Machine Learning (CE717-2)

Prof. Soleymani

Assignment #2, Due on Mehr 30th

Nima Mohammadi

nima.mohammadi@ut.ac.ir

1 Error Analysis

(a) **False**. Less complex model generally does not help with high training error.

(b) **True**. The transformation with nth order polynomial basis function results in higher dimensions which benefits more from larger number of training samples, compared to RBF.

(c) **True**. Using more training data can be helpful in case of high variance by reducing the effect of over-fitting, but it does not help with high bias and a high-bias algorithm attain nearly the same error with larger training set. More samples degrade the speed of the algorithm and not really help with bias.

(d) **False**. It's not necessarily true as providing more training samples can diminish this effect.

(e) **False**. Overfitting is a famous counter-example where training error is minimized, though validation error is still increased due to less generalization capability of the model.

2 Linear Regression

2.1. We replace the sum with matrix multiplication and the cost function would be as below:

$$
\begin{aligned}
J(w) &= (Xw - y)^T (Xw - y) \\
J(w) &= ((Xw)^T - y^T)(Xw - y) \\
J(w) &= (Xw)^T Xw - (Xw)^T y - y^T (Xw) + y^T y \\
J(w) &= w^T X^T Xw - 2(Xw)^T y + y^T y
\end{aligned} \tag{1}
$$

Taking the partial derivative of the cost function $J(w)$ with respect to w, we then determine the minimum by setting the derivative to zero:

$$
\begin{aligned}
&\frac{\partial J(w)}{\partial w} = 2X^T Xw - 2X^T y = 0 \\
&\Rightarrow 2x^T Xw = 2X^T y \\
&\rightarrow \hat{w} = \frac{2X^T y}{2X^T X} = (X^T X)^{-1} X^T y
\end{aligned} \tag{2}
$$

2.2. First in case the number of features is high, the operation of taking the inverse of the matrix is very slow. So provided $X_{n\times n}$, calculating $(X^T X)^{-1}$ is of $O(n^3)$. So we may opt to use the gradient descent method instead of the closed form. Second problem occurs when the intermediate matrix, for which we are to find the inverse, is non-invertible. We may still use pseudo-inverse.

2.3. In case one or more rows, or "features", can be expressed as a linear combination of some other rows, then the determinant will be zero and the matrix is called singular or ill-conditioned. In this case we can remove those (redundant) features, hopefully resulting in an invertible matrix. We may also use the pseudo-inverse instead of the inverse of matrix. Also when the inverse of $X^T X$ does not exist, gradient descent may be preferred.

2.4. We have autocorrelation matrix $R = E_x\left[xx^T\right]$ and correlation vector $c = E_{xy}\left[xy\right]$, then the optimal w^* would be:

$$
\begin{aligned}
E_x[xx^T]w^* &= E_{xy}[xy] \\
Rw^* &= c \\
w^* &= R^{-1}c
\end{aligned}
$$

The cost function would be:

$$
J(w) = E[y^2] - 2c^T w + w^T Rw
$$

Using the factorization

$$
w^T Rw - 2c^T w = (Rw - c)^T R^{-1}(Rw - c) - c^T R^{-1} c
$$

Then the cost function can be decomposed to

$$J(w) = \left[E[y^2] - c^T R^{-1} c\right] + \left[(Rw - c)^T R^{-1} (Rw - c)\right]$$

where the left term is for structural error and the right term is for the approximation error.

2.5.

$$J(w) = \sum_{i=1}^{n} \left(y^{(i)} - w^T x^{(i)}\right)^2 + \frac{\lambda}{2} \|w\|^2$$

$$\begin{aligned}
\frac{\partial J(w)}{\partial w} &= 0 \\
\frac{\partial}{\partial w}(\left\|y - Xw\right\|^2 + \frac{\lambda}{2} \|w\|^2) &= 0 \\
\frac{\partial}{\partial w}(y - Xw)^T (y - Xw) + \frac{\lambda}{2} w^T w &= 0 \\
-X^T y + X^T X w + \lambda w &= 0 \\
-X^T y + (X^T X + \lambda I) w &= 0 \\
(X^T X + \lambda I) w &= X^T y \\
w &= (X^T X + \lambda I)^{-} 1 X^T y
\end{aligned}$$

---

3 Nonlinear Regression

3.1. It is linear with respect to α.

3.2. There is no suitable change of variable as parameters are in the exponents.

3.3. According to laws of logarithms, the transformation as below:

$$y = \log(x_1^{\alpha_1} x_2^{\alpha_2}) = \alpha_1 \log(x_1) + \alpha_2 \log(x_2) + \epsilon$$

And the change of variables:

$$\begin{aligned}
z_1 &= \log(x_1) \\
z_2 &= \log(x_2)
\end{aligned}$$

---

4 Unrestricted Regression

4.1.

$$E_{x,y}[(y-h(x)^2)] = \int\int (y-h(x)^2)p(x,y)dxdy$$
$$= \int 2yp(x,y)dy - \int 2h(x)p(x,y)dy = 0$$

$$h^*(x) = \frac{\int yp(x,y)dy}{\int p(x,y)dy} = \int \frac{yp(x,y)}{p(x)}dy = \int yp(y|x)dy = E_{y|x}[y]$$

4.2.

$$E_{x,y}[|h(x)-y|] = \int\int |h(x)-y|p(x,y)dxdy$$

$$\frac{\partial E_{x,y}[|h(x)-y|]}{\partial h(x)} = \frac{\partial}{\partial x}\left(\int_x\int_{-\infty}^{h(x)}(h(x)-y)p(x,y)dydx + \int_x\int_{-h(x)}^{+\infty}(y-h(x))p(x,y)dydx\right)$$
$$= \int_x\int_{-\infty}^{h(x)} p(x,y)dydx + \int_x\int_{-h(x)}^{+\infty} p(x,y)dydx = 0$$

And by the second axiom of probability, we know that the probability of the entire sample space is equal to one. This along the equality above suggest h to be the median.