Applied Machine Learning EX 1 (10 pts, 5%)

due Saturday April 14 (submit pdf on Canvas; LaTeXing is recommended but not required)

1. In 2D, what are the x_1 and x_2 intercepts for the line $w_1 x_1 + w_2 x_2 + b = 0$? In n-dimensions, what are the x_i intercept for $\mathbf{w} \cdot \mathbf{x} + b = 0$? What's the distance from the origin to $\mathbf{w} \cdot \mathbf{x} + b = 0$?

2. What are your understanding of hyperplane, half-plane, and half-space in the context of linear classifiers?

3. The perceptron algorithm on slides assumed augmented space (implicit bias). State the perceptron algorithm with explicit bias, the corresponding linear separability condition and the convergence theorem.

4. State the exact definition of "linear separability": a dataset D is said to be linearly separable under a feature map $\mathbf{\Phi}$ (which converts every input $\mathbf{x}$ to a feature vector $\mathbf{\Phi}(\mathbf{x})$), if there exists $\mathbf{u} : \|\mathbf{u}\| = 1$ and $\delta > 0$, such that for every example $(\mathbf{x}, y) \in D$ where $y = \pm 1$, ______________________________.

5. Under the above definition, prove the perceptron converges regardless of initial weight vector.

6. Follow-up: is this new convergence bound worse than the original bound, R^2/δ^2? If so, does it imply that the perceptron always converges slower with a non-zero initial weight vector?

7. For each of the following, find a feature map $\mathbf{\Phi}$ that makes it linearly separable. Draw a picture for each.

   (a) $D = \{((0,2), +1), ((-2,0), +1), ((0,-2), +1), ((2,0), +1), ((0,0), -1)\}$

   (b) $D = \{((2,2), +1), ((1,1), -1)\}$

8. For the XOR data set, run perceptron for 8 iterations ($\mathbf{w}^{(0)} = \mathbf{0}$) and verify the perceptron cycling theorem.

9. For real-valued features (such as "lot size" in classifying whether this home is a "hot home" on `redfin.com`), we often transform each feature to be zero-mean and unit-variance. Geometrically, why this would help perceptron training?

10. If we extend the definition of $\mathbf{\Phi}$ to allow it to have access to the index i of each example $(\mathbf{x}^{(i)}, y^{(i)}) \in D$, so that $\mathbf{\Phi}$ maps $\mathbf{x}^{(i)}$ to $\mathbf{\Phi}(\mathbf{x}^{(i)}, i)$, then every dataset becomes (trivially) linearly separable. Why?

Debriefing (required):

1. Approximately how many hours did you spend on this assignment?
2. Would you rate it as easy, moderate, or difficult?
3. Did you work on it mostly alone, or mostly with other people?
4. How deeply do you feel you understand the material it covers (0%–100%)?
5. Any other comments?