination of feature space, mapping of points. substitute the relation above into the original $\hat{f}$ equation, we have:

$$\sum_{i=1}^{n}(y^{(i)} - < f, \phi(x^{(i)}) >_{\mathcal{H}})^2 + \lambda \|f\|_{\mathcal{H}}^2 = \|Y - K\boldsymbol{\alpha}\|^2 + \lambda \boldsymbol{\alpha}^T K \boldsymbol{\alpha}$$

Taking derivative over α and make it equal to 0 to get $\underset{\alpha}{\text{argmin}}$:

$$\frac{d}{d\alpha}(\|Y - K\boldsymbol{\alpha}\|^2 + \lambda \boldsymbol{\alpha}^T K \boldsymbol{\alpha}) = 0$$
$$\Rightarrow 2K(Y - K\boldsymbol{\alpha}) + 2\lambda \mathbf{I_d} K \boldsymbol{\alpha} = 0$$
$$\Rightarrow (K + \lambda \mathbf{I_d})\boldsymbol{\alpha} = 2Y$$
$$\Rightarrow \hat{\boldsymbol{\alpha}} = (K + \lambda \mathbf{I_d})^{-1} Y$$

recall:

$$f = \sum_{i=1}^{n} \alpha^{(i)} \phi(x^{(i)}) = \sum_{i=1}^{n} \alpha^{(i)} k(x^{(i)}, \cdot)$$

our $\hat{f}$ is then:

$$\hat{f} = K^T \hat{\boldsymbol{\alpha}} = K^T (K + \lambda \mathbf{I_d})^{-1} Y$$

In which K is the Kernel matrix W.R.T X