1 Probabilistic Model of Linear Regression

Both ordinary least squares and ridge regression have interpretations from a probabilistic standpoint. In particular, assuming a generative model for our data and a particular noise distribution, we will derive least squares and ridge regression as the maximum likelihood and maximum *a-posteriori* parameter estimates, respectively. This problem will walk you through a few steps to do that. (Along with some side digressions to make sure you get a better intuition for ML and MAP estimation.)

(a) Assume that X and Y are both one-dimensional random variables, i.e. $X, Y \in \mathbb{R}$. Assume an affine model between X and Y: $Y = Xw_1 + w_0 + Z$, where $w_1, w_0 \in \mathbb{R}$, and $Z \sim N(0, 1)$ is a standard normal (Gaussian) random variable. Assume w_1, w_0 are fixed parameters (i.e., they are not random). **What is the conditional distribution of Y given X?**

(b) Given n points of training data $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ generated in an iid fashion by the probabilistic setting in the previous part, **derive the maximum likelihood estimator for w_1, w_0 from this training data.**

(c) Now, consider a different generative model. Let $Y = Xw + Z$, where $Z \sim U[-0.5, 0.5]$ is a continuous random variable uniformly distributed between -0.5 and 0.5. Again assume that w is a fixed parameter. **What is the conditional distribution of Y given X?**

(d) Given n points of training data $\{(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)\}$ generated in an i.i.d. fashion in the setting of the part (c) **derive a maximum likelihood estimator of w.** Assume that $X_i > 0$ for all $i = 1, \ldots, n$. (Note that MLE for this case need not be unique; but you are required to report only one particular estimate.)

(e) Take the model $Y = Xw + Z$, where $Z \sim U[-0.5, 0.5]$. **Use a computer to simulate n training samples $\{(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)\}$ and illustrate what the likelihood of the data looks like as a function of w after $n = 5, 25, 125, 625$ training samples. Qualitatively describe what is happening as n gets large.**

(You have total freedom in using any python libraries for this problem part. No restrictions.)

(f) (One-dimensional Ridge Regression) Now, let us return to the case of Gaussian noise. Given n points of training data $\{(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)\}$

generated according to $Y_i = X_i W + Z_i$, where $Z_i \sim N(0,1)$ are iid standard normal random variables. Assume $W \sim N(0, \sigma^2)$ is also a standard normal and is independent of both the Z_i's and the X_i's. **Use Bayes' Theorem to derive the posterior distribution of W given the training data. What is the mean of the posterior distribution of W given the data?**

Hint: Compute the posterior up-to proportionality and try to identify the distribution by completing the square.

(g) Consider n training data points $\{(\vec{x}_1, Y_1), (\vec{x}_2, Y_2), \cdots, (\vec{x}_n, Y_n)\}$ generated according to $Y_i = \vec{w}^\top \vec{x}_i + Z_i$ where $Y_i \in \mathbb{R}, \vec{w}, \vec{x}_i \in \mathbb{R}^d$ with $\vec{w}$ fixed, and $Z_i \sim N(0,1)$ iid standard normal random variables. **Argue why the maximum likelihood estimator for $\vec{w}$ is the solution to a least squares problem.**

(h) (Multi-dimensional ridge regression) Consider the setup of the previous part: $Y_i = \vec{W}^\top \vec{x}_i + Z_i$, where $Y_i \in \mathbb{R}, \vec{W}, \vec{x}_i \in \mathbb{R}^d$, and $Z_i \sim N(0,1)$ iid standard normal random variables. Now we treat $\vec{W}$ as a random vector and assume a prior knowledge about its distribution. In particular, we use the prior information that the random variables W_j are i.i.d. $\sim N(0, \sigma^2)$ for $j = 1, 2, \ldots, d$. **Derive the posterior distribution of $\vec{W}$ given all the $\vec{x}_i, Y_i$ pairs. What is the mean of the posterior distribution of the random vector $\vec{W}$?**

Hint: Use hints from part (f) and the following identities: For $\mathbf{X} = \begin{bmatrix} \vec{x}_1^\top \\ \vdots \\ \vec{x}_n^\top \end{bmatrix}$ and $\vec{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$ we have $\mathbf{X}^\top \mathbf{X} = \sum_{i=1}^n \vec{x}_i \vec{x}_i^\top$ and $\mathbf{X}^T \vec{Y} = \sum_{i=1}^n \vec{x}_i Y_i$.

(i) Consider $d = 2$ and the setting of the previous part. **Use a computer to simulate and illustrate what the *a-posteriori* probability looks like for the W model parameter space after $n = 5, 25, 125$ training samples for different values of σ^2.** (You have total freedom in using any python libraries for this problem part. No restrictions.)