Tutorial Linear Regression

Thursday, 19 November 2020 17:33

Problem 1: Assume that we are given a dataset, where each sample x_i and regression target y_i is generated according to the following process

$$x_i \sim \text{Uniform}(-10, 10)$$
$$y_i = ax_i^3 + bx_i^2 + cx_i + d + \epsilon_i, \quad \text{where} \quad \epsilon_i \sim \mathcal{N}(0,1) \quad \text{and} \quad a, b, c, d \in \mathbb{R}.$$

The 3 regression algorithms below are applied to the given data. Your task is to say what the bias and variance of these models are (low or high). Provide a 1-2 sentence explanation to each of your answers.

a) Linear regression

d = 1

> Bias: high. Variance: low.
>
> A straight line cannot capture a degree 3 polynomial (thus underfitting the data).

b) Polynomial regression with degree 3

d = 3

> Bias: low. Variance: low.
>
> The model is same as the data generating process. We can achieve a good fit.

c) Polynomial regression with degree 10

d = 10

> Bias: low. Variance: high.
>
> Since we are using a polynomial regression with a degree much higher compared to the data generating process, the model will overfit the data.

Problem 2: Given is a training set consisting of samples $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ with respective regression targets $\boldsymbol{y} = \{y_1, y_2, \ldots, y_N\}$ where $\boldsymbol{x}_i \in \mathbb{R}^D$ and $y_i \in \mathbb{R}$.

Alice fits a linear regression model $f(\boldsymbol{x}_i) = \boldsymbol{w}^T \boldsymbol{x}_i$ to the dataset using the closed form solution for linear regression (normal equations).

Bob has heard that by transforming the inputs $\boldsymbol{x}_i$ with a vector-valued function $\boldsymbol{\Phi}$, he can fit an alternative function, $g(\boldsymbol{x}_i) = \boldsymbol{v}^T \boldsymbol{\Phi}(\boldsymbol{x}_i)$, using the same procedure (solving the normal equations). He decides to use a linear transformation $\boldsymbol{\Phi}(\boldsymbol{x}_i) = \boldsymbol{A}^T \boldsymbol{x}_i$, where $\boldsymbol{A} \in \mathbb{R}^{D \times D}$ has full rank.

a) Show that Bob's procedure will fit the same function as Alice's original procedure, that is $f(\boldsymbol{x}) = g(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^D$ (given that $\boldsymbol{w}$ and $\boldsymbol{v}$ minimize the training set error).

> Alice uses the normal equation directly and obtains
>
> $$\boldsymbol{w}^* = (\boldsymbol{X}^\mathrm{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathrm{T}\boldsymbol{y}.$$
>
> Bob fits the model to the transformed data and obtains
>
> $$\begin{aligned} \boldsymbol{v}^* &= ((\boldsymbol{X}\boldsymbol{A})^\mathrm{T}(\boldsymbol{X}\boldsymbol{A}))^{-1}(\boldsymbol{X}\boldsymbol{A})^\mathrm{T}\boldsymbol{y} \\ &= (\boldsymbol{A}^\mathrm{T}\boldsymbol{X}^\mathrm{T}\boldsymbol{X}\boldsymbol{A})^{-1}\boldsymbol{A}^\mathrm{T}\boldsymbol{X}^\mathrm{T}\boldsymbol{y} \qquad (\dagger) \\ &= \boldsymbol{A}^{-1}(\boldsymbol{X}^\mathrm{T}\boldsymbol{X})^{-1}(\boldsymbol{A}^\mathrm{T})^{-1}\boldsymbol{A}^\mathrm{T}\boldsymbol{X}^\mathrm{T}\boldsymbol{y} \qquad (\ddagger) \\ &= \boldsymbol{A}^{-1}(\boldsymbol{X}^\mathrm{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathrm{T}\boldsymbol{y} = \boldsymbol{A}^{-1}\boldsymbol{w}^*. \end{aligned}$$
>
> (Note that $\boldsymbol{\Phi}$ transforms the column vectors $\boldsymbol{x}_i$ via $\boldsymbol{A}^\mathrm{T}\boldsymbol{x}_i$ but $\boldsymbol{X}$ contains the features as rows. Therefore the transformed feature matrix is $\boldsymbol{X}\boldsymbol{A}$)
>
> Now it is immediate to see that
>
> $$g(\boldsymbol{x}_i) = \boldsymbol{v}^{*\mathrm{T}}\boldsymbol{\Phi}(\boldsymbol{x}_i) = \boldsymbol{w}^{*\mathrm{T}}\boldsymbol{A}^{-\mathrm{T}}\boldsymbol{A}\boldsymbol{x}_i = \boldsymbol{w}^{*\mathrm{T}}\boldsymbol{x}_i = f(\boldsymbol{x}_i).$$

b) Can Bob's procedure lead to a lower training set error than Alice's if the matrix $\boldsymbol{A}$ is not invertible? Explain your answer.

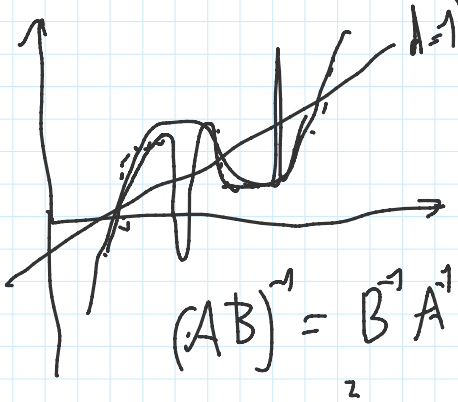
> Any weights $\boldsymbol{v}^*$ Bob finds are also feasible for Alice by letting $\boldsymbol{w} = \boldsymbol{A}\boldsymbol{v}^*$. Therefore Bob can only access a subset of the parameter space and cannot achieve a lower loss value than Alice. (It could still be equal but it cannot be better.)
>
> Note that we are only talking about training error in this example, not test error. Bob might manage to find a model that generalizes better than Alice's, but Alice will always be able to fit the training data at least as well as Bob.

$x_1, \ldots x_n \qquad y_1, \ldots, y_n$

find f s.t. $f(x_i) \approx y_i$

and f generalizes to new x

$f(x) = w^T \Phi(x)$ $\qquad \Phi(x) = \begin{pmatrix} x^0 \\ x^1 \\ \vdots \\ x^d \end{pmatrix}$



$(AB)^{-1} = B^{-1}A^{-1}$

$E_{LS} = \frac{1}{2}\sum_i (y_i - f(x_i))^2$

$= (Xw - y)^T (Xw - y) \frac{1}{2}$

$w^* = X^\dagger y = (X^T X)^{-1} X^T y$

$\Phi = \begin{pmatrix} x_1^T A \\ \vdots \\ \end{pmatrix}$

$v^* = (XA)^\dagger y$

$w = A v^*$

$g(x) = v^{*T} A^T x$ $\qquad f(x) = w^T x = v^{*T} A^T x$