17s1: COMP9417 Machine Learning and Data Mining

Lectures: Linear Models for Regression
Topic: Questions from lecture topics
Version: with answers
Last revision: Mon Mar 20 00:03:45 AEDT 2017

Introduction

Some questions and exercises from the second week's course lectures, covering aspects of learning linear models (models "linear in the parameters") for regression, i.e., numeric prediction, tasks.

Question 1 A *univariate linear regression model* is a linear equation $y = a + bx$. Learning such a model requires fitting it to a sample of training data so as to minimize the error function $E(a, b) = \sum_{i=1}^{n}(y_i - (a + bx_i))^2$. To find the best parameters a and b that minimize this error function we need to find the error *gradients* $\frac{\partial E(a,b)}{\partial a}$ and $\frac{\partial E(a,b)}{\partial b}$. So derive these expressions by taking partial derivatives, divide by n (the number of (x, y) data points in the training sample) set them to zero, and solve for a and b.

Answer

In the following note that $\bar{z}$ denotes the mean of z.

$$
\begin{aligned}
\frac{\partial E(a,b)}{\partial a} &= \frac{\partial}{\partial a}\sum_{i=1}^{n}(y_i - (a+bx_i))^2 \\
&= \sum_{i=1}^{n} 2(y_i - (a+bx_i))(\frac{\partial}{\partial a}y_i - \frac{\partial}{\partial a}(a+bx_i) \\
&= \sum_{i=1}^{n} 2(y_i - (a+bx_i)) - (\frac{\partial}{\partial a}a + \frac{\partial}{\partial a}bx_i) \\
&= \sum_{i=1}^{n} 2(y_i - (a+bx_i)) - 1 \\
&= -2\sum_{i=1}^{n}(y_i - (a+bx_i)) \\
0 &= \frac{-2}{n}(\sum_{i=1}^{n}(y_i - (a+bx_i))) \\
0 &= \frac{1}{n}(\sum_{i=1}^{n}(y_i - (a+bx_i))) \\
a &= \frac{1}{n}\sum_{i=1}^{n} y_i - b\frac{1}{n}\sum_{i=1}^{n} x_i \\
a &= \bar{y} - b\bar{x}
\end{aligned}
$$

Similarly, the expression for b is as follows (the full expansion once the derivative is found can be skipped).

$$
\begin{aligned}
\frac{\partial E(a,b)}{\partial b} &= \frac{\partial}{\partial b}\sum_{i=1}^{n}(y_i - (a+bx_i))^2 \\
&= \sum_{i=1}^{n} 2(y_i - (a+bx_i))(\frac{\partial}{\partial b}y_i - \frac{\partial}{\partial b}(a+bx_i) \\
&= \sum_{i=1}^{n} 2(y_i - (a+bx_i)) - (\frac{\partial}{\partial b}a + \frac{\partial}{\partial b}bx_i) \\
&= -2\sum_{i=1}^{n}(y_i - (a+bx_i))x_i \\
b &= \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} y_1 \sum_{i=1}^{n} x_i}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\frac{1}{n}\sum_{i=1}^{n} x_i)^2}
\end{aligned}
$$

Looking at this, the denominator is recognisable as the variance of x, and the whole expression turns out to be: $\frac{\text{covariance of } x \text{ and } y}{\text{variance of } x}$.

Question 2 A linear regression model is represented by the linear equation $y = a + bx$. Show that the mean point $(\bar{x}, \bar{y})$ must line on the regression line.

Answer
In Question 1 we derived the expression $a = \bar{y} - b\bar{x}$. Suppose we ask the question: what is the value of the output y for the linear regression model when the input is $\bar{x}$, i.e., the mean of x ?
We formulate the regression model like this, then substitute the expression for a:

$$\begin{aligned} y &= a + b\bar{x} \\ &= \bar{y} - b\bar{x} + b\bar{x} \\ &= \bar{y} \end{aligned}$$

which completes the proof.

Question 3 *Mean Square Error*, or MSE, of an estimator such as a regression model can be decomposed. Show that MSE = (variance) + $(\text{bias})^2$.

Background Suppose we have a repeatable setting in which a regression model is trained and has its predictions tested to determine error. Let f (standing for $f(x)$) be the actual value and y (standing for $\hat{y}$) be the predicted value. Then, to simplify notation, the MSE can be written as follows:

$$\text{MSE} = E[f - y]^2$$

You will need some properties of the expected value, or expectation operator $E[\,]$ (information on these properties is available on the web, e.g., `http://en.wikipedia.org/wiki/Expected_value`).

