Homework #1

Claire Jung

May 1, 2023

Problem 1

Show:

$$E[(y - \mathbf{w*}^T\mathbf{x})\mathbf{a}^T\mathbf{x})] = 0$$

The idea is to first show the correlation is zero, then we need to show that $E[y - \mathbf{w*}^T\mathbf{x}] = 0$. But first, let us revisit a derivation from class.

We know that

$$E[(\mathbf{w*}^T\mathbf{x} - \hat{\mathbf{w}}^T\mathbf{x})(\mathbf{y} - \mathbf{w*}^T\mathbf{x})] = 0$$

multiplying it out and considering the assumption that $E[\hat{\mathbf{w}}^T\mathbf{x}] = \mathbf{w*}^T\mathbf{x}$,

$$E[\mathbf{y}E[\hat{\mathbf{w}}^T\mathbf{x}]] - E[E[\hat{\mathbf{w}}^T\mathbf{x}]^2] - E[E[\hat{\mathbf{w}}^T\mathbf{x}]\mathbf{y}] + E[E[\hat{\mathbf{w}}^T\mathbf{x}]\hat{\mathbf{w}}^T\mathbf{x}] = 0$$

as $\mathbf{y}$ is deterministic, and $E[E[x]] = E[x]$ the expression clearly cancels out. Armed with the fact that we can now state $E[\hat{\mathbf{w}}^T\mathbf{x}] = \mathbf{w*}^T\mathbf{x}$ Using the definition of correlation from the problem, we can set $U = y - E[\hat{\mathbf{w}}^T\mathbf{x}]$ and $V = \mathbf{a}^T\mathbf{x}$. The correlation will be zero if $(U - E[U]) = 0$. Plugging in U:

$$y - E[\hat{\mathbf{w}}^T\mathbf{x}] - E[y - E[\hat{\mathbf{w}}^T\mathbf{x}]] = y - E[\hat{\mathbf{w}}^T\mathbf{x}] - E[y] - E[\hat{\mathbf{w}}^T\mathbf{x}] = 0$$

As the correlation is zero, we can state $E[UV] = E[U]E[V]$. As $\mathbf{w*}^T\mathbf{x}$ is the best possible fit, with zero error, the sum of $\mathbf{y} - \mathbf{w*}^T\mathbf{x}$ is zero and therefore, $E[\mathbf{y} - \mathbf{w*}^T\mathbf{x}] = 0$ Thus, $E[UV] = 0$

Another way of considering the solution, minimizing the least squared loss via gradient descent yields the argmin_w of $\sum_i^N(\mathbf{y}_i - \hat{\mathbf{w}}^T\mathbf{x})\mathbf{x}_i$ and $\mathbf{w*}$ is the $\mathbf{w}$ that has zero loss. Therefore, the summation is zero and so is the expectation value.

Problem 2

Already solved in Problem 1. If $E[UV] = 0$ and $E[U] = 0$, then $\rho(U, V) = 0$.

Problem 3

Show,

$$\hat{\mathbf{w}} \cdot \mathbf{x} = \hat{\tilde{\mathbf{w}}} \cdot \tilde{\mathbf{x}}$$

Naively, as both $\tilde{w}$ and $\hat{w}$ were trained on the same y and minimized the same cost, their predictions should be the same. Thus $\hat{\tilde{w}} \cdot x = \hat{w} \cdot x$

The cost is minimized for a trained w thus the derivative is zero and we can write an expression for the jth component of w as:

$$\sum_i^N (y_i - \hat{w} \cdot x_i) x_{ij} = 0$$

Similarly for $\tilde{w}$

$$\sum_i^N (y_i - \hat{\tilde{w}} \cdot \tilde{x}_i) \tilde{x}_{ij} = 0$$

As $\tilde{x}_{ij} = c_j x_{ij}$,

$$\sum_i^N (y_i - \hat{\tilde{w}} \cdot \tilde{x}_i) c_j x_{ij} = 0 \rightarrow \sum_i^N (y_i - \hat{\tilde{w}} \cdot \tilde{x}_i) x_{ij} = 0$$

combining the two summations,

$$\sum_i^N (y_i - \hat{\tilde{w}} \cdot \tilde{x}_i) x_{ij} = \sum_i^N (y_i - \hat{w} \cdot x_i) x_{ij}$$

$$\sum_i^N (y_i - \hat{\tilde{w}} \cdot \tilde{x}_i) = \sum_i^N (y_i - \hat{w} \cdot x_i)$$

$$\sum_i^N (\hat{\tilde{w}} \cdot \tilde{x}_i) = \sum_i^N (\hat{w} \cdot x_i)$$

$$\hat{\mathbf{w}} \cdot \mathbf{x} = \hat{\tilde{\mathbf{w}}} \cdot \tilde{\mathbf{x}}$$

Problem 4

We can write the ML estimate as,

$$ML = -\frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - f(x_i; w))^2 - \frac{N}{2} \log(2\pi\sigma^2)$$

maximizing ML, we can set $\frac{\partial ML}{\partial \sigma^2} = 0$. Thus,

$$\frac{\partial ML}{\partial \sigma^2} = \frac{1}{2\sigma^4} \sum_{i=1}^N (y_i - f(x_i; w))^2 - \frac{N}{2\sigma^2} = 0$$

Thus, we can write σ^2 as:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i; w))^2$$

Both the second and third degree polynomials fit the data well, but the second degree polynomial model has a lower validation loss. There is a clear inflection point in the data, which is why the linear model performs so poorly.

The third order polynomial model fits the data better for training, which is to be expected for a higher order polynomial. But, the validation loss is what matters, thus the second order polynomial model is the best model.

Model A is quadratic.

Problem 5

The validation loss values for the asymmetric loss are higher for all the models compared to the symmetric models. Also, visualizing the fits it is clear that the symmetric loss models are superior.

Again, the second order polynomial model was the best model, thus model B is a quadratic.

Model A had a lower test error than model B, therefore model A is a better choice.