Homework 1

Deadline: Monday, Jan. 16, at 11:59pm.

Submission: You must submit your solutions as a PDF file through MarkUs[1]. You can produce the file however you like (e.g. LaTeX, Microsoft Word, scanner), as long as it is readable.

Weekly homeworks are individual work. See the Course Information handout[2] for detailed policies.

1. **Regularized linear regression.** For this problem, we will use the linear regression model from lecture:

$$y = \sum_{j=1}^{D} w_j x_j + b$$

In lecture, we saw that regression models with too much capacity can overfit the training data and fail to generalize. One way to improve generalization, which we'll cover properly later in this course, is **regularization**: adding a term to the cost function which favors some explanations over others. For instance, we might prefer that weights not grow too large in magnitude. We can encourage them to stay small by adding a penalty

$$\mathcal{R}(\mathbf{w}) = \frac{\lambda}{2}\mathbf{w}^\top \mathbf{w} = \frac{\lambda}{2}\sum_{j=1}^{D} w_j^2$$

to the cost function, for some $\lambda > 0$. In other words,

$$\mathcal{E}_{\text{reg}} = \underbrace{\frac{1}{2N}\sum_{i=1}^{N}\left(y^{(i)} - t^{(i)}\right)^2}_{=\mathcal{E}} + \underbrace{\frac{\lambda}{2}\sum_{j=1}^{D} w_j^2}_{=\mathcal{R}},$$

where i indexes the data points and $\mathcal{E}$ is the same squared error cost function from lecture. Note that in this formulation, *there is no regularization penalty on the bias parameter.*

   (a) **[3 pts]** Determine the gradient descent update rules for the regularized cost function $\mathcal{E}_{\text{reg}}$. Your answer should have the form:

$$w_j \leftarrow \cdots$$
$$b \leftarrow \cdots$$

   This form of regularization is sometimes called "weight decay". Based on this update rule, why do you suppose that is?

   (b) **[3 pts]** It's also possible to solve the regularized regression problem directly by setting the partial derivatives equal to zero. In this part, for simplicity, we will drop the bias term from the model, so our model is:

$$y = \sum_{j=1}^{D} w_j x_j.$$

[1] `https://markus.teach.cs.toronto.edu/csc321-2017-01`

[2] `http://www.cs.toronto.edu/~rgrosse/courses/csc321_2017/syllabus.pdf`

In tutorial, and in Section 3.1 of the Lecture 2 notes, we derived a system of linear equations of the form

$$\frac{\partial \mathcal{E}}{\partial w_j} = \sum_{j'=1}^{D} A_{jj'} w_{j'} - c_j = 0.$$

It is possible to derive constraints of the same form for $\mathcal{E}_{\text{reg}}$. Determine the appropriate values for $A_{jj'}$ and c_j.

2. **Visualizing the cost function.** In lecture, we visualized the linear regression cost function in weight space and saw that the contours were ellipses. Let's work through a simple example of that. In particular, suppose we have a linear regression model with two weights and no bias term:

$$y = w_1 x_1 + w_2 x_2,$$

and the usual loss function $\mathcal{L}(y, t) = \frac{1}{2}(y-t)^2$ and cost $\mathcal{E}(w_1, w_2) = \frac{1}{N} \sum_i \mathcal{L}(y^{(i)}, t^{(i)})$. Suppose we have a training set consisting of $N = 3$ examples:

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

Let's sketch one of the contours.

(a) **[2pts]** Write the cost in the form

$$\mathcal{E} = c_1(w_1 - d_1)^2 + c_2(w_2 - d_2)^2 + \mathcal{E}_0.$$

(b) **[2pts]** Since $c_1, c_2 > 0$, this corresponds to an axis-aligned ellipse. Sketch the ellipse by hand for $\mathcal{E} = 1$. Label the center and radii of the ellipse. If you've forgotten how to plot axis-aligned ellipses, see Khan Academy[3].

[3] https://www.khanacademy.org/math/algebra-home/alg-conic-sections/alg-center-and-radii-of-an-ellipse/v/conic-sections-intro-to-ellipses