University of Illinois Spring 2018

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, $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}$ 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 $\frac{1}{2}$)

Solution:
$\mathcal{L} = \frac{1}{2} \cdot \sum\limits_{(x^{(i)}, y^{(i)}) \in \mathcal{D}} (\hat{y}^{(i)} - y^{(i)})^2$

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

Solution: $(\hat{y}^{(i)} - y^{(i)})$

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

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

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

Solution: $\nabla_{\mathbf{w}} \mathcal{L} =$

$$\begin{bmatrix} \frac{\partial \mathcal{L}}{\partial \mathbf{w}_1} \\ \frac{\partial \mathcal{L}}{\partial \mathbf{w}_2} \\ \vdots \\ \frac{\partial \mathcal{L}}{\partial \mathbf{w}_K} \end{bmatrix}$$

Writing it out in terms of (b) and (c).

$$\begin{bmatrix} \sum\limits_{(\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}} (\hat{y}^{(i)} - y^{(i)}) \cdot \mathbf{x}_1^{(i)} \\ \sum\limits_{(\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}} (\hat{y}^{(i)} - y^{(i)}) \cdot \mathbf{x}_2^{(i)} \\ \vdots \\ \sum\limits_{(\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}} (\hat{y}^{(i)} - y^{(i)}) \cdot \mathbf{x}_K^{(i)} \end{bmatrix}$$

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

Solution:
$\sum\limits_{(x^{(i)}, y^{(i)}) \in \mathcal{D}} (\hat{y}^{(i)} - y^{(i)})$

(f) For convenience, we group $\mathbf{w}$ and b together into $\mathbf{u}$, then we denote $\mathbf{z} = [\mathbf{x} \quad 1]$. (*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.

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

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 $p(y^{(i)}|x^{(i)}, w)$.

   Solution:
   From the given assumption, $\epsilon^{(i)} = y^{(i)} - w \cdot x^{(i)}$ and $\epsilon^{(i)}$ is Gaussian distributed.
   Substitute $\epsilon^{(i)} = y^{(i)} - w \cdot x^{(i)}$ into the pdf of a zero-mean unit variance Gaussian distribution.
   $p(y^{(i)}|x^{(i)}, w) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(y^{(i)} - w \cdot x^{(i)})^2)$

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

   Solution: By definition of negative log likelihood.

$$L = -\log\left(\prod_{(x^{(i)},y^{(i)})\in\mathcal{D}} p(y^{(i)}|x^{(i)}, w)\right)$$

$$L = -\log\left(\prod_{(x^{(i)},y^{(i)})\in\mathcal{D}} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(y^{(i)} - w \cdot x^{(i)})^2)\right)$$

$$L = \frac{|\mathcal{D}|}{2}\log(2\pi) + \frac{1}{2}\sum_{(x^{(i)},y^{(i)})\in D}(y^{(i)} - w \cdot x^{(i)})^2$$