CS 446: Machine Learning
Homework 2

Due on Tuesday, January 30, 2018, 11:59 a.m. Central Time

1. **[6 points]** Linear Regression Basics
   Consider a linear model of the form $\hat{y}^{(i)} = \mathbf{w}^\intercal \mathbf{x}^{(i)} + b$, where $\mathbf{w}, \mathbf{x} \in \mathbb{R}^K$ and $b \in \mathbb{R}$. Next, we are given a training dataset, $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}$ denoting the corresponding input-target example pairs.

   (a) What is the loss function, $\mathcal{L}$, for training a linear regression model? (Don't forget the $\frac{1}{2}$)

   Your answer: Let $N = |\mathcal{D}|$, then

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\left(\mathbf{w}^\intercal \dot{\mathbf{x}}^{(i)} + b - y^{(i)}\right)^2 = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\left(\hat{y}^{(i)} - y^{(i)}\right)^2$$

   (b) Compute $\frac{\partial \mathcal{L}}{\partial \hat{y}^{(i)}}$.

   Your answer: As before, let $N = |\mathcal{D}|$. Expanding the summation from the previous answer we have that

$$\begin{aligned}\mathcal{L} &= \frac{1}{2N}\left[\left(\mathbf{w}^\intercal \dot{\mathbf{x}}^{(0)} + b - y^{(0)}\right)^2 + \cdots + \left(\mathbf{w}^\intercal \dot{\mathbf{x}}^{(i)} + b - y^{(i)}\right)^2 + \cdots + \left(\mathbf{w}^\intercal \dot{\mathbf{x}}^{(N)} + b - y^{(N)}\right)^2\right] \\ &= \frac{1}{2N}\left[\left(\hat{y}^{(1)} - y^{(1)}\right)^2 + \cdots + \left(\hat{y}^{(i)} - y^{(i)}\right)^2 + \cdots + \left(\hat{y}^{(N)} - y^{(N)}\right)^2\right]\end{aligned}$$

   Note that when taking the derivative w.r.t $\hat{y}^{(i)}$ all the terms except the ith one can be treated as constants and, therefore, their derivative will be 0. Then

$$\frac{\partial \mathcal{L}}{\partial \hat{y}^{(i)}} = \frac{1}{2N}\left[0 + \cdots + 2\left(\hat{y}^{(i)} - y^{(i)}\right) + \cdots + 0\right] = \frac{1}{N}\left(\hat{y}^{(i)} - y^{(i)}\right)$$

   (c) Compute $\frac{\partial \hat{y}^{(i)}}{\partial \mathbf{w}_k}$, where $\mathbf{w_k}$ denotes the k^{th} element of $\mathbf{w}$.

   Your answer:
   Lets expand the dot product $\mathbf{w}^\intercal \mathbf{x}^{(i)}$, I will ignore the bias term since it doesn't have any effect when taking the derivative.

$$\hat{y}^{(i)} = \mathbf{w}^\intercal \mathbf{x}^{(i)} = \sum_{k=1}^{K} \mathbf{w}_k \mathbf{x}_k^{(i)} = \mathbf{w}_1 \mathbf{x}_0^{(i)} + \cdots + \mathbf{w}_k \mathbf{x}_k^{(i)} + \cdots + \mathbf{w}_K \mathbf{x}_K^{(i)}$$

   Then when taking the derivative most of the terms are treated as constants

$$\frac{\partial \hat{y}^{(i)}}{\partial \mathbf{w}_k} = \mathbf{x}_k^{(i)}$$

(d) Putting the previous parts together, what is $\nabla_{\mathbf{w}}\mathcal{L}$?

Your answer: As before, let $N = |\mathcal{D}|$. Now lets first look at what $\frac{\partial\mathcal{L}}{\partial\mathbf{w_k}}$ is

$$\frac{\partial\mathcal{L}}{\partial\mathbf{w_k}} = \sum_{i=1}^{N}\frac{\partial\mathcal{L}}{\partial\hat{y}^{(i)}}\times\frac{\partial\hat{y}^{(i)}}{\partial\mathbf{w_k}} = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{w}^{\intercal}\dot{\mathbf{x}}^{(i)} + b - y^{(i)}\right)\mathbf{x}_k^{(i)} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}^{(i)} - y^{(i)}\right)\mathbf{x}_k^{(i)}$$

Finally,

$$\nabla_{\mathbf{w}}\mathcal{L} = \left\langle\frac{\partial\mathcal{L}}{\partial\mathbf{w_0}},\ldots,\frac{\partial\mathcal{L}}{\partial\mathbf{w_K}}\right\rangle$$

(e) Compute $\frac{\partial\mathcal{L}}{\partial b}$.

