Problem Set 1*

Tasha Pais†

1 Linear Regression

Problem 1

Given the original regression problem, we have:

$$y^{(i)} = w^T x^{(i)} \tag{1}$$

where $y^{(i)}$ is the label for the i^{th} data point, w is the weight vector, and $x^{(i)}$ is the feature vector for the i^{th} data point.

Now, suppose we transform the labels as:

$$\tilde{y}^{(i)} = a y^{(i)} + b \tag{2}$$

for some constants a and b.

The new regression problem becomes:

$$\tilde{y}^{(i)} = \tilde{w}^T x^{(i)} \tag{3}$$

Given the transformation, we can express $\tilde{y}^{(i)}$ in terms of the original model:

$$\tilde{y}^{(i)} = a(w^T x^{(i)}) + b \tag{4}$$

$$\tilde{y}^{(i)} = a w^T x^{(i)} + b \tag{5}$$

Now, recall that we assumed the first dimension of $x^{(i)}$ is always 1. This means that the first element of the weight vector w (or $\tilde{w}$) acts as the bias term. Let's denote the first element of w as w_1 and the first element of $\tilde{w}$ as $\tilde{w}_1$.

From the equation above, we can deduce:

$$\tilde{w}_1 = w_1 a + b \tag{6}$$

and for $j > 1$:

$$\tilde{w}_j = a w_j \tag{7}$$

This gives us the mapping g from w^* to $\tilde{w}$ given a and b:

$$\tilde{w}_1 = w_1 a + b \tag{8}$$

$$\tilde{w}_j = a w_j \quad \text{for} \quad j > 1 \tag{9}$$

In essence, each weight in $\tilde{w}$ is a scaled version of the corresponding weight in w^* by the factor a except for the bias term, which gets an additional shift by b.

*Due: `September 27, 2023`, Student(s) worked with: `Arul Elango`

†NetID: `tdp74`, Email: `tdp74@scarletmail.rutgers.edu`

Problem 2

Given the original regression problem:

$$y^{(i)} = w^{*T} x^{(i)} \tag{10}$$

Now, the inputs are transformed as:

$$\bar{x}_j^{(i)} = c_j x_j^{(i)} \tag{11}$$

for some nonzero constants $c_1, \ldots, c_d \in \mathbb{R}$.

The new regression problem with the transformed inputs is:

$$y^{(i)} = \bar{w}^T \bar{x}^{(i)} \tag{12}$$

Given the transformation, we can express $y^{(i)}$ in terms of the original model:

$$y^{(i)} = w^{*T} x^{(i)} = \bar{w}^T (c \odot x^{(i)}) \tag{13}$$

where $\odot$ denotes element-wise multiplication.

From the equation above, we can deduce the relationship between the weights of the transformed model and the original model:

$$\bar{w}_j = \frac{w_j^*}{c_j} \quad \text{for} \quad j = 1, 2, \ldots, d \tag{14}$$

Thus, we can obtain $\bar{w}$ directly from w^* without retraining on the new dataset. The mapping h from w^* to $\bar{w}$ given the constants $c_1, \ldots, c_d$ is:

$$\bar{w}_j = \frac{w_j^*}{c_j} \quad \text{for} \quad j = 1, 2, \ldots, d \tag{15}$$

Problem 3

Given the model:

$$y^{(i)} = w_{\text{true}}^T x^{(i)} + \epsilon_i \tag{16}$$

where $\epsilon_i \sim N(0, \sigma_i^2)$ is a sample-specific Gaussian noise.

Likelihood: MLE seeks the parameter values under which the observed data is most probable. The likelihood is a measure of how well the model with parameters w explains or fits the observed data. For a single data point $(x^{(i)}, y^{(i)})$, the term $p(y^{(i)}|x^{(i)}, w)$ represents the probability (under the model with parameters w) of observing the output $y^{(i)}$ given the input $x^{(i)}$.

The likelihood of observing $y^{(i)}$ given $x^{(i)}$ and w is:

$$p(y^{(i)}|x^{(i)}, w) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(y^{(i)} - w^T x^{(i)})^2}{2\sigma_i^2}\right) \tag{17}$$

The joint likelihood for the entire dataset is the product of the individual likelihoods since the samples are independently generated:

$$L(w) = \prod_{i=1}^{N} p(y^{(i)}|x^{(i)}, w) \tag{18}$$

The notation $\prod_{i=1}^{N}$ is the product notation, analogous to the Σ notation for summation. It means that we're multiplying together the individual likelihoods $p(y^{(i)}|x^{(i)}, w)$ for all N data points in the dataset.

Optimization: To find the maximum likelihood estimate, we'll maximize the likelihood (or equivalently, the log-likelihood). The objective is to find the parameter values w that maximize the likelihood function. Formally, this is represented as:

$$\hat{w}_{MLE} = \arg\max_w L(w) \tag{19}$$

where $\hat{w}_{MLE}$ is the estimate of w that maximizes the likelihood function $L(w)$.

Often, it's more convenient to work with the log-likelihood due to its mathematical properties. The objective in terms of the log-likelihood is:

$$\hat{w}_{MLE} = \arg\max_w \log L(w) \tag{20}$$

To achieve this optimization, one would typically differentiate the log-likelihood with respect to w, set the result to zero, and solve for w to find the value that maximizes the function. Depending on the nature of the likelihood function, this might yield a closed-form solution, or it might require numerical methods for optimization.

The expanded formula using the joint likelihood equation above is:

$$\log L(w) = \sum_{i=1}^{N} \left(-\frac{1}{2} \log\left(2\pi\sigma_i^2\right) - \frac{(y^{(i)} - w^T x^{(i)})^2}{2\sigma_i^2} \right) \tag{21}$$

$$\hat{w}_{MLE} = \arg\max w \left[\sum_{i=1}^{N} \left(-\frac{1}{2} \log\left(2\pi\sigma_i^2\right) - \frac{(y^{(i)} - w^T x^{(i)})^2}{2\sigma_i^2} \right) \right] \tag{22}$$

Closed-form solution: To maximize this with respect to w, we can set its gradient to zero. The gradient of a function gives the direction of steepest ascent. In the context of a scalar-valued function of a vector (like the likelihood function with respect to the parameter vector w, the gradient is a vector of the function's partial derivatives with respect to each component of w.

Given the log-likelihood function:

$$\log L(w) = \sum_{i=1}^{N} \left(-\frac{1}{2} \log\left(2\pi\sigma_i^2\right) - \frac{(y^{(i)} - w^T x^{(i)})^2}{2\sigma_i^2} \right) \tag{23}$$

The maximum likelihood estimate $\hat{w}_{MLE}$ is given by:

$$\hat{w}_{MLE} = \arg\max_w \left[\sum_{i=1}^{N} \left(-\frac{1}{2} \log\left(2\pi\sigma_i^2\right) - \frac{(y^{(i)} - w^T x^{(i)})^2}{2\sigma_i^2} \right) \right] \tag{24}$$

To find the value of w that maximizes this function, we differentiate with respect to w and set the result to zero.

This leads to an equation of the form:

$$\mathbf{X}^T \Sigma^{-1} \mathbf{y} = \mathbf{X}^T \Sigma^{-1} \mathbf{X} w \tag{25}$$

From the above equation, we can express $\hat{w}_{MLE}$ in closed form as:

$$\hat{w}_{MLE} = (\mathbf{X}^T \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}^T \Sigma^{-1} \mathbf{y} \tag{26}$$

This solution provides the maximum likelihood estimate for w under the given model with non-identically distributed Gaussian noise.