CS4801: Principle of Machine Learning
Assignment 2

no deadline

This homework consists of problems covering regression. A few instructions to make life easier for all of us:

- Assignment need not to be submitted.
- No quiz for this assignment

Exercise 1: Regression

(a) (6 points) Derive the solution for following optimization problem

i. Least Square Regression solves

$$\min_{\mathbf{w}} \frac{1}{2}(\mathbf{X}\mathbf{w}-\mathbf{y})^T(\mathbf{X}\mathbf{w}-\mathbf{y}).$$

Show that the solution is

$$\mathbf{w}^{ls} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

ii. Ridge Regression solves

$$\min_{\mathbf{w}} \frac{1}{2}(\mathbf{X}\mathbf{w}-\mathbf{y})^T(\mathbf{X}\mathbf{w}-\mathbf{y}) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2.$$

Show that the solution is

$$\mathbf{w}^{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

iii. Lasso

$$\min_{\mathbf{w}} \frac{1}{2}(\mathbf{X}\mathbf{w}-\mathbf{y})^T(\mathbf{X}\mathbf{w}-\mathbf{y}) + \lambda\|\mathbf{w}\|_1.$$

Show that when $\mathbf{X}^T\mathbf{X}$ is an identity matrix, the solution is

$$\mathbf{w}^{lasso} = sign(\mathbf{w}^{ls})(|\mathbf{w}^{ls}| - \lambda)^+.$$

[hints : $sign(\mathbf{w})$ gives sign of the element of vector $\mathbf{w}$, i.e., $sign([9,-8,0]) = [1,-1,0]$. The modulo function $|\mathbf{w}|$ gives absolution values , i.e., $|[9,-8,0]| = [9,8,0]$ and $(x_i)^+ = max\{0, x_i\}$]

(b) (4 points) Ordinary Least Square (OLS) regression solves

$$\min_{\mathbf{w}} E(\mathbf{w}) = \frac{1}{2}(\mathbf{X}\mathbf{w}-\mathbf{y})^T(\mathbf{X}\mathbf{w}-\mathbf{y}).$$

In gradient descent approach in every iteration the parameter $\mathbf{w}$ is updated using following equation:

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \alpha\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}}|_{\mathbf{w}=\mathbf{w}^t} = \mathbf{w}^t - \alpha\mathbf{X}^T(\mathbf{X}\mathbf{w}^t - \mathbf{y}).$$

Derive the update equation of parameter $\mathbf{w}$ to solve following problem using gradient descent approach.

i. Ridge Regression : $E(\mathbf{w}) = \frac{1}{2}(\mathbf{X}\mathbf{w}-\mathbf{y})^T(\mathbf{X}\mathbf{w}-\mathbf{y}) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$

ii. Weighted least square regression minimize weighted average of square loss of all data points, where weight for the i^{th} data point is r_i then weighted least square regression solves $E(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^n r_i(\mathbf{x}_i^T\mathbf{w} - y_i)^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$

(c) (2 points) Consider the regression problem where the input $\mathbf{x} \in \mathbb{R}^p$ and the output $y \in \mathbb{R}$. Assume that the likelihood is specified in terms of the unknown parameter $\mathbf{w} \in \mathbb{R}^p$ as $p(y|X = \mathbf{x}, \mathbf{w}) \sim \mathcal{N}(\langle \mathbf{w}, \mathbf{x} \rangle, g(\mathbf{x}))$, where $g : \mathbb{R}^p \to \mathbb{R}^+$ is a known non-negative function. We are given the training sample $(\mathbf{x}_i, y_i)_{i=1}^n$ drawn independently according to the joint probability measure P on $X \times Y$. Assume a prior distribution on $\mathbf{w} : \mathbf{w} \sim \mathcal{N}(0, \Lambda)$, where Λ is a diagonal matrix with the diagonal entries given by $\lambda_{ii} \geq 0$). Find the maximum a posteriori estimate of w?

(d) Discuss whether MAP estimates of w are less prone to over-fitting than MLE. In case of regression, what assumption on prior distribution will make the MAP estimates is same as the MLE?

Exercise 2: Bias Variance Trade-off

(a) (3 points) What is bias-variance trade-off? Prove that Err = irreducible error+ bias^2 + variance.