Your answer:

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial b} &= \sum_{i=1}^{N}\frac{\partial\mathcal{L}}{\partial\hat{y}^{(i)}}\times\frac{\partial\hat{y}^{(i)}}{\partial b}\\ &= \sum_{i=1}^{N}\frac{1}{N}\left(\mathbf{w}^{\intercal}\dot{\mathbf{x}}^{(i)} + b - y^{(i)}\right)\times 1\\ &= \frac{1}{N}\sum_{i=1}^{N}(\hat{y}^{(i)} - y^{(i)})\end{aligned}$$

(f) For convenience, we group $\mathbf{w}$ and b together into $\mathbf{u}$, then we denote $\mathbf{z} = [\mathbf{x} \ \ 1]$. (*i.e.* $\hat{y} = \mathbf{u}^{\intercal}[x, 1] = \mathbf{w}^{\intercal}x + b$). What are the optimal parameters $\mathbf{u}^* = [\mathbf{w}^*, b^*]$? Use the notation $\mathbf{Z} \in \mathbb{R}^{|D|\times(K+1)}$ and $\mathbf{y} \in \mathbb{R}^{|D|}$ in the answer. Where, each row of $\mathbf{Z}, \mathbf{y}$ denotes an example input-target pair in the dataset.

Your answer:
Using the new notation

$$\mathbf{u}^* = (\mathbf{Z}^{\intercal}\mathbf{Z})^{-\mathbf{1}}\,\mathbf{Z}^{\intercal}\mathbf{y}$$

2. **[2 points]** Linear Regression Probabilistic Interpretation
Consider that the input $x^{(i)} \in \mathbb{R}$ and target variable $y^{(i)} \in \mathbb{R}$ to have to following relationship.

$$y^{(i)} = w \cdot x^{(i)} + \epsilon^{(i)}$$

where, ϵ is independently and identically distributed according to a Gaussian distribution with zero mean and unit variance.

(a) What is the conditional probability $p(y^{(i)}|x^{(i)}, w)$.

Your answer:

$$p\left(y^{(i)}|x^{(i)}, w\right) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(y^{(i)} - w\cdot x^{(i)}\right)^2\right)$$

(b) Given a dataset $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}$, what is the negative log likelihood of the dataset according to our model? (Simplify.)

Your answer:
The likelihood is

$$\prod_{i=1}^{N} p\left(y^{(i)}|x^{(i)}, w\right) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(y^{(i)} - w \cdot x^{(i)}\right)^2\right)$$

So when taking the negative log

$$\begin{aligned}
-\log\left(\prod_{i=1}^{N} p\left(y^{(i)}|x^{(i)}, w\right)\right) &= -\sum_{i=1}^{N} \log\left(p\left(y^{(i)}|x^{(i)}, w\right)\right) \\
&= -\sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(y^{(i)} - w \cdot x^{(i)}\right)^2\right)\right) \\
&= -\sum_{i=1}^{N} \left(\log\left(\frac{1}{\sqrt{2\pi}}\right) + \log\left(\exp\left(-\frac{1}{2}\left(y^{(i)} - w \cdot x^{(i)}\right)^2\right)\right)\right) \\
&= -N \log\left(\frac{1}{\sqrt{2\pi}}\right) - \sum_{i=1}^{N} \left(-\frac{1}{2}\left(y^{(i)} - w \cdot x^{(i)}\right)^2\right) \\
&= \sum_{i=1}^{N} \frac{1}{2}\left(y^{(i)} - w \cdot x^{(i)}\right)^2 - N \log\left(\frac{1}{\sqrt{2\pi}}\right) \\
&= \sum_{i=1}^{N} \frac{1}{2}\left(y^{(i)} - w \cdot x^{(i)}\right)^2 + \frac{N}{2} \log\left(2\pi\right) \blacksquare
\end{aligned}$$