1) [CA session] Problem 1: Least-Squares Linear Regression

In last week’s module we studied the linear regression algorithm, which solves a regression problem using a linear predictor via optimizing the objective

$\text{TrainLoss}(w) = \frac{1}{|D_{\text{train}}|} \sum_{(x,y) \in D_{\text{train}}} (w \cdot \phi(x) - y)^2.$ \hspace{1cm} (1)

The training loss was minimized via gradient descent, which works iteratively to decrease the training loss. As mentioned in the module, we can actually solve for the optimal weights $w^*$ in closed-form. In this problem we will derive the normal equations used to solve for this estimator.
2) [CA session] Problem 2: Non-linear features

Consider the following two training datasets of \((x, y)\) pairs:

- \(D_1 = \{(-1, +1), (0, -1), (1, +1)\}\).
- \(D_2 = \{(-1, -1), (0, +1), (1, -1)\}\).

Observe that neither dataset is linearly separable if we use \(\phi(x) = x\), so let’s fix that. Define a two-dimensional feature function \(\phi(x)\) such that:

- There exists a weight vector \(w_1\) that classifies \(D_1\) perfectly (meaning that \(w_1 \cdot \phi(x) > 0\) if \(x\) is labeled +1 and \(w_1 \cdot \phi(x) < 0\) if \(x\) is labeled −1); and
- There exists a weight vector \(w_2\) that classifies \(D_2\) perfectly.

Note that the weight vectors can be different for the two datasets, but the features \(\phi(x)\) must be the same.

Some additional food for thought: Is every dataset linearly separable in some feature space? In other words, given pairs \((x_1, y_1), \ldots, (x_n, y_n)\), can we find a feature extractor \(\phi\) such that we can perfectly classify \((\phi(x_1), y_1), \ldots, (\phi(x_n), y_n)\) for some linear model \(w\)? If so, is this a good feature extractor to use?
3) [CA session] Problem 3: Backpropagation

Consider the following function

\[ \text{Loss}(x, y, z, w) = 2(xy + \max\{w, z\}) \]

Run the backpropagation algorithm to compute the four gradients (each with respect to one of the individual variables) at \( x = 3, \ y = -4, \ z = 2 \) and \( w = -1 \). Use the following nodes: addition, multiplication, max, multiplication by a constant.
4) [breakout, optional] Problem 4: Non-linear decision boundaries

Suppose we are performing classification where the input points are of the form \((x_1, x_2) \in \mathbb{R}^2\). We can choose any subset of the following set of features:

\[
\mathcal{F} = \left\{ x_1^2, x_2^2, x_1 x_2, x_1, x_2, \frac{1}{x_1}, \frac{1}{x_2}, 1, 1[x_1 \geq 0], 1[x_2 \geq 0] \right\}
\] (2)

For each subset of features \(F \subseteq \mathcal{F}\), let \(D(F)\) be the set of all decision boundaries corresponding to linear classifiers that use features \(F\).

For each of the following sets of decision boundaries \(E\), provide the minimal \(F\) such that \(D(F) \supseteq E\). If no such \(F\) exists, write ‘none’.

- \(E\) is all lines [CA hint]:

\[\text{______________________________} \] (3)

- \(E\) is all circles centered at the origin:

\[\text{______________________________} \] (4)

- \(E\) is all circles:

\[\text{______________________________} \] (5)

- \(E\) is all axis-aligned rectangles:

\[\text{______________________________} \] (6)

- \(E\) is all axis-aligned rectangles whose lower-right corner is at \((0, 0)\):

\[\text{______________________________} \] (7)
5) [breakout, optional] Problem 5: K-means

Consider doing ordinary $K$-means clustering with $K = 2$ clusters on the following set of 3 one-dimensional points:

$$\{-2,0,10\}.$$ \hspace{1cm} (8)

Recall that $K$-means can get stuck in local optima. Describe the precise conditions on the initialization $\mu_1 \in \mathbb{R}$ and $\mu_2 \in \mathbb{R}$ such that running $K$-means will yield the global optimum of the objective function. Notes:

- Assume that $\mu_1 < \mu_2$.
- Assume that if in step 1 of $K$-means, no points are assigned to some cluster $j$, then in step 2, that centroid $\mu_j$ is set to $\infty$.
- Hint: try running $K$-means from various initializations $\mu_1, \mu_2$ to get some intuition; for example, if we initialize $\mu_1 = 1$ and $\mu_2 = 9$, then we converge to $\mu_1 = -1$ and $\mu_2 = 10$. 