Leren: Opdracht 1

Deadline: 30 Oktober 2014
Deze opdracht moet individueel gemaakt worden

1. Een internet winkelier gebruikt gegevens van klanten om te achterhalen wat voor klanten wat voor producten kopen. Hij zoekt klanten op op facebook en koppelt deze informatie aan de producten die ze kopen.
   (a) Wat is in dit geval een "trainingsvoorbeeld"?
   (b) Geef enkele voorbeelden van variabelen die het trainingsvoorbeeld beschrijven.
   Het gaat er hier om wat eigenlijk de data zouden kunnen zijn vor zo'n probleem. Deze bestaan uit gegevens over de klant en gegevens over producten die de klant wel (of juist niet) gekocht of bekeken heeft. Gegevens overt klanten kunnen zijn: leeftijd, geslacht, tijd waarop die actief is, dag van de week. Als postcode bekend is kan dat aangevuld worden met gegevens over woonruimte. Qua producten kan het gaan om specifieke producten of eigenschappen van producten (vers, bio, alcoholisch, ...). Een "voorbeeld" kan zijn een klant die iets koopt met alle eigenschappen van klant en product, of een klant met eigenschappen en alle producten met eigenschappen, of alleen producten die zijn gekocht in 1 koop of door 1 klant.
   (c) Geef een voorbeeld van een "supervised" en van een "unsupervised" leertaak die interessant kunnen zijn voor de winkelier. Hoe zou het resultaat van het leren eruit kunnen zien? Geef een voorbeeld ter illustratie.

2. Given the following data:

X	5	5	3
Y	6	6	10

   (a) Manually (using only a calculator) do linear regression. First, initialize the parameters θ_0 and θ_1 such that the regression function passes through the origin (0, 0) and has an angle of 45 degrees with the x-axis. Calculate the Mean Squared Error for this hypothesis.
   (b) Use a "learning rate" of 0.1 do one iteration of gradient descent. Give the intermediate results of your calculations, the new θ_0 and θ_1 and also compute the mean-squared error of the hypothesis after 1 iteration.
   (c) Do another iteration and again compute the Mean Squared Error. Is gradient descent having the intended effect?
   The line is characterized by the function $y = 0 + 1.x$ or $y = x$, so gradient descent is initialized with this. Predictions y for the three examples become equal to their x value and we get

X	*5*	*5*	*3*
Y	*6*	*6*	*10*
(h(x)	*5*	*5*	*3*
error	-1	-1	-7
squared error	*1*	*1*	*49*

   This gives a mean squared error of $51/3 = 17$ *(or if we divide by* $2m$ *instead of m: 8.5). The update rules for* θ_0 *and* θ_1 *are:*
   $\theta_0 = theta_0 - \alpha 1/m \Sigma_i (h(x_i) - y_i)$ *and*
   $\theta_0 = theta_0 - \alpha 1/m \Sigma_i (h(x_i) - y_i).x_i$
   We take $\alpha = 0.1$ *and we get:*
   $\theta_0 = 0 - (0.1.0.33. - 9) = \theta_0 - (0.1 - 3) = 0.3$
   $\theta_1 = 1 - (0.1.0.33.((-1.5) + (-1.5) + (-7.3)) = 1 - 0.033.(-5 - 5 - 21) = 1 - 0.033. - 31 = 1 + 1.03 = 2.03$
   Our new regression function $h(x)$ *becomes:* $y = 0.3 + 2.03x$. *With this we can calculate the MSE and make a new gradient descent step:*

X	*5*	*5*	*3*
Y	*6*	*6*	*10*
(h(x)	*10.5*	*10.5*	*6.4*
error	*4.5*	*4.5*	*3.6*
squared error	*20.3*	*20.3*	*10.2*

So the mean squared error is now 16.9.

3. Can we extend the gradient descent approach to finding univariate nonlinear, quadratic functions? E.g. can we apply the same approach to finding hypotheses of the form $y = \theta_0 + \theta_1 x + \theta_2 x^2$? If so, find the gradient and update rule. If not, explain the problem.

   The cost function $J(\theta)$*, the MSE divided by 2, now becomes* $\frac{1}{2m}\Sigma(\theta_0 + \theta_1 x + \theta_2 x^2 - y)^2$*. We find the gradient for* θ_1 *and* θ_2 *by taking the derivative. For* θ_2 *we get:*
   $\Sigma(\theta_0 + \theta_1 x + \theta_2 x_i^2 - y)^2)x^2$
   or: $\Sigma(\theta_0 x_i^2 + \theta_1 x_i^3 + \theta_2 x_i^4 - y_i x_i^2)$
   Note that this is a function with local minima.

4. Derive an equation that can be used to find the optimal value of the parameter θ_1 for linear regression from a set of data (x, y) directly, without doing gradient descent. This can be done by setting the value of the derivative of the cost for θ_1 equal to 0. Assume that the value of θ_0 is constant.

   MSE: $\frac{1}{m}(\Sigma_i^m(\theta_0 + \theta_1 x - y)^2 m)$ (We do not divide this by 2 here.)

   We take the partial derivative to θ_1. The derivative of a sum is equal to the sum of the derivatives of the terms in the sum, so

   $\frac{1}{m}(\Sigma_i^m 2(\theta_0 + \theta_1 x - y).x) = 0$

   solving for θ_1 and simplifying gives:

   $\theta_1 = \frac{1}{m}\frac{\Sigma xy - \theta_0 \Sigma x}{\Sigma x^2}$
   where summation runs over all examples.

   (Note: this is in fact a single variable version of the "normal equation". It can be generalized to n variables e.g. by expressing it in matrix notation.)