Applied Machine Learning: EX (8%)

due Saturday April 20 (submit pdf on Canvas; LaTeXing is recommended but not required; graded by completeness, not correctness. Please watch all videos in this Unit and study the slides.)

1. This course uses "semi-blind" test sets. What other ML contests are like this? What's the advantage over non-blind test? What's a "truly blind" test set and how to use it? Do you know a contest doing that?

2. In 2D, what are the x_1 and x_2 intercepts for the line $w_1x_1 + w_2x_2 + b = 0$? 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 $\mathbf{w} \cdot \mathbf{x} + b = 0$?

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

4. The perceptron algorithm from the slides assumed augmented space (implicit bias). Describe the perceptron algorithm **with explicit bias**. Hint: read the Chapter in CIML, which uses explicit bias.

5. Now state the exact definition of "linear separability" with explicit bias: 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 $y = \pm 1$, _____.

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

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

8. For real-valued features (such as "lot size"), we often transform each feature to be zero-mean and unit-variance. Geometrically, why this would help perceptron training?

9. 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?

10. You want to build a spam filter with one feature per word, and the vocabulary size is V. Each email has at most n words, and there are $|D|$ training emails. You train T passes on D and make U updates in total ($U \ll T|D|$). Compare the running times of averaged perceptron in the naive and smart implementations.

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?