Solutions Exercises Pattern Recognition 2018

1 Linear Regression

(a)

$$\boldsymbol{X}^\top \boldsymbol{X} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 31 & 25 & 27 & 23 & 32 & 22 & 29 \end{bmatrix} \begin{bmatrix} 1 & 31 \\ 1 & 25 \\ 1 & 27 \\ 1 & 23 \\ 1 & 32 \\ 1 & 22 \\ 1 & 29 \end{bmatrix} = \begin{bmatrix} 7 & 189 \\ 189 & 5193 \end{bmatrix}$$

$$\boldsymbol{X}^\top \boldsymbol{t} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 31 & 25 & 27 & 23 & 32 & 22 & 29 \end{bmatrix} \begin{bmatrix} 80 \\ 105 \\ 120 \\ 105 \\ 70 \\ 120 \\ 100 \end{bmatrix} = \begin{bmatrix} 700 \\ 18540 \end{bmatrix}$$

$$\boldsymbol{w} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{t} = \frac{1}{630} \begin{bmatrix} 5193 & -189 \\ -189 & 7 \end{bmatrix} \begin{bmatrix} 700 \\ 18540 \end{bmatrix} = \frac{1}{630} \begin{bmatrix} 131040 \\ -2520 \end{bmatrix} = \begin{bmatrix} 208 \\ -4 \end{bmatrix}$$

So the fitted model is

$$y(x) = 208 - 4x$$

(b) $w_0 = 208$: this is the expected performance at a temperature of 0 degrees. This doesn't make any sense: the model is only supposed to hold for temperatures between 20 and 35 degrees.

$w_1 = -4$: this is the change in expected performance when the temperature increases with one degree.

(c)

$$y(x = 20) = 208 - 4 \times 20 = 128.$$

(d) Some bookkeeping:

n	x_n	t_n	y_n	$t_n - y_n$	$(t_n - y_n)^2$	$t_n - \bar{t}$	$(t_n - \bar{t})^2$
1	31	80	84	-4	16	-20	400
2	25	105	108	-3	9	5	25
3	27	120	100	20	400	20	400
4	23	105	116	-11	121	5	25
5	32	70	80	-10	100	-30	900
6	22	120	120	0	0	20	400
7	29	100	92	8	64	0	0
$\sum$	189	700	700	0	710	0	2150

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{710}{2150} \approx 0.67.$$

2 Linear Models for Classification

(a) `whtvict` and `stranger` ($\alpha = 0.05$) In addition: `aggcirc` and `multstab` ($\alpha = 0.1$)

(b) The fitted probability is $-0.18679 - 0.08692 = -0.27371$. Negative probabilities are not possible according to the axioms of probability. This highlights a shortcoming of the linear probability model.

(c) 0.35639

(d) It appears that black defendants have a lower probability of getting the death penalty, but the coefficient of `blkdef` is not significantly different from zero (p-value: 0.43) at any conventional significance level. On the other hand, if you kill a white person, you have a higher probability of getting the death penalty, and the coefficient of `whtvict` is significant (p-value: 0.013) at $\alpha = 0.05$. One could argue that this is also a form of racial discrimination.

(e) The fitted probability is

$$\hat{p}(t = 1|\boldsymbol{x}) = (1 + e^{-\boldsymbol{w}_{\text{ML}}^\top \boldsymbol{x}})^{-1} = \left(1 + e^{3.5675+0.5308}\right)^{-1} = 0.0166$$

(f) The fitted response function is given by

$$\hat{p}(t = 1|\boldsymbol{x}) = (1 + e^{-\boldsymbol{w}_{\text{ML}}^\top \boldsymbol{x}})^{-1}.$$

Applying the chain rule twice, and noting that $\frac{d\,e^z}{d\,z} = e^z$, we get

$$\frac{\partial \hat{p}(t = 1|\boldsymbol{x})}{\partial x_i} = -(1 + e^{-\boldsymbol{w}_{\text{ML}}^\top \boldsymbol{x}})^{-2} \times e^{-\boldsymbol{w}_{\text{ML}}^\top \boldsymbol{x}} \times -w_i = w_i \times \frac{e^{-\boldsymbol{w}_{\text{ML}}^\top \boldsymbol{x}}}{(1 + e^{-\boldsymbol{w}_{\text{ML}}^\top \boldsymbol{x}})^2}$$

