Boğaziçi University
Institute for Data Science & Artificial Intelligence

DSAI 512 Fall 2025
HW #1

Asst. Prof. Dr. Hüseyin Oktay ALTUN

Due: 20/10/2025

General submission information:

- Each homework must be submitted as a single `.pdf` file. Submissions in other formats (e.g., `.docx`, `.txt`) will not be graded. Handwritten solutions are acceptable if scanned clearly. If you prepare your homework in LaTeX and ensure a clean, well-formatted layout, you may earn a bonus of up to 20 points.
- If the homework includes both written and coding parts, two separate upload sections will appear on Moodle: one for the `.pdf` (theoretical part) and one for the `.ipynb` file (coding part). If the homework contains only written or only coding questions, only the relevant section will be available.
- Do not submit multiple files or a compressed (`.zip`) folder. Only a single, final file will be accepted.
- Name your file as `name_surname_hw1.pdf` or `name_surname_hw1.ipynb` (e.g., `oktay_altun_hw1.pdf`). Include your name and student ID clearly at the top of your file.
- **Academic integrity:** Cite all external sources you use (books, lecture notes, online materials, etc.). Plagiarism or unreferenced copying will result in a zero grade and potential disciplinary action.
- Upload your homework to Moodle before the deadline. Late submissions may not be accepted unless stated otherwise.
- If deemed necessary, the instructor of the course may conduct interviews related to the assignment and request explanations on how the questions were solved.

❗ If you have questions about the assignment, you may ask them in the class WhatsApp group. *If your question includes your solution method, do not post it publicly send it privately to the course assistant.*

Good luck, and submit on time!

1. **(30 points)**
   Consider a perceptron model in two dimensions defined as
   $$h(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^\top \mathbf{x})$$
   where $\mathbf{w} = [w_0, w_1, w_2]^\top$ and $\mathbf{x} = [1, x_1, x_2]^\top$. Although $\mathbf{x}$ has three coordinates, the first coordinate is fixed at 1, so the decision boundary lies on a two-dimensional plane.

   (a) Show that the regions on the (x_1, x_2) plane where $h(\mathbf{x}) = +1$ and $h(\mathbf{x}) = -1$ are separated by a straight line. Express this line as
   $$x_2 = ax_1 + b$$
   and derive the slope a and intercept b in terms of w_0, w_1, w_2.

   (b) Sketch the separating line for $\mathbf{w} = [2, 1, 4]^\top$ and for $\mathbf{w} = -[2, 1, 4]^\top$. Comment on whether the two lines and corresponding decision regions differ or coincide.

   In higher dimensions, the separating boundary generalizes from a line to a hyperplane.

2. **(70 points)**

   In the *update rule proof* covered in Lecture 1, we showed that a single update of the Perceptron Learning Algorithm (PLA)
   $$\mathbf{w}(t+1) = \mathbf{w}(t) + y(t)\mathbf{x}(t)$$
   moves the weight vector in the *right direction*: it increases the alignment between the weight vector $\mathbf{w}(t)$ and the misclassified point $\mathbf{x}(t)$. That local argument gave us geometric intuition for why the PLA works.

   Now, let us take the next conceptual step proving that the PLA *eventually converges* when the data are linearly separable. This proof may look intimidating at first, but if you follow it step by step, each part is straightforward. Throughout, assume that $\mathbf{w}(0) = \mathbf{0}$ and that the dataset is linearly separable.

   (a) **(Margin definition)** Let $\mathbf{w}^*$ denote an optimal weight vector that perfectly separates all data points, i.e., $y_n(\mathbf{w}^{*T}\mathbf{x}_n) > 0$ for all n. Define
   $$\rho = \min_{1 \le n \le N} y_n(\mathbf{w}^{*T}\mathbf{x}_n).$$
   Show that $\rho > 0$ and explain its geometric meaning.
   Toy check (draw & compute ρ): Consider the following 2D dataset (with $x_0 = 1$ fixed) and $\mathbf{w}^* = [2,\ 1,\ -2]^\top$:
   $$\begin{aligned} x_1 &= (-1,\ 3), & y_1 &= -1,\\ x_2 &= (-2,\ 1), & y_2 &= -1,\\ x_3 &= (1,\ 2), & y_3 &= -1,\\ x_4 &= (2,\ 1), & y_4 &= +1,\\ x_5 &= (-1,\ -1), & y_5 &= +1. \end{aligned}$$

(i) In the (x_1, x_2)-plane, *draw* these five points and the decision boundary induced by $\mathbf{w}^*$, clearly label $\mathbf{w}^*$ (as the normal to the boundary).

