4033/5033 Assignment: MLE and WLS

Sasank Sribhashyam

In this assignment, we will derive a connection between a variant of the weighted least square problem in HW1 and MAP estimation. Consider the following regularized weighted least square (RWLS) problem

$$\min_{\beta} \sum_{i=1}^{n} w_i \cdot (x_i^T \beta - y_i)^2 + \lambda \beta^T \beta, \tag{1}$$

where w_i is the weight for instance x_i.

Task 1. Design a density estimation problem with proper probabilistic assumptions and prove this problem is equivalent to the above RWLS problem. In your answer,

1. Explain every notation you will be using before writing down any assumption or analysis. The explanation of each notation should include its dimension and interpretation. Below are two example answers. Please follow their formats as much as possible. – X is a p-dimensional feature vector of an instance. – ε is a random noise following a Gaussian distribution $N(0, v^2)$.

2. List all the assumptions.

3. Elaborate the arguments. You don't need to show the exact MAP estimate is same as the RWLS problem. You only need to show the MAP estimation problem (i.e., maximizing posterior) is equivalent to another problem which has the same form as RWLS. Tip: In the lectures, we assume ε follows the same distribution $N(0, v^2)$ for every instance. But its distribution may be different for different instances.

To derive a connection between the regularized weighted least squares (RWLS) problem and a density estimation problem, we can start by defining the probabilistic assumptions and notations for the density estimation problem. Then, we will show how this problem is equivalent to the RWLS problem.

Let's define the notations before proceeding:

Notations:

- X: This is a p-dimensional feature vector representing an instance. Its dimension is $(p, 1)$, and it contains the input features for each instance.
- β: The parameter vector we want to estimate. It is also p-dimensional, with dimensions $(p, 1)$.
- y_i: The observed target value for the ith instance. It is a scalar.
- w_i: The weight for the ith instance. It is a scalar.
- ϵ_i: The random noise associated with the ith instance. It follows a Gaussian distribution with mean 0 and variance σ_i^2. It is a scalar.
- λ: The regularization parameter, a positive scalar.

Now, let's define the density estimation problem as finding the joint probability distribution of the data under these assumptions:

Assumptions:

1. The target variable y_i is generated from a linear model with added Gaussian noise ϵ_i for each instance i:

$$y_i = X_i^T \beta + \epsilon_i$$

2. The noise ϵ_i for each instance follows a Gaussian distribution $N(0, \sigma_i^2)$.

$$\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$$

3. **Prior on β:** We place a Gaussian prior distribution on the parameter vector β:

$$\beta \sim \mathcal{N}(0, \lambda^{-1} I_p)$$

   where λ is the regularization parameter, and I_p is the $p \times p$ identity matrix.

With these notations and assumptions in place, we can now formulate the density estimation problem and establish its equivalence to the RWLS problem.

Arguements:

We want to estimate the posterior distribution of β given the data $\{y_i, X_i\}$ for $i = 1$ to n. Using Bayes' theorem, the posterior distribution is proportional to the likelihood and the prior:

$$P(\beta|\{y_i, X_i\}) \propto P(\{y_i\}|\{X_i\}, \beta) \cdot P(\beta)$$

Now, let's break down the terms:

1. **Likelihood** ($P(\{y_i\}|\{X_i\}, \beta)$): Given our assumptions, the likelihood term is the product of n independent Gaussian distributions for each instance i:

$$P(y_i|X_i, \beta) \propto \exp\left(-\frac{1}{2v_i^2}(Y_i - X_i^\top \beta)^2\right)$$

$$P(\{y_i\}|\{X_i\}, \beta) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{1}{2}\frac{(y_i - X_i^T \beta)^2}{\sigma_i^2}\right)\right)$$

2. **Prior** ($P(\beta)$): We assume a Gaussian prior for β with mean 0 and covariance matrix proportional to the identity matrix (to encourage sparsity):

$$P(\beta) \propto \exp\left(-\frac{\lambda}{2}\beta^\top \beta\right)$$

$$P(\beta) = \left(\frac{1}{\sqrt{(2\pi)^p |\lambda I|}}\right) \exp\left(-\frac{1}{2}\beta^T (\lambda I)^{-1} \beta\right)$$

Now, let's rewrite the posterior distribution:

$$P(\beta|\{y_i, X_i\}) \propto \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{1}{2} \frac{(y_i - X_i^T\beta)^2}{\sigma_i^2} \right) \right) \cdot \left(\frac{1}{\sqrt{(2\pi)^p |\lambda I|}} \right) \exp\left(-\frac{1}{2} \beta^T (\lambda I)^{-1} \beta \right)$$

Taking the logarithm for convenience:

$$\log(P(\beta|\{y_i, X_i\})) \propto -\frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - X_i^T\beta)^2}{\sigma_i^2} - \frac{1}{2} \beta^T (\lambda I)^{-1} \beta + \text{constant}$$

This log-posterior is equivalent in form to the RWLS problem with a specific choice of weights and regularization term:

$$\min_{\beta} \sum_{i=1}^{n} w_i (y_i - X_i^T\beta)^2 + \lambda \beta^T \beta$$

Here, we identify that: - w_i corresponds to $\frac{1}{2\sigma_i^2}$ - λ corresponds to $\frac{1}{2\sigma^2}$, where σ^2 is a common variance for all instances.

We have demonstrated that the MAP estimation problem, when formulated with the given probabilistic assumptions, is equivalent in form to the RWLS problem with appropriately chosen weights and regularization term.