Machine Learning

Tutorial – Numeric Prediction

BITS F464 / BITS C464

September 11, 2016

Q1

Show that $MSE = variance + bias^2$

(MSE: mean-squared-error)

A1

We know that $var[X] = E[X^2] - [E[X]]^2$

Let $X = V - \theta$; so, we can write the above equation as
$var[V - \theta] = E[(V - \theta)^2] - (E[V - \theta])^2$

According to rule of variance, $var[V - \theta] = var[V]$ (as θ is constant)

Again on the R.H.S. we have
$E[(V - \theta)^2] = MSE$
and $E[V - \theta] = bias$

So, $variance = MSE - bias^2$

Q2

To find a linear model $Y = a + bX$ that minimizes MSE given points (x_1, y_1), (x_2, y_2), ..., (x_n, y_n); we first want to find the gradients $\frac{\partial MSE}{\partial b}$ and $\frac{\partial MSE}{\partial a}$. Find out the expression.

A2

When we fit a linear model, we have the equation $y_i = a + bx_i + e_i$; where e_i is the error in estimation of y_i given some x_i

MSE can be written in terms of this error e_i as

$$MSE = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n} \{y_i - (a + bx_i)\}^2 \tag{1}$$

If we partially differentiate eq.(1) w.r.t. a, we get

$$\frac{\partial MSE}{\partial a} = \frac{-2}{n}\sum_{i=1}^{n}\{y_i - (a + bx_i)\} \tag{2}$$

Similarly, if we partially differentiate eq.(1) w.r.t. b, we get

$$\frac{\partial MSE}{\partial b} = \frac{-2}{n}\sum_{i=1}^{n}x_i\{y_i - (a + bx_i)\} \tag{3}$$

Note that n is the number of samples.

Q3

We want to fit a linear model $Y = a + bX$. Derive the expressions for a and b that do this.

To obtain the expression for a and b, we need to equate eq.(2) and eq.(3) (refer the answer to Q2) to 0 respectively.

So, the expression for a can be written as

$$a = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{b}{n}\sum_{i=1}^{n} x_i = \overline{Y} - b\overline{X} \tag{4}$$

Similarly, the expression for b is

$$b = \frac{\sum_{i=1}^{n} y_i x_i - \frac{1}{n}\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2} \tag{5}$$

We can make eq.(4) by replacing b with its corresponding expression in eq.(5), and then simplifying the final expression for a.

Q4

We want to fit a linear model $Y = a + bX$ by minimising mean-square error. We know that a minimum occurs when $\nabla_a = 0$ and $\nabla_b = 0$. Show, by setting $\nabla_a = 0$ that the point $(\overline{X}, \overline{Y})$ lies on the regression line (that is, $\overline{Y} = a + b\overline{X}$)

A4

This has already been obtained in the expression for a (see eq.(4) in answer to Q3)

$$a = \overline{Y} - b\overline{X}$$

Hence, the data point $(\overline{X}, \overline{Y})$ lies on the regression line.

Q5

We want to fit a linear model $Y = a + bX$ by finding a and b using gradient descent. Write the iterative update equations for a and b in terms of the gradients.

A5

The update equations for b and a are

$$b_{k+1} = b_k - \eta \nabla_b \tag{6}$$

$$a_{k+1} = a_k - \eta \nabla_a \tag{7}$$

where, η is the learning rate.

Q6

What changes with gradient descent we want to fit a non-linear model $Y = a + bX + cX^2$?

A6

Algorithmically, nothing. We just have to find the extra gradient ∇_c and use the corresponding update equation for c as given below.

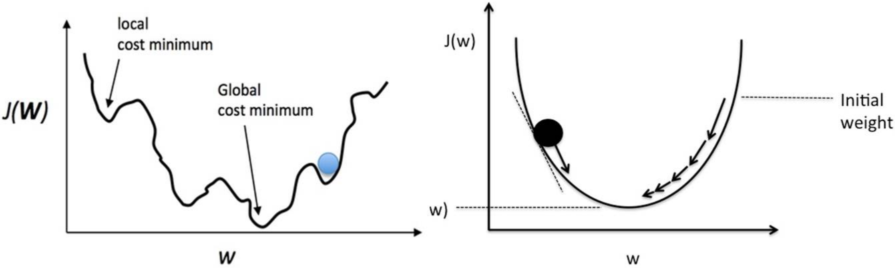
$$c_{k+1} = c_k - \eta \nabla_c \tag{8}$$

Q7

What happens the the cost function being minimised has multiple local minima?

A7

Gradient descent can get stuck in a local minimum that can be far away from the global minimum. Random restarts will not provably fix this, although it may help.



For convex functions (right-hand side figure) there is a unique local minimum and gradient descent will find this.

Q8

Is the cost function being minimised in least-squares regression convex?

A8

Yes. For the least-square regression, the cost function be always convex; so that the gradient descent reaches the unique local minimum (i.e. global minimum).

Q9

We want to fit a linear model $Y = a + bX$. For the special case that the errors $e_i \sim_{\mathrm{i}.i.d.} N(0, \sigma^2)$ show that the least-square estimate for b is the same as the maximum likelihood estimate for b.

A9

Since each e_i of the error vector **e** is uncorrelated with every other e_i and normally distributed with zero mean and the same variance of σ^2, we can write the likelihood function as

$$L = \frac{1}{(2\pi\sigma^2)^{n/2}} exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(y_i - a - bx_i)^2\right) \tag{9}$$

Taking natural logarithm of L gives us

$$ln\ L = -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - a - bx_i)^2 \tag{10}$$

[Continued on next slide.]

Minimizing L will lead to equating the partial differential of above equation w.r.t. b to zero. Which achives the following expression:

$$-\frac{1}{2\sigma^2}2\sum_{i=1}^{n}(y_i - a - bx_i)\frac{\partial}{\partial b}(y_i - a - bx_i) = 0 \tag{11}$$

After simplifying the above equation we get

$$\sum_{i=1}^{n} xi(y_i - a - bx_i) = 0 \tag{12}$$

This is same expression for b which we had obtained for LSE.

Q10

Derive equations for gradient descent when a regularisation term is added to the usual MSE cost function.

A10

Let consider the regularisation term be a function of the parameter (say θ) which we are updating using gradient descent. So, the cost function will look like

$$J(\theta) = MSE + \lambda f(\theta) \tag{13}$$

where, λ is a constant.

Partially differentiate the cost function w.r.t. θ,

$$\frac{\partial J}{\partial \theta} = \frac{\partial MSE}{\partial \theta} + \lambda \frac{\partial f(\theta)}{\partial \theta} \tag{14}$$