Hence we see that the marginal effect of an increase in x_i depends on the value of x_i and also on the value of the other variables. However, the quantity

$$\frac{e^{-\boldsymbol{w}_{\text{ML}}^\top \boldsymbol{x}}}{(1 + e^{-\boldsymbol{w}_{\text{ML}}^\top \boldsymbol{x}})^2}$$

is always positive, so the sign of the influence of an increase in x_i can be read from the sign of w_i. **Note**: in fact,

$$f(z) = \frac{e^z}{(1+e^z)^2}$$

is the probability density function of the standard logistic distribution, and the cumulative distribution function of the standard logistic distribution is given by

$$F(z) = \frac{e^z}{1+e^z},$$

which is the logistic response function (activation function or transfer function in neural network terminology).

3 Logistic Regression

Note: $\exp(x) \equiv e^x$. I use both notations interchangeably.

(a) Not surprising at all. Explanatory variable x denotes the additional time taken by public transport. The more additional time, the higher the probability that a person will take the car. This is exactly what the positive coefficient says.

(b) If traveling by car and public transport takes the same time ($x = 0$), then there is a preference for public transport, because

$$\frac{e^{-0.24}}{1+e^{-0.24}} \approx 0.44 < 0.5.$$

(c) Fill in $x = 30$:

$$\hat{p}(t = 1 \mid x = 30) = \frac{\exp(-0.24 + 0.053 \cdot 30)}{1 + \exp(-0.24 + 0.053 \cdot 30)} \approx 0.794$$

(d) The marginal effect of an increase in x is

$$\frac{\partial \hat{p}(t = 1|x)}{\partial x} = 0.053 \times \frac{e^{0.24-0.053x}}{(1 + e^{0.24-0.053x})^2}$$

For $x = 5$ this evaluates to 0.016, for $x = 30$ to 0.009. So an increase from 5 to 6 minutes time difference has a larger effect than an increase from 30 to 31 minutes time difference.

(e) If $-0.24 + 0.053x > 0$, predict that someone will take the car, otherwise predict public transport. Further simplification gives: if $x > 4.53$ then car, otherwise public transport. Since travel time is measured in whole minutes, an appropriate verbal description would be: *If, for a given person, travelling by public transport takes 5 minutes or more longer than travelling by car, predict that this person will take the car, otherwise predict that this person will take public transport.*

4 Optimization/Linear Regression

(a) The error function is:

$$E(w_0, w_1) = (4 - w_0 - w_1)^2 + (8 - w_0 - 2w_1)^2 + (6 - w_0 - 3w_1)^2$$

(b) The partial derivatives are:

$$\frac{\partial E}{\partial w_0} = -2(18 - 3w_0 - 6w_1)$$
$$\frac{\partial E}{\partial w_1} = -2(38 - 6w_0 - 14w_1)$$

We get two linear equations with two unknowns:

$$18 - 3w_0 - 6w_1 = 0 \tag{1}$$
$$38 - 6w_0 - 14w_1 = 0 \tag{2}$$

Solving for w_0 and w_1 we find: $w_0 = 4$, $w_1 = 1$. So $y(x) = 4 + x$.

(c) The second derivatives are:

$$\frac{\partial^2 E}{\partial w_0^2} = 6 \qquad \frac{\partial^2 E}{\partial w_1^2} = 28 \qquad \frac{\partial^2 E}{\partial w_0 \partial w_1} = 12$$

Putting these in the Hessian matrix we get

$$\mathbf{H} = \begin{bmatrix} 6 & 12 \\ 12 & 28 \end{bmatrix}$$

We find $\mathbf{H}_{11} = 6 > 0$ and $\det(\mathbf{H}) = 6 \cdot 28 - 12 \cdot 12 = 24 > 0$. Since both are positive, we conclude that $\mathbf{H}$ is positive definite. This means the point $(w_0 = 4, w_1 = 1)$ is a (local) minimum. If fact, since the Hessian matrix is positive definite everywhere (the second derivatives do not depend on the values of w_0 and w_1), the error function is globally convex (or concave up) so that $(w_0 = 4, w_1 = 1)$ is the unique global minimum.

