ECE 469/ECE 568 - Machine Learning

Southern Illinois University

September 27, 2024

Regression viewed as Maximum Likelihood Estimation

- Regression can be viewed from a probabilistic perspective.
- Our goal is to make predictions for the output variable y given some new value of the input variable x on the basis of a set of training data set consisting of N for a single feature x. The data set for feature x is denoted as $(x_1, \cdots, x_N)^T$ and their corresponding output values as $(y_1, \cdots, y_N)^T$.
- Recall, our polynomial curve/model is

$$g(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M$$

- Note that we would like to train this model based on our date set such that we compute the optimal values for the weights $\mathbf{w} = [w_0, w_1, \cdots, w_k, \cdots, w_M]^T$.

Regression viewed as Maximum Likelihood Estimation

- We wish to express our uncertainty over the value of the output variable using a probability distribution.
- To this end, we can assume that given the value of x, the corresponding value of y has a Gaussian distribution with a mean equal to the model value $g(x, \mathbf{w})$.

$$y|(x, \mathbf{w}) \sim \mathcal{N}(g(x, \mathbf{w}), \sigma^2)$$

- The variance is also inverse of the precision parameter $\sigma^2 = \beta^{-1}$.
- The above probabilistic approach can be viewed as

$$y = g(x, \mathbf{w}) + \epsilon \quad \text{where } \ \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Regression viewed as Maximum Likelihood Estimation

- Recall that for our model: $y = g(\mathbf{x}, \mathbf{w}) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$, the output available is conditionally Gaussian given the input variables $\mathbf{x}$: That is $y|\mathbf{x} \sim \mathcal{N}(g(\mathbf{x}, \mathbf{w}), \sigma^2)$.
- Then it follows that the conditional mean of the output variable given the input variables $\mathbf{x}$ is

$$\mathbb{E}[y|\mathbf{x}] = g(\mathbf{x}, \mathbf{w})$$

- Notice that the optimal prediction, for a new value of $\mathbf{x}$, is given by the conditional mean of the output variable.

Regression viewed as Maximum Likelihood Estimation

- We can now use the training data $\{x, y\}$ to determine the values of the unknown parameters $\mathbf{w}$ and σ^2 (or β^{-1}) by the maximum likelihood estimation technique.
- If the data are drawn i.i.d. from the Gaussian distribution, then the likelihood function is given by

$$f(\mathbf{y}|\mathbf{x}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_n - g(x_n, \mathbf{w}))^2}{2\sigma^2} \right) \right)$$

- The log-likelihood function is

$$\ln(f(\mathbf{y}|\mathbf{x}, \mathbf{w}, \sigma^2)) = -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_n - g(x_n, \mathbf{w}))^2 - \frac{N}{2} \ln(\sigma^2) - \frac{N}{2} \ln(2\pi)$$

Regression viewed as Maximum Likelihood Estimation

- To estimate $\mathbf{w}$, we can maximize the log-likelihood function.
- The log-likelihood function is

$$\mathbf{w}_{\mathrm{MLE}} = \mathrm{argmax}_{\mathbf{w}} \ln(f(\mathbf{y}|\mathbf{x}, \mathbf{w}, \sigma^2))$$

where

$$\ln(f(\mathbf{y}|\mathbf{x}, \mathbf{w}, \sigma^2)) = -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_n - g(x_n, \mathbf{w}))^2 - \frac{N}{2} \ln(\sigma^2) - \frac{N}{2} \ln(2\pi)$$

- We have already shown that this maximization is equivalent to minimization of the sum squared error.

$$\mathbf{w}_{\mathrm{MLE}} = \left(\mathrm{argmin}_{\mathbf{w}} \sum_{i=1}^{N} (y_n - g(x_n, \mathbf{w}))^2 \right)$$

Regression viewed as Maximum Likelihood Estimation

- Similarly, we can also use maximum likelihood estimation to estimate the variance of the Gaussian conditional distribution.
- Maximizing $\ln(f(\mathbf{y}|\mathbf{x}, \mathbf{w}, \sigma^2))$ with respect to σ^2 yields

$$\sigma^2_{\text{MLE}} = \frac{1}{N} \sum_{n=1}^{N} (y_n - g(x_n, \mathbf{w}_{\text{MLE}}))^2$$

- After estimating the parameters $\mathbf{w}$ and σ^2, we can now make predictions for new values of x.
- This is because now we have a probabilistic model.
- The predictive distribution that gives the probability distribution over y can be found by substituting the corresponding estimates for σ^2_{MLE} and $\mathbf{w}_{\text{MLE}}$.

$$y \sim \mathcal{N}(g(x, \mathbf{w}_{\text{MLE}}), \sigma^2_{\text{MLE}})$$

Bias-variance trade-off

- Recall that quantifying the mean square error (MSE) incorporates both the bias and the variance.

$$\mathrm{MSE} \quad = \quad \underbrace{\mathbb{E}[(\hat{x} - \mathbb{E}[\hat{x}])^2]}_{\mathrm{Var}[\hat{x}]} + \underbrace{(\mathbb{E}[\hat{x}] - x)^2}_{\mathrm{bias}^2(\hat{x})}$$

- We would like our estimators to have small MSE, and the estimators should manage to keep both their bias and variance smaller too.

Bias-variance trade-off

- Generally, relationship between bias and variance is tightly linked to the machine learning concepts of capacity, underfitting and overfitting.
- When the test/generalization error is measured by MSE, increasing capacity leads to increase variance and decrease bias.