CS 760: Machine Learning Spring 2024

Homework 2: Linear Regression

AUTHORS: Jed Pulley

DO NOT POLLUTE! AVOID PRINTING, OR PRINT 2-SIDED MULTIPAGE.

Problem 2.1

To minimize the MSE of θ, we computer our estimator $\hat{\theta}$:

$$\hat{\theta} = \arg\min_{\theta} \|\mathsf{y} - \mathbf{X}\theta\|_{\mathbf{2}}^{\mathbf{2}}$$

Expanding that out, we find:

$$\begin{aligned} \|\mathsf{y} - \mathbf{X}\theta\|_{\mathbf{2}}^{\mathbf{2}} &= (\mathsf{y} - \mathbf{X}\theta)^{\mathsf{T}}(\mathsf{y} - \mathbf{X}\theta) \\ &= \mathsf{y}^{\mathsf{T}}\mathsf{y} - \mathsf{y}^{\mathsf{T}}\mathbf{X}\theta - \theta^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathsf{y} + \theta^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\theta \\ &= \mathsf{y}^{\mathsf{T}}\mathsf{y} - \mathbf{2}\theta^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathsf{y} + \theta^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\theta \end{aligned}$$

From there, we take the derivative w.r.t. θ:

$$\frac{d}{d\theta} = 2\mathbf{X}^{\mathsf{T}}\mathbf{X}\theta - \mathbf{2}\mathbf{X}^{\mathsf{T}}\mathsf{y}$$

After setting this derivative to zero and solving for θ, we find that our solution is:

$$\hat{\theta} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{\mathsf{T}}\mathsf{y}$$

Problem 2.2

Since know that ϵ is distributed normally, we can use the general multivariate normal distribution formula to derive our expression for the MLE of θ^*:

$$\mathbb{P}(\mathsf{y}, \mathbf{X}|\theta, \mathbf{\Sigma}^*) = \frac{\mathbf{1}}{\mathbf{2}\pi^{\frac{\mathbf{n}}{\mathbf{2}}}|\mathbf{\Sigma}^*|^{\frac{\mathbf{1}}{\mathbf{2}}}}\mathbf{e}^{-\frac{\mathbf{1}}{\mathbf{2}}(\mathsf{y}-\mathbf{X}\theta)^{\mathbf{T}}\mathbf{\Sigma}^{*-\mathbf{1}}(\mathsf{y}-\mathbf{X}\theta)}$$

To find the MLE, we first take the natural log, then we take the derivative and set it to zero. After that, we solve for $\hat{\theta}$:

$$\hat{\theta} = (\mathbf{X}^\intercal \mathbf{\Sigma}^{*-1} \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{\Sigma}^{*-1} \mathsf{y}$$

This is very similar to the solution of Problem 2.1 above, with the exception that it's now being weighted by the inverse covariance matrix Σ^{*-1}

Problem 2.3

Thanks to the central limit theorem, we can assume that θ^* is distributed normally under large sample sizes. The mean of this distribution is found to be θ^*, but our covariance matrix is a little trickier to find. Since we don't know exactly the structure of Σ^*, we can estimate it using the Fisher Information matrix, which is given below:

$$I(\theta^*) = -\mathbb{E}[\frac{\partial^2}{\partial \theta^2} ln(\mathbb{P}(\mathsf{y}, \mathbf{X}|\theta, \mathbf{\Sigma}^*))]$$

Therefore, the distribution of θ^* can be shows as:

$$\theta^* \sim \mathcal{N}(\theta^*, I(\theta^*)^{-1})$$

Problem 2.4

Given a new feature vector x, our new response is: $\hat{y} = x^T \hat{\theta}$. Letting $H = (\mathbf{X}^\intercal \mathbf{\Sigma}^{*-1} \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{\Sigma}^{*-1}$, given the above equation for $\hat{\theta}$, we can show that the MLE of our new response is $\hat{y} = x^T \theta^* + x^T H \epsilon$

Problem 2.5

Using H that we defined above, we can show the distribution of $\hat{y}$ below:

$\hat{y} \sim \mathcal{N}(x^T \theta^*, x^T H \Sigma^* H^T x)$

Problem 2.6

We'll take a similiar approach as in Problem 2.2, so first we take the natural log, take the derivative w.r.t. Σ^*, set it to zero, and then solve for Σ^*. Doing so, we get:

$$\Sigma^* = -X(\mathsf{y} - \mathsf{X}\theta)$$

Problem 2.7

(a)