Machine Learning – Written Assignment 2

Philip Hartout

October 6, 2016

1. (a) The hypothesis function is given by

$$h_\theta(\vec{x}) = \sum_{i=1}^{m} \vec{\theta}^T \cdot \vec{x}^{(i)}$$

(b) The cost function using the explicit summation over all training examples is given by:

$$\frac{1}{2m}\sum_{i=0}^{m}\left[\left(\vec{\theta}^T \cdot \vec{x}^{(i)} - y^i\right)^2\right], \text{ where } \vec{y} = \begin{bmatrix} y^0 \\ \vdots \\ y^m \end{bmatrix}$$

(c) The vectorised expression for the gradient of the cost function is as follows:

$$\frac{\partial J(\theta)}{\partial \theta} = \begin{bmatrix} \frac{1}{m}\sum_{i=1}^{m}(\vec{\theta}^T \cdot \vec{x}^{(i)} - y^{(i)})x_0^{(i)} \\ \vdots \\ \frac{1}{m}\sum_{i=1}^{m}(\vec{\theta}^T \cdot \vec{x}^{(i)} - y^{(i)})x_n^{(i)} \end{bmatrix}$$

(d) The vectorised expression of the θ update rule in the gradient descent procedure is as follows:

$$\forall j, \quad \theta_j := \theta_j - \alpha\frac{1}{m}\sum_{i=1}^{m}(\vec{\theta}^T \cdot \vec{x}^{(i)} - y^{(i)})x_j^{(i)}$$

(e) We define the matrix X containing all the data vectors $x_1, x_2, \ldots, x_n$:

$$X = \begin{bmatrix} x_0^1 & \ldots & x_n^1 \\ \vdots & \ddots & \vdots \\ x_0^m & \ldots & x_n^m \end{bmatrix}$$

We can then deduce the following formula for the batch gradient descent in a fully vectorised form:

$$\vec{\theta} := \vec{\theta} - \alpha\frac{1}{m}X^T(X\vec{\theta} - \vec{y})$$

2.

3. (a) We estimate μ and σ using $\hat{\mu} = E(X) \approx 9.167$ and $\hat{\sigma}^2 = V(X) \approx 54.8$, so $X \sim \mathcal{N}(9.167, 54.8)$.

(b) $P(X \leq 20) \approx 0.928$.

(c) We calculate $f_{X_1,\ldots,X_6}(x_1, \ldots, x_6)$ as follows, given $X_1, \ldots, X_6$ are independent:

$$f_{X_1,\ldots,X_6}(x_1, \ldots, x_6) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) \cdot \ldots \cdot f_{X_6}(x_6).$$

Here, in particular:

$$\begin{aligned} f_{X_1,\ldots,X_6}(2,5,7,7,9,25) &= f_{X_1}(2) \cdot f_{X_2}(5) \cdot \ldots \cdot f_{X_6}(25) \\ &\iff = 0.166 \cdot 0.287 \cdot 0.385 \cdot 0.385 \cdot 0.491 \cdot 0.984 \\ &\iff \approx 3.412 \cdot 10^{-3}. \end{aligned}$$

(d) It is going to be smaller, because $f_{X_6}(25) = P(X_6 \leq 25)$ has a higher chance to occur, which increases the previous result significantly.

(e) $Cov(X, Y) \approx 17.56667$, based on the following data:

x	2	5	7	7	9	25
y	4	4	5	6	8	10

.

(f) The covariance is a number measuring the spread of the data around the mean (in squared units), while the MSE measures the vertical spread of the data around the regression line (in squared vertical units). So they are related because they are both measures of spread, but they don't refer to the same spread: the MSE refers to the spread with respect to the regression line whereas the covariance refers to the spread of the data with respect to the mean.

4.