Let's verify from "first principles" that the Hessian is positive definite. This means that

$$\mathbf{z}^\top \mathbf{H} \mathbf{z} > 0$$

for every $\mathbf{z} \neq \mathbf{0}$. Now

$$[z_1 \ \ z_2] \begin{bmatrix} 6 & 12 \\ 12 & 28 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = 6z_1^2 + 24z_1z_2 + 28z_2^2$$

If we divide this by 6, we get

$$z_1^2 + 4z_1z_2 + 4\frac{2}{3}z_2^2 = (z_1 + 2z_2)^2 + \frac{2}{3}z_2^2.$$

Since this is a sum of squares, it is bigger than zero unless both z_1 and z_2 are zero.

For those of you who are interested we discuss below why we have a local minimum if the Hessian matrix is positive definite. The treatment is copied from the book of Bishop and the equation numbers refer to the corresponding equations in Bishop. Consider a second-order Taylor expansion (quadratic approximation) of $E(\mathbf{w})$ around an arbitrary point $\widehat{\mathbf{w}}$

$$E(\mathbf{w}) \approx E(\widehat{\mathbf{w}}) + (\mathbf{w} - \widehat{\mathbf{w}})^\top \mathbf{b} + \frac{1}{2}(\mathbf{w} - \widehat{\mathbf{w}})^\top \mathbf{H}(\mathbf{w} - \widehat{\mathbf{w}}) \tag{5.28}$$

where

$$\mathbf{b} \equiv \nabla E|_{\mathbf{w}=\widehat{\mathbf{w}}} \tag{5.29}$$

is the gradient of $E(\mathbf{w})$ evaluated at $\mathbf{w} = \widehat{\mathbf{w}}$, and

$$\mathbf{H}_{ij} \equiv \left.\frac{\partial^2 E}{\partial w_i \partial w_j}\right|_{\mathbf{w}=\widehat{\mathbf{w}}} \tag{5.30}$$

is the so-called Hessian matrix of second order partial derivatives of $E(\mathbf{w})$, also evaluated at $\mathbf{w} = \widehat{\mathbf{w}}$.

Let $\mathbf{w}^\star$ be a *stationary point* of the error function, i.e. $\nabla E = \mathbf{0}$ at $\mathbf{w}^\star$.
Then (5.28) becomes

$$E(\mathbf{w}) \approx E(\mathbf{w}^\star) + \frac{1}{2}(\mathbf{w} - \mathbf{w}^\star)^\top \mathbf{H}(\mathbf{w} - \mathbf{w}^\star) \tag{5.32}$$

Hence, we have

$$E(\mathbf{w}) - E(\mathbf{w}^\star) \approx \frac{1}{2}(\mathbf{w} - \mathbf{w}^\star)^\top \mathbf{H}(\mathbf{w} - \mathbf{w}^\star)$$

Now, suppose

$$(\mathbf{w} - \mathbf{w}^\star)^\top \mathbf{H}(\mathbf{w} - \mathbf{w}^\star) > 0 \qquad \text{for all } (\mathbf{w} - \mathbf{w}^\star) \neq \mathbf{0}, \tag{5.37}$$

that is, $\mathbf{H}$ is positive definite.

Then we have $E(\mathbf{w}) - E(\mathbf{w}^\star) > 0$, that is, $E(\mathbf{w}) > E(\mathbf{w}^\star)$, so $\mathbf{w}^\star$ is a local minimum.

5 Optimization/Linear Regression

(a) The partial derivatives are:

$$\frac{\partial E}{\partial w_0} = -\sum_{n=1}^{N}(t_n - w_0 - w_1 x_n)$$
$$\frac{\partial E}{\partial w_1} = -\sum_{n=1}^{N}x_n(t_n - w_0 - w_1 x_n)$$

So we have:

$$\nabla E(\mathbf{w}) = \begin{bmatrix} \frac{\partial E}{\partial w_0} \\ \frac{\partial E}{\partial w_1} \end{bmatrix} = \begin{bmatrix} -\sum_{n=1}^{N}(t_n - w_0 - w_1 x_n) \\ -\sum_{n=1}^{N}x_n(t_n - w_0 - w_1 x_n) \end{bmatrix}$$

