**CSE 404: Introduction to Machine Learning (Fall 2019)**

Homework #4

Note: (1) LFD refers to the textbook "Learning from Data".

1. (20 points) Recall the objective function for linear regression can be expressed as

$$E(w) = \frac{1}{N}\|Xw - y\|^2,$$

as in (3.3) of LFD. Minimizing this function with respect to $w$ leads to the optimal $w$ as $(X^TX)^{-1}X^Ty$. This solution holds only when $X^TX$ is nonsingular. To overcome this problem, the following objective function is commonly minimized instead:

$$E_2(w) = \|Xw - y\|^2 + \lambda\|w\|^2,$$

where $\lambda > 0$ is a user-specified parameter. Please do the following:

(a) (10 points) Derive the optimal $w$ that minimize $E_2(w)$.
**Solution:** $\nabla E_2(w) = 2X^T(Xw - y) + 2\lambda w = 0 \Rightarrow w = (X^TX + \lambda I)^{-1}X^Ty$

(b) (10 points) Explain how this new objective function can overcome the singularity problem of $X^TX$.
**Solution:**
$\forall v \neq 0, v^T(X^TX + \lambda I)v = v^T(X^TX)v + \lambda v^Tv.$
Since $v^T(X^TX)v \geq 0$ and $\lambda v^Tv > 0$, $X^TX + \lambda I$ is positive definite. Hence it is invertible.

2. (20 points) Exercise 3.3 in LFD (page 87).

**Solution:**

(a) $H^T = (X(X^TX)^{-1}X^T)^T = X((X^TX)^{-1})^TX^T = X(X^TX)^{-1}X^T = H$

(b) When K = 1, it holds.

Assume when K = n it holds. For K = n + 1, $H^{n+1} = H^nH = HH = H$

Hence $H^K = H$ holds for any positive integer K.

(c) When K = 1, it holds.

Assume when K = n it holds. For K = n + 1, $(I-H)^{n+1} = (I-H)^n(I-H) = (I-H)(I-H) = (I-H)$

Hence $(I-H)^K = (I-H)$ holds for any positive integer K.

(d) $trace(H) = trace(X(X^TX)^{-1}X^T) = trace(X^TX(X^TX)^{-1}) = trace(I_{(d+1)\times(d+1)}) = d+1$