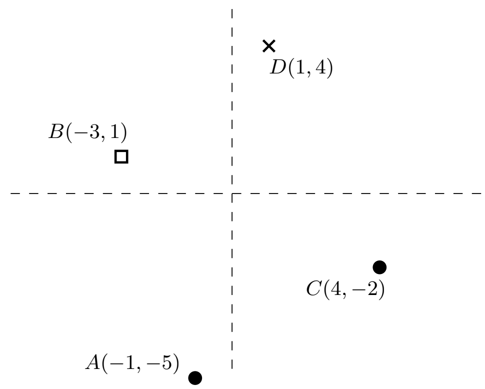
CMSC5724: Exercise List 7

Problem 1. Consider the training set P of points shown below:



where the two dots have label 1, the cross has label 2, and the box has label 3. Run multiclass Perceptron to find a generalized linear classifier to separate P.

Answer: At the beginning, $\vec{w_1} = \vec{w_2} = \vec{w_3} = [0, 0]$.
Round 1: Violation point D, $\ell = 2, z = 1$. Hence, $\vec{w_1} = [-1, -4], \vec{w_2} = [1, 4]$, $\vec{w_3} = [0, 0]$.
Round 2: Violation point B, $\ell = 3, z = 2$. Hence, $\vec{w_1} = [-1, -4], \vec{w_2} = [4, 3]$, $\vec{w_3} = [-3, 1]$.
Round 3: Violation point C, $\ell = 1, z = 2$. Hence, $\vec{w_1} = [3, -6], \vec{w_2} = [0, 5]$, $\vec{w_3} = [-3, 1]$.
No more violations.

Problem 2. Calculate the margin of the classifier you obtained in the previous problem.

Answer: Let W be the set of weight vectors obtained.

$$margin(A \mid W) = \min\left(\frac{\vec{w_1}\cdot\vec{A} - \vec{w_2}\cdot\vec{A}}{\sqrt{2\times\sum_1^3 |w_i|^2}}, \frac{\vec{w_1}\cdot\vec{A} - \vec{w_3}\cdot\vec{A}}{\sqrt{2\times\sum_1^3 |w_i|^2}}\right) = \min\left(\frac{27-(-25)}{\sqrt{2\times 80}}, \frac{27-(-2)}{\sqrt{2\times 80}}\right) = \frac{29}{\sqrt{2\times 80}}$$

Similarly,

$$margin(B \mid W) = \min\left(\frac{10-(-15)}{\sqrt{2\times 80}}, \frac{10-5}{\sqrt{2\times 80}}\right) = \frac{5}{\sqrt{2\times 80}}$$
$$margin(C \mid W) = \min\left(\frac{24-(-10)}{\sqrt{2\times 80}}, \frac{24-(-14)}{\sqrt{2\times 80}}\right) = \frac{34}{\sqrt{2\times 80}}$$
$$margin(D \mid W) = \min\left(\frac{20-(-21)}{\sqrt{2\times 80}}, \frac{20-1}{\sqrt{2\times 80}}\right) = \frac{19}{\sqrt{2\times 80}}$$

Therefore, the margin equals $\frac{5}{\sqrt{2\times 80}}$.

Problem 3. Suppose we run multiclass Perceptron on $k = 2$. Let $\{\vec{w_1}, \vec{w_2}\}$ be the set of weight vectors returned. Prove: $\vec{w_1} = -\vec{w_2}$.

Answer: It suffices to prove that $\vec{w_1} + \vec{w_2} = \vec{0}$ after every round. This obviously holds at the beginning because $\vec{w_1} = \vec{w_2} = \vec{0}$. Suppose that $\vec{w_1} + \vec{w_2} = \vec{0}$ before the next round starts. Let p be the violation point used in the round to do adjustments. Since we always add $\vec{p}$ to a weight vector but subtract $\vec{p}$ from the other weight vector, $\vec{w_1} + \vec{w_2}$ is still $\vec{0}$ at the end of the round.

Problem 4. Continuing on Problem 3, prove: the "margin" of $W = \{\vec{w_1}, \vec{w_2}\}$ as defined in multiclass Perceptorn is precisely the "margin" as defined in (the traditional) Perceptorn (i.e., the smallest distance from a point in the training set P to the separation plane).

Answer: It suffices to prove: for each point p in the training set, $margin(p \mid W)$ is precisely the distance from p to the separation plane.

Without loss of generality, assume that p is classified as class 1, i.e., $\vec{w_1} \cdot \vec{p} > \vec{w_2} \cdot \vec{p}$. We have:

$$\begin{aligned} margin(p \mid W) &= \frac{\vec{w_1} \cdot \vec{p} - \vec{w_2} \cdot \vec{p}}{\sqrt{2(|\vec{w_1}|^2 + |\vec{w_2}|^2)}} \\ &= \frac{2\vec{w_1} \cdot \vec{p}}{\sqrt{4|\vec{w_1}|^2}} \\ &= \frac{\vec{w_1} \cdot \vec{p}}{|\vec{w_1}|} \end{aligned}$$

which is the distance from p to the separation plane, as promised.

Problem 5 (Multi-Class Generalization Theorem). Let $\mathcal{X}$ be an instance space, $\mathcal{Y} = \{1, 2., ..., k\}$ be a label space, and $\mathcal{D}$ a distribution over $\mathcal{X} \times \mathcal{Y}$. Let S be a set of independent samples drawn from $\mathcal{D}$. A *classifier* h is a function $h : \mathcal{X} \to \mathcal{Y}$. For every such h, define

$$\begin{aligned} \text{er}(h) &= \Pr_{(x,y)\sim\mathcal{D}}[h(x) \neq y] \\ \text{er}_S(h) &= \frac{|\{(x,y) \in S \mid h(x) \neq y\}|}{|S|}. \end{aligned}$$

Let $\mathcal{H}$ be a finite set of classifiers. Fix a value δ such that $0 < \delta \leq 1$. Prove: with probability at least $1 - \delta$, we have the property that

$$\text{er}(h) \leq \text{er}_S(h) + \sqrt{\frac{\ln(1/\delta) + \ln|\mathcal{H}|}{2|S|}}$$

holds true for every $h \in \mathcal{H}$.

Answer: The proof is precisely the same as our proof for the generalization theorem presented in Lecture 1.