CS446: Machine Learning, Fall 2017, Homework 1

Name: Triveni Putti (tputti2)

Worked individually

Problem 2

Solution: (a) It is given in the question that

$$p(y|\mathbf{x}, \mathbf{w}) = Ber(y|sigm(\mathbf{w}^T\mathbf{x})).$$

Hence,

$$p(y = 1|\mathbf{x}, \mathbf{w}) = Ber(y = 1|sigm(\mathbf{w}^T\mathbf{x})).$$

By definition of Bernoulli distribution,

$$Ber(y|\theta) = \theta^y(1-\theta)^{1-y}$$

$$Ber(y = 1|\theta) = \theta^1(1-\theta)^{1-1} = \theta$$

Taking $\theta = sigm(\mathbf{w}^T\mathbf{x})$,

$$Ber(y = 1|sigm(\mathbf{w}^T\mathbf{x})) = sigm(\mathbf{w}^T\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}^T\mathbf{x}}}$$

Hence,

$$p(y = 1|\mathbf{x}, \mathbf{w}) = Ber(y = 1|sigm(\mathbf{w}^T\mathbf{x})) = \frac{1}{1 + e^{-\mathbf{w}^T\mathbf{x}}}.$$

$$p(y = 0|\mathbf{x}, \mathbf{w}) = 1 - p(y = 1|\mathbf{x}, \mathbf{w}) = 1 - \frac{1}{1 + e^{-\mathbf{w}^T\mathbf{x}}} = \frac{e^{-\mathbf{w}^T\mathbf{x}}}{1 + e^{-\mathbf{w}^T\mathbf{x}}}$$

Dividing numerator and denominator by $e^{-\mathbf{w}^T\mathbf{x}}$, we get

$$p(y = 0|\mathbf{x}, \mathbf{w}) = \frac{1}{1 + e^{\mathbf{w}^T\mathbf{x}}}.$$

Solution: (b) Derivative of Sigmoid function

$$\frac{d}{dz}sigm(z) = \frac{d}{dz}\frac{1}{1 + e^{-z}}$$

By quotient rule,

$$= \frac{0 * (1 + e^{-z}) - 1 * (-e^{-z})}{(1 + e^{-z})^2} = \frac{e^{-z}}{(1 + e^{-z})^2}$$

Solution: (c) Likelihood function of Logistic Regression for a data set D_n

$$= p(D_n|\theta_i)$$

Because we assume that independent sampling is done we can write this as,

$$= \prod_{i=1}^{n} p(y_i|\theta_i)$$

where θ_i is the set of parameters, which is $sigm(\mathbf{w}^T x_i)$ here and y_i are data points which can be equal to 0 or 1.

$$= \prod_{i=1}^{n} Ber(y_i|\theta_i)$$

$$= \prod_{i=1}^{n} \theta_i^{y_i}(1-\theta_i)^{1-y_i}$$

Hence the likelihood function of logistic regression is

$$= \prod_{i=1}^{n} \theta_i^{y_i}(1-\theta_i)^{1-y_i}$$

where $\theta_i = \frac{1}{1+e^{-\mathbf{w}^T x_i}}$ and $y_i = \{0, 1\}$

Solution: (d) Log Likelihood function
Taking logarithm of the final expression in (c), we get

$$log(\prod_{i=1}^{n} \theta_i^{y_i}(1-\theta_i)^{1-y_i})$$

$$= \sum_{i=1}^{n}(y_i log(\theta_i) + (1-y_i)log(1-\theta_i))$$

where $\theta_i = sigm(\mathbf{w}^T x_i)$. To find the update rule for gradient descent, we first take the gradient of the log likelihood expression i.e. $\nabla LL(\mathbf{w}_k)$

$$\nabla LL(\mathbf{w}_k) = \frac{d}{d\mathbf{w}_k} LL(\mathbf{w}_k)$$

Using chain rule, this can be written as

$$= \frac{d}{d\theta_i} LL(\mathbf{w}_k) * \frac{d\theta_i}{d\mathbf{w}_k}$$

Let us solve this in two parts. Computing the first part,

$$\frac{d}{d\theta_i} \sum_{i=1}^{n}(y_i log(\theta_i) + (1-y_i)log(1-\theta_i))$$

Since $y_i = 0$ or 1, it can be treated as a constant. On differentiating the inside expression, we get

$$\sum_{i=1}^{n}(y_i * \frac{1}{\theta_i} + (1-y_i) * \frac{1}{1-\theta_i} * -1)$$

Simplifying the expression and substituting $\theta_i = sigm(\mathbf{w}_k^T x_i)$ we get,

$$\sum_{i=1}^{n}(\frac{y_i}{sigm(\mathbf{w}_k^T x_i)} - \frac{(1-y_i)}{(1-sigm(\mathbf{w}_k^T x_i))})$$

Now, computing the second part

$$\frac{d\theta_i}{d\mathbf{w}_k} = \frac{d}{d\mathbf{w}_k} sigm(\mathbf{w}_k^T x_i) = \frac{d}{d\mathbf{w}_k}(\frac{1}{1+e^{-\mathbf{w}_k^T x_i}})$$

Using the result of part (b) in this question,

$$= \frac{e^{-\mathbf{w}_k^T x_i} x_i}{(1+e^{-\mathbf{w}_k^T x_i})^2} = x_i.sigm(\mathbf{w}_k^T x_i).(1 - sigm(\mathbf{w}_k^T x_i))$$

Since, $sigm(\mathbf{w}_k^T x_i) = \frac{1}{(1+e^{-\mathbf{w}_k^T x_i})}$ and $1 - sigm(\mathbf{w}_k^T x_i) = \frac{e^{-\mathbf{w}_k^T x_i} x_i}{(1+e^{-\mathbf{w}_k^T x_i})}$ Now, let's combine both the expressions into one.

$$\sum_{i=1}^{n}(\frac{y_i}{sigm(\mathbf{w}_k^T x_i)} - \frac{(1-y_i)}{(1-sigm(\mathbf{w}_k^T x_i))}).x_i.sigm(\mathbf{w}_k^T x_i).(1 - sigm(\mathbf{w}_k^T x_i)$$

$$= \frac{(y_i - sigm(\mathbf{w}_k^T x_i))}{sigm(\mathbf{w}_k^T x_i).(1 - sigm(\mathbf{w}_k^T x_i)}.x_i.sigm(\mathbf{w}_k^T x_i).(1 - sigm(\mathbf{w}_k^T x_i)$$

After cancelling some terms, we get

$$= (y_i - sigm(\mathbf{w}_k^T x_i)).x_i$$

Hence the update rule for Gradient Descent is

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \eta.(y_i - sigm(\mathbf{w}_k^T x_i)).x_i$$

where η is the stepsize of Gradient Descent.

References

Knuth, D. E., Larrabee, T. and Roberts, P. M. (1998). Mathematical writing .