CSC321H5 Homework 2.

Deadline: Thursday, Jan. 23, by 9pm

Submission: You must submit your solutions as a PDF file through MarkUs. You can produce the file however you like (e.g. LaTeX, Microsoft Word, scanner), as long as it is readable.

Late Submission: Please see the syllabus for the late submission criteria.

Question 1. Loss Function

Suppose we have a prediction problem where the target t corresponds to an angle, measured in radians. A reasonable loss function we might use is:

$$L(y, t) = 1 - \cos(y - t)$$

Suppose we make predictions using a linear model $y = \mathbf{w}^T\mathbf{x} + b$.

(Update: Jan 19, please note change in point breakdown)

Part (a) – 2 pts

Compute $\frac{\partial L}{\partial y}$. Show your work.

Part (b) – 3 pts

Compute $\frac{\partial L}{\partial w_j}$. Show your work.

Part (c) – 2 pts

Compute $\frac{\partial \mathcal{E}}{\partial w_j}$, where $\mathcal{E} = \frac{1}{N}\sum_{i=1}^{N} L(y^{(i)}, t^{(i)})$. Show your work.

Part (d) – 6 pts

Derive a sequence of vectorized mathematical expressions for the gradients of the cost with respect to $\mathbf{w}$. As usual, the inputs are organized into a matrix $\mathbf{X}$ with one row per training example. The expressions should be something you can translate into a Python program without requiring a `for`-loop. Your answer should look like:

$$\mathbf{y} = \cdots$$
$$\frac{\partial \mathcal{E}}{\partial \mathbf{y}} = \cdots$$
$$\frac{\partial \mathcal{E}}{\partial \mathbf{w}} = \cdots$$

You can use $\sin(\mathbf{A})$ to denote the sin function applied elementwise to $\mathbf{A}$.

Question 2. Feature Maps

Suppose we have the following 1-D dataset for which we would like to build a binary classification model.

- $x^{(1)} = -1$, $t^{(1)} = 1$
- $x^{(2)} = 1$, $t^{(2)} = 0$
- $x^{(3)} = 3$, $t^{(3)} = 1$

Part (a) – 4 pts

Prove that this dataset is not linearly separable. That is, it is impossible to find values for w, b so that the model $y = \sigma(wx + b)$ correctly classifies all three data points.

(Hint: To be able to correctly classify these examples, we would need to have $-w + b > 0$, $w + b < 0$ and $3w + b > 0$.)

Part (b) – 3 pts

Now suppose we apply the following feature map

$$\phi(x) = \begin{pmatrix} \phi_1(x) \\ \phi_2(x) \end{pmatrix} = \begin{pmatrix} x \\ x^2 \end{pmatrix}$$

So that we instead use the model $y = \sigma(w_1\phi_1(x) + w_2\phi_2(x))$. Assume we have no bias term, so that the parameters are w_1 and w_2.

Write down the constraint on w_1 and w_2 corresponding to each training example, if we want w_1 and w_2 to correctly classify all three training examples.

Part (c) – 2 pts

Find a pair of values (w_1, w_2) for the model in part(b) by hand, so that the model correctly classifies all the examples. Remember that there is no bias term.

Question 3

Suppose we used a sigmoid activation function for a 1-D, binary classification model. That is, we use the following model:

$$\begin{aligned} z &= wx + b \\ y &= \sigma(z) \end{aligned}$$

Part (a) – 4 pts

Suppose we use the square loss $\mathcal{L}_{SE}(y, t) = \frac{1}{2}(y - t)^2$. Show that $\frac{\partial \mathcal{L}_{SE}}{\partial w} = (y - t)y(1 - y)x$.

Hint: Use chain rule $\frac{\partial \mathcal{L}}{\partial w} = \frac{\partial \mathcal{L}}{\partial y}\frac{dy}{dz}\frac{\partial z}{\partial w}$, and compute each derivative separately.

Part (b) – 4 pts

Suppose we use the cross-entropy loss $\mathcal{L}_{CE}(y, t) = -t\log(y) - (1 - t)\log(1 - y)$. Compute $\frac{\partial \mathcal{L}_{CE}}{\partial w}$.