(ii) Compute the minimum margin ρ. Show your calculation steps briefly and mark on your plot the point(s) that achieve ρ.

(b) **(Growing alignment)** Show that for each update,

$$\mathbf{w}(t)^\top \mathbf{w}^* \geq \mathbf{w}(t-1)^\top \mathbf{w}^* + \rho.$$

Here, $\mathbf{w}(t)^\top \mathbf{w}^*$ measures how well the current weight vector aligns with the ideal separator, each update makes them more aligned. Then, using induction, conclude that

$$\mathbf{w}(t)^\top \mathbf{w}^* \geq t\rho.$$

(*Hint: use the PLA update rule and the fact that* $y(t)\mathbf{w}^{*T}\mathbf{x}(t) \geq \rho$.)

(c) **(Bound on the weight norm)** Show that

$$\|\mathbf{w}(t)\|^2 \leq \|\mathbf{w}(t-1)\|^2 + \|\mathbf{x}(t-1)\|^2.$$

(*Hint: expand* $\|\mathbf{w}(t)\|^2$ *using the update rule and recall that* $\mathbf{x}(t-1)$ *was misclassified by* $\mathbf{w}(t-1)$.)

(d) **(Bounding the growth rate)**

At this stage, we already know two things from the previous parts:

- The **alignment** with the true separator $\mathbf{w}^*$ grows linearly in t (part b).
- The **length** of the weight vector $\|\mathbf{w}(t)\|$ cannot explode in one step. It increases at most by $\|\mathbf{x}(t-1)\|^2$ each time (part c).

Now we want to understand how this length behaves *after many updates*. Even if it grows at most a little each time, after t updates, it might still accumulate. To capture that growth safely, we will use **induction**.

Let

$$R = \max_n \|\mathbf{x}_n\|$$

be the largest norm of any input vector in the dataset. Since every $\|\mathbf{x}(t-1)\| \leq R$, each update can increase $\|\mathbf{w}(t)\|^2$ by at most R^2.

Your task: use the inequality from part (c)

$$\|\mathbf{w}(t)\|^2 \leq \|\mathbf{w}(t-1)\|^2 + \|\mathbf{x}(t-1)\|^2$$

and apply it repeatedly (step by step, using induction) to show that after t updates, the total increase is bounded by the sum of these small increments.

Hint: Think of it this way each update adds at most R^2 to $\|\mathbf{w}\|^2$, and there are t updates. When you put that together, you should get a clean and simple expression for $\|\mathbf{w}(t)\|^2$ in terms of t and R.

(*Goal: derive a general bound on how fast* $\|\mathbf{w}(t)\|$ *can grow over time.*)

(e) **(Combining the results)** Combine (b) and (d) to show that

$$\frac{\mathbf{w}(t)^\top \mathbf{w}^*}{\|\mathbf{w}(t)\| \, \|\mathbf{w}^*\|} \geq \sqrt{t}\, \frac{\rho}{R\|\mathbf{w}^*\|}.$$

Recall that the left-hand side represents the cosine of the angle between $\mathbf{w}(t)$ and $\mathbf{w}^*$ and is therefore at most 1. Use this to derive an upper bound on the number of updates t before convergence.

(f) **(Final conclusion)** Conclude that the PLA converges after at most

$$t \leq \frac{R^2\|\mathbf{w}^*\|^2}{\rho^2}$$

updates. This bound is not tight in practice, but it guarantees convergence when the data are separable.

⭐ Don't be discouraged by the algebraic steps the main idea is simple: each update makes $\mathbf{w}(t)$ more aligned with $\mathbf{w}^*$, and since its growth is bounded, it must stop making mistakes eventually.