Problem 1

Linear Regression [30 pts]

Suppose that, $y = w_0 + w_1x_1 + w_2x_2 + \epsilon, where\ \epsilon \sim N(0, \sigma^2)$

a) **[10 pts]** Write down an expression for $P(y|x_1, x_2)$.

$$P(y|x_1, x_2) = N(w_0 + w_1x_1 + w_2x_2 + \epsilon, \sigma^2) \ = \frac{1}{\sigma\sqrt{2\pi}} * e^{-\frac{1}{2}\left(\frac{y - w_0 + w_1x_1 + w_2x_2 + \epsilon}{\sigma}\right)^2}$$

b) **[10 pts]** Assume you are given a set of training observations $(x_1^{(i)}, x_2^{(i)}, y_1^{(i)})$ for $i = 1, \ldots, n$. Write down the conditional log-likelihood of this training data. Drop any constants that do not depend on the parameters w_0, w_1, or w_2.

$$l(w) = \prod_{i=1}^{n} p\left(y^{(i)}|x^{(i)}, w\right) = \sum_{i=1}^{n} \ln p\left(y^{(i)}\middle|x_1^{(i)}, x_2^{(i)}\right)$$

$$l(w) = \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y^i - w_0 - w_1x_1^i - w_2x_2^i\right)^2$$

If we drop any constants that do not depend on the parameters w_0, w_1, or w_2, we can rewrite $l(w)$ as,

$$l(w) = \frac{1}{2} \sum_{i=1}^{n} \left(y^i - w_0 - w_1x_1^i - w_2x_2^i\right)^2$$

c) **[10 pts]** Based on your answer, show that finding the MLE of that conditional log- likelihood is equivalent to minimizing least squares,

$$\frac{1}{2}\sum_{i=1}^{n}\left\{y^{(i)} - \left(w_0 + w_1x_1^{(i)} + w_2x_2^{(i)}\right)\right\}^2$$

$$= \ l(w|x, y) = -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{\left(y^i - w_0 - w_1x_1^i - w_2x_2^i\right)^2}{2}$$

$$= \frac{1}{2}\sum_{i=1}^{n}\left(y^i - w_0 - w_1x_1^i - w_2x_2^i\right)^2$$

Problem 2

Regularization [30 pts]

a) **[15 pts]** Find the partial derivative of the regularized least squares problem:

$$\frac{1}{2}\sum_{i=1}^{n}\left\{y^{(i)}-\left(w_0+w_1x_1^{(i)}+w_2x_2^{(i)}\right)\right\}^2+\frac{\lambda}{2}\|[w1,w2]\|_2^2$$

with respect to $w0$, $w1$, and $w2$. Although there is a closed-form solution to this problem, there are situations in practice where we solve this via gradient descent. Write down the gradient descent update rules for $w0$, $w1$, and $w2$.

Let us denote $h(w) = \frac{1}{2}\sum_{i=1}^{n}\frac{\partial h(w)}{\partial w_i}(y^i - w_0 - w_1x_1^i - w_2x_2^i)^2 + \frac{\lambda}{2}(\frac{\partial h(w)}{\partial w_i}(w_1^2 + w_2^2))$

The partial derivative of h(w) with respect to w_0 is:

$$\frac{\partial h(w)}{\partial w_0} = \frac{1}{2}\sum_{i=1}^{n}\frac{\partial h(w)}{\partial w_0}(y^i - w_0 - w_1x_1^i - w_2x_2^i)^2 + \frac{\lambda}{2}(\frac{\partial h(w)}{\partial w_0}(w_1^2 + w_2^2))$$

$$= \frac{1}{2}\sum_{i=1}^{n} -2(y^i - w_0 — w_1x_1^i - w_2x_2^i)$$

$$= -\sum_{i=1}^{n}(y^i - w_0 — w_1x_1^i - w_2x_2^i)$$

The partial derivative of h(w) with respect to w_1 is:

$$\frac{\partial h(w)}{\partial w_1} = \frac{1}{2}\sum_{i=1}^{n}\frac{\partial h(w)}{\partial w_1}(y^i - w_0 - w_1x_1^i - w_2x_2^i)^2 + \frac{\lambda}{2}(\frac{\partial h(w)}{\partial w_1}(w_1^2 + w_2^2))$$

$$= \frac{1}{2}\sum_{i=1}^{n} -2x_1^i(y^i - w_0 — w_1x_1^i - w_2x_2^i) + \frac{\lambda}{2}(2w_1)$$

$$= \sum_{i=1}^{n} x_1^i(y^i - w_0 — w_1x_1^i - w_2x_2^i) + (\lambda w_1)$$

The partial derivative of h(w) with respect to w_2 is:

$$\frac{\partial h(w)}{\partial w_2} = \frac{1}{2}\sum_{i=1}^{n}\frac{\partial h(w)}{\partial w_2}(y^i - w_0 - w_1x_1^i - w_2x_2^i)^2 + \frac{\lambda}{2}(\frac{\partial h(w)}{\partial w_2}(w_1^2 + w_2^2))$$

$$= \frac{1}{2}\sum_{i=1}^{n} -2x_2^i(y^i - w_0 — w_1x_1^i - w_2x_2^i) + \frac{\lambda}{2}(2w_2)$$

$$= \sum_{i=1}^{n} x_2^i(y^i - w_0 — w_1x_1^i - w_2x_2^i) + (\lambda w_2)$$

The gradient descent update rules for w_0 is:

$$w_0^{(i+1)} = w_0^{(i)} + \sum_{i=1}^{n} y^i - w_0 - w_1 x_1^i - w_2 x_2^i)$$

b) **[15 pts]** Suppose that $w_1, w_2 \sim N(0, \tau^2)$. Prove that, the MAP estimate of w_0, w_1, and w_2 with this prior is equivalent to minimizing the above regularized least squares problem with $\lambda = \frac{\sigma^2}{\tau^2}$.

Hint: Derive the equations for the two optimization problems and show they are equivalent.

$$= \frac{1}{2}\sum_{i=1}^{n}\left(y^i - w_0 - w_1 x_1^i - w_2 x_2^i\right)^2 + \frac{\sigma^2}{2\tau^2}(w_1^2 + w_2^2)$$

$$= argmax\ P(y|x, w)p(w) = logP(y|x, w) + \log P(w)$$

$$= argmin - logP(y|x, w) + \log P(w)$$

$$= -\frac{1}{2}\left(\frac{(y - w_0 - w_1 x_1 - w_2 x_2)^2}{\sigma^2} + \frac{w_1^2}{\tau^2} + \left(\frac{w_2}{\tau}\right)^2 + \frac{w_0^2}{\sigma^2}\right)$$

$$= (y - w_0 - w_1 x_1 - w_2 x_2)^2$$