MATH 156 HOMEWORK 1

Due Apr. 7

Exercise 1 (Method of Least Squares)**.** Suppose we have matrices $\mathbf{Y} \in \mathbb{R}^{d\times n}$ and $\mathbf{X} \in \mathbb{R}^{d\times r}$. We seek to find a matrix $\hat{\mathbf{B}} \in \mathbb{R}^{r\times n}$ where

$$\hat{\mathbf{B}} = \operatorname*{argmin}_{\mathbf{B}\in\mathbb{R}^{r\times n}} \|\mathbf{Y}-\mathbf{X}\mathbf{B}\|_F^2 + \lambda\|\mathbf{B}\|_F^2. \tag{1}$$

Here $\lambda \geq 0$ is called the L_2-regularization parameter. (This is an instance of unconstrained quadratic optimization problem.)

(i) Show that

$$\begin{aligned} \|\mathbf{Y}-\mathbf{X}\mathbf{B}\|_F^2 + \lambda\|\mathbf{B}\|_F^2 &= \operatorname{tr}(\mathbf{Y}-\mathbf{X}\mathbf{B})^T(\mathbf{Y}-\mathbf{X}\mathbf{B}) + \lambda\operatorname{tr}(\mathbf{B}^T\mathbf{B}) \quad (2)\\ &= \operatorname{tr}(\mathbf{Y}^T\mathbf{Y}) - 2\operatorname{tr}(\mathbf{Y}^T\mathbf{X}\mathbf{B}) + \operatorname{tr}(\mathbf{B}^T\mathbf{X}^T\mathbf{X}\mathbf{B}) + \lambda\operatorname{tr}(\mathbf{B}^T\mathbf{B}). \quad (3)\end{aligned}$$

(ii) Show that (use Exercise 2)

$$\frac{\partial}{\partial\mathbf{B}}\left(\|\mathbf{Y}-\mathbf{X}\mathbf{B}\|_F^2 + \lambda\|\mathbf{B}\|_F^2\right) = 2(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I})\mathbf{B} - 2\mathbf{X}^T\mathbf{Y}, \qquad \frac{\partial^2}{\partial\mathbf{B}^2}\left(\|\mathbf{Y}-\mathbf{X}\mathbf{B}\|_F^2 + \lambda\|\mathbf{B}\|_F^2\right) = 2(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}). \tag{4}$$

(iii) From **(ii)**, conclude that the quadratic function $\mathbf{B} \mapsto \|\mathbf{Y}-\mathbf{X}\mathbf{B}\|_F^2$ is convex, and and hence its every critical point is a local minimum. (See Ref)

(iv) Suppose $\lambda = 0$ and $\mathbf{X}^T\mathbf{X}$ is invertible[1]. Then from **(ii)** and **(iii)**, conclude that the quadratic function $\mathbf{B} \mapsto \|\mathbf{Y}-\mathbf{X}\mathbf{B}\|_F^2 + \lambda\|\mathbf{B}\|_F^2$ has a unique global minimum $\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ [2]

(v) Suppose $\lambda > 0$. Then argue that $\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}$ is always invertible[3], and the quadratic function $\mathbf{B} \mapsto \|\mathbf{Y}-\mathbf{X}\mathbf{B}\|_F^2 + \lambda\|\mathbf{B}\|_F^2$ has a unique global minimum $\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}$.