The key point about the expectation is that it provides a way to describe characteristics of the distribution of possible values of random variables (or functions of random variables). The most well-known is the mean μ, or expected value, of a random variable, which is defined as the sum of *all* possible values of the random variable, weighted by the probability of that value occurring:

$$E[X] = \sum_{i=1}^{N} x_i \, p(x_i) = \mu$$

Analogously, the expected value of a function of one (or more) random variable(s) is the sum of the possible values of the function for each outcome, weighted by the probability of that outcome occurring. For example, $E[Y] = E[f(X)]$, for the function $Y = f(X)$:

$$E[f(X)] = \sum_{i=1}^{N} f(x_i) \, p(x_i)$$

Here are some properties following from this definition.

$$
\begin{aligned}
E[X+c] &= E[X]+c && \text{for } c \text{ a constant} \\
E[X+Y] &= E[X]+E[Y] && \text{where } X \text{ and } Y \text{ are random variables defined on the same probability space} \\
E[aX] &= aE[X] && \text{for } a \text{ not a random variable}
\end{aligned}
$$

These can be used, for example, to decompose the expectation of a linear model $E[a+bX] = a+bE[X]$. Some further properties of the expectation you may need to use:

$$
\begin{aligned}
E[fE[y]] &= fE[y] && \text{for } f \text{ a deterministic function, } y \text{ a random variable} \\
E[E[y]] &= E[y] \\
E[E[y]^2] &= E[y]^2 \\
E[yf] &= fE[y] && \text{again, since } f \text{ is deterministic, and } y \text{ a random variable} \\
E[yE[y]] &= E[y]^2 && \text{by the definition of expectation.}
\end{aligned}
$$

Answer

Suppose we have a repeatable setting in which a regression model is trained and has its predictions tested to determine error. Let f (standing for $f(x)$) be the actual value and y (standing for $\hat{y}$) be the predicted value. Then the MSE can be decomposed as follows:

$$\begin{aligned}
\text{MSE} &= E[f-y]^2 \\
&= E[(f-E[y]+E[y]-y)^2] \\
&= E[(f-E[y]+E[y]-y)(f-E[y]+E[y]-y)] \\
&= E[f^2 - fE[y] + fE[y] - fy \\
&\quad - fE[y] + E[y]^2 - E[y]^2 + yE[y] \\
&\quad + fE[y] - E[y]^2 + E[y]^2 - yE[y] \\
&\quad - yf + yE[y] - yE[y] + y^2] \\
&= E[y]^2 - 2fE[y] + f^2 + E[y^2 - 2yE[y] + E[y]^2] + \\
&\quad E[fE[y] + fE[y] - fy - fy - E[y]^2 - E[y]^2 + yE[y] + yE[y]]
\end{aligned}$$

We can see that this derivation starts with a small “trick”, by adding and subtracting the expectation $E[y]$. Then follows the rather tedious expansion of the terms. By re-arranging we find the expressions for the Bias^2 and Variance.

$$\begin{aligned}
\text{Variance} &= E[(y-E[y])^2] \\
&= E[y^2 - 2yE[y] + E[y]^2]
\end{aligned}$$

and

$$\begin{aligned}
\text{Bias}^2 &= (E[y]-f)^2 \\
&= E[y]^2 - 2fE[y] + f^2
\end{aligned}$$

However, we still have something left over at the end:

$$E[2fE[y] - 2fy - 2E[y]^2 + 2yE[y]]$$

which can be written as:

$$2E[fE[y] - fy - E[y]^2 + yE[y]]$$

and then:

$$2(E[fE[y]] - E[fy] - E[E[y]^2] + E[yE[y]])$$

How do we get rid of this ? If a term in the expectation is deterministic it can be brought out. So we get $2E[fE[y] - fy - E[y]^2]$, which expands to

$$2[E[fE[y]] - E[E[y]^2] - E[yf] + E[yE[y]]]$$

Now $E[fE[y]] = fE[y]$, since f is deterministic. Also, since $E[E[z]] = E[z]$ we get $E[E[y]^2] = E[y]^2$. Lastly, $E[yf] = fE[y]$ since f is deterministic, and $E[yE[y]] = E[y]^2$, by the definition of expectation. Thus

$$\begin{aligned}
&2[fE[y] - fE[y] - E[y]^2 + E[y]^2] \\
&= 2(0) \\
&= 0
\end{aligned}$$