(b) For a single observation (t_n, x_n) the gradient is:

$$\nabla E_n(\mathbf{w}) = \begin{bmatrix} \frac{\partial E_n}{\partial w_0} \\ \frac{\partial E_n}{\partial w_1} \end{bmatrix} = \begin{bmatrix} -(t_n - w_0 - w_1 x_n) \\ -x_n(t_n - w_0 - w_1 x_n) \end{bmatrix}$$

For the given data point and weight vector $\mathbf{w}^{(0)}$ we get:

$$\nabla E_n(\mathbf{w}^{(0)}) = \begin{bmatrix} -(3 - 1.6 - 0.8 \times 3) \\ -3(3 - 1.6 - 0.8 \times 3) \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

With $\eta = 0.1$, the new weights become:

$$\mathbf{w}^{(1)} = \begin{bmatrix} 1.6 \\ 0.8 \end{bmatrix} - 0.1 \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 1.5 \\ 0.5 \end{bmatrix}$$

(c) With $\mathbf{w}^{(0)}$ the prediction for $x_n = 3$ was

$$y(x_n = 3) = 1.6 + 0.8 \times 3 = 4$$

So the squared prediction error for the data point is $(y(x_n) - t_n)^2 = (4-3)^2 = 1$. With the new weight vector the prediction is:

$$y(x_n = 3) = 1.5 + 0.5 \times 3 = 3$$

This gives a prediction error of zero which is obviously an improvement.

With $\eta = 0.2$ the new weight vector becomes:

$$\mathbf{w}^{(1)} = \begin{bmatrix} 1.6 \\ 0.8 \end{bmatrix} - 0.2 \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 1.4 \\ 0.2 \end{bmatrix}$$

The prediction becomes:

$$y(x_n = 3) = 1.4 + 0.2 \times 3 = 2$$

6 Support Vector Machines

(a) The support vectors are the attribute vectors with positive lagrange multiplier, so row 4, 6 and 7 in the data table:

$$\mathbf{x}_4 = \begin{bmatrix} 3 \\ 3 \end{bmatrix} \quad \mathbf{x}_6 = \begin{bmatrix} 4 \\ 6 \end{bmatrix} \quad \mathbf{x}_7 = \begin{bmatrix} 6 \\ 4 \end{bmatrix}$$

(b) To compute the value of the SVM bias term b, we use the formula

$$b = t_m - \sum_{n=1}^{N} a_n t_n \mathbf{x}_m^\top \mathbf{x}_n,$$

with any support vector, for example $\mathbf{x}_6 = [4 \;\; 6]^\top$. This yields:

$$b = 1 + \frac{1}{4}[4 \;\; 6] \begin{bmatrix} 3 \\ 3 \end{bmatrix} - \frac{1}{8}[4 \;\; 6] \begin{bmatrix} 4 \\ 6 \end{bmatrix} - \frac{1}{8}[4 \;\; 6] \begin{bmatrix} 6 \\ 4 \end{bmatrix} = -4$$

(c) To predict the class label for given attribute vectors, we use the formula

$$y(\mathbf{x}) = b + \sum_{n=1}^{N} a_n t_n \mathbf{x}^\top \mathbf{x}_n,$$

with $\mathbf{x} = [0 \;\; 7]^\top$. This yields:

$$y(\mathbf{x}) = -4 - \frac{1}{4}[0 \;\; 7] \begin{bmatrix} 3 \\ 3 \end{bmatrix} + \frac{1}{8}[0 \;\; 7] \begin{bmatrix} 4 \\ 6 \end{bmatrix} + \frac{1}{8}[0 \;\; 7] \begin{bmatrix} 6 \\ 4 \end{bmatrix} = -\frac{1}{2}$$

Since $y(\mathbf{x}) < 0$ we predict class -1.

(d) The weight vector is:

$$\mathbf{w} = \sum_{n=1}^{N} a_n t_n \mathbf{x}_n = -\frac{1}{4} \begin{bmatrix} 3 \\ 3 \end{bmatrix} + \frac{1}{8} \begin{bmatrix} 4 \\ 6 \end{bmatrix} + \frac{1}{8} \begin{bmatrix} 6 \\ 4 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}$$

The equation for the maximum margin decision boundary is:

$$\frac{1}{2}x_1 + \frac{1}{2}x_2 - 4 = 0$$