Exercise 2 (Matrix derivatives)**.** Show the following matrix derivatives: (Ref: The Matrix Cookbook)

(i) (First order) $\quad \frac{\partial}{\partial\mathbf{X}}\operatorname{tr}(\mathbf{B}\mathbf{X}\mathbf{C}) = \mathbf{B}^T\mathbf{C}^T, \qquad \frac{\partial}{\partial\mathbf{X}}\operatorname{tr}(\mathbf{B}\mathbf{X}^T\mathbf{C}) = \mathbf{C}\mathbf{B}.$

(ii) (Second order) $\quad \frac{\partial}{\partial\mathbf{X}}\operatorname{tr}(\mathbf{X}^T\mathbf{B}\mathbf{X}) = \mathbf{B}\mathbf{X}+\mathbf{B}^T\mathbf{X}, \qquad \frac{\partial}{\partial\mathbf{X}}\operatorname{tr}(\mathbf{B}^T\mathbf{X}^T\mathbf{C}\mathbf{X}\mathbf{B}) = \mathbf{C}^T\mathbf{X}\mathbf{B}\mathbf{B}^T + \mathbf{C}\mathbf{X}\mathbf{B}\mathbf{B}^T.$

Exercise 3. Fix $\mathbf{w} = [w_0, w_1, \dots, w_M] \in \mathbb{R}^{M+1}$ and $\sigma \geq 0$. Let $\hat{Y}_1, \dots, \hat{Y}_N$ be independent Gaussian RVs where $\hat{Y}(x_i; \mathbf{w}, \sigma) \sim N(\boldsymbol{\phi}(\mathbf{x})^T\mathbf{w}, \sigma^2)$ for $i \in \{1, \dots, N\}$, where $\boldsymbol{\phi}(x) = [1, x, \dots, x^M]^T$. Show that the joint likelihood function for observing the values $y_1, \dots, y_N$ is given by

$$L(y_1,\dots,y_N;\mathbf{w},\sigma) = (2\pi\sigma^2)^{-N/2}\exp\left[-\frac{1}{2\sigma^2}\|\mathbf{Y}-\mathbf{X}\mathbf{w}\|_F^2\right], \tag{5}$$

where

$$\mathbf{Y} = [y_1,\dots,y_N]^T \in \mathbb{R}^N, \qquad \mathbf{X} = \begin{bmatrix} 1 & x_1 & \cdots & x_1^M \\ & \vdots & & \\ 1 & x_2 & \cdots & x_N^M \end{bmatrix} \in \mathbb{R}^{N\times M}. \tag{6}$$

Exercise 4. Using the Jupyter notebook provided in the course repository, reproduce the Figures 2-6 in the lecture note, where the data are independently generated from the following random variable

$$Y = \cos(2\pi X) + \varepsilon, \tag{7}$$

where $X \sim \text{Uniform}([0,1])$ and $\varepsilon \sim N(0, 0.16)$ are independent. (Include screen shots of the plots you generate in your solution)

---

[1] $\mathbf{X}^T\mathbf{X}$ is symmetric and positive semidefinite, and it is invertible iff the singular values of $\mathbf{X}$ are all nonzero.

[2] The matrix $\mathbf{X}^\dagger := (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is called the *Moore-Penrose pseudo-inverse* of $\mathbf{X}$. If $\mathbf{X}$ is square and invertible, then $\mathbf{X}^\dagger = \mathbf{X}^{-1}(\mathbf{X}^T)^{-1}\mathbf{X}^T = \mathbf{X}^{-1}$. So the the psuedo-inverse can be regarded as a generalization of matrix inverse for non-square matrices.

[3] Hint: Show that $\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}$ is positive definite if $\lambda > 0$. Use the fact that the eigenvalues of a positive definite matrix $\mathbf{A}$ has to be positive (why?), so $\mathbf{A}\mathbf{y} \neq \mathbf{0}$ for any $\mathbf{y}$ (why?) so $\mathbf{A}$ is invertible (why?).

Exercise 5. Using the Jupyter notebook provided in the course repository, reproduce the Figures 8-9 in the lecture note, where the data are independently generated from the following random variable

$$Y = \cos(2\pi X) + \varepsilon, \tag{8}$$

where $X \sim \text{Uniform}([0,1])$ and $\varepsilon \sim N(0, \sigma^2)$ are independent. Compare the results of the maximum likelihood and the Bayesian polynomial regression. (Include screen shots of the plots you generate in your solution))