CS 3190: Foundations of Data Analysis **Due: Wednesday 4.15 at 2:30pm**

Homework 6: Clustering and Classification

Instructions: Your answers are due at 2:30pm. You must turn in a pdf through canvas I recommend using latex (`http://www.cs.utah.edu/~jeffp/teaching/latex/`) for producing the assignment answers. If the answers are too hard to read you will lose points, entire questions may be given a 0 (e.g. **sloppy pictures with your phone's camera are not ok, but very careful ones are**)

Please make sure your name appears at the top of the page.

You may discuss the concepts with your classmates, but write up the answers entirely on your own. **Be sure to show all the work involved in deriving your answers! If you just give a final answer without explanation, you may not receive credit for that question.**

---

1. **[40 points]** Consider this set of 3 sites: $S = \{s_1 = (-3,-1), s_2 = (1,1), s_3 = (-2,2)\} \subset \mathbb{R}^2$. We will consider the following 5 data points $X = \{x_1 = (-2,0), x_2 = (-2,1), x_3 = (-1,1), x_4 = (0,0), x_5 = (-3,-2)\}$.
   (a) Provide a simple plot of both point sets S and X with each type in a different color and marker style.

   For each of the following points compute the closest site (under Euclidean distance):
   (b) $\phi_S(x_1) =$
   (c) $\phi_S(x_2) =$
   (d) $\phi_S(x_3) =$
   (e) $\phi_S(x_4) =$
   (f) $\phi_S(x_5) =$

   Now consider that we have 3 Gaussian distributions defined with each site s_j as a center μ_j. The corresponding standard deviations are $\sigma_1^2 = 0.3$, $\sigma_2^2 = 1.0$ and $\sigma_3^2 = 1.0$, and we assume they are univariate so the covariance matrices are $\Sigma_j = \begin{bmatrix} \sigma_j^2 & 0 \\ 0 & \sigma_j^2 \end{bmatrix}$.

   (g) Write out the probability density function (its likelihood $f_j(x)$ for each of the Gaussians).

   Now we want to assign each x_i to each site in a soft assignment. For each site s_j define the weight of a point as $w_j(x) = f_j(x)/(\sum_{j=1}^{3} f_j(x))$. For each of the following points calculate the weight for each site

   (h) $w_1(x_1), w_2(x_1), w_3(x_1) =$
   (i) $w_1(x_2), w_2(x_2), w_3(x_2) =$
   (j) $w_1(x_3), w_2(x_3), w_3(x_3) =$
   (k) $w_1(x_4), w_2(x_4), w_3(x_4) =$
   (l) $w_1(x_5), w_2(x_5), w_3(x_5) =$

2. **[15 points]** Construct a data set X with $n = 4$ points in $\mathbb{R}^2$ and a set S of $k = 2$ sites so that Lloyd's algorithm will have converged, but there is another set S', of size $k = 2$, so that $\mathsf{cost}(X, S') < \mathsf{cost}(X, S)$.

   (a) Provide a single plot that contains your choice of X, S, and S' each in its own format.

   (b) Explain why S' is better than S, but that Lloyd"s algorithm will not move from S.

3. **[30 points]** Consider a family of linear classifiers defined by the sign of function $g_{w,b}(x) = \langle w, x \rangle + b$, where $x \in \mathbb{R}^2$ and so $w \in \mathbb{R}^2$ and $b \in \mathbb{R}$. Given a data point x_i and label $y_i \in \{-1, +1\}$. We require that $\|w\| = 1$.

   Now consider a uncertainty zone misclassification goal Λ (in place of Δ). In this setting, we want to penalize a classifier with a cost of 3/4 for any point within a distance of 2 of the classification boundary – even if it has the correct sign. So the cost is

$$\Lambda(g_{w,b}, (x_i, y_i)) = \begin{cases} 1 & \text{if } (x_i, y_i) \text{ is misclassified and } |g_{w,b}(x_i)| > 2 \\ 3/4 & \text{if } 0 \leq |g_{w,b}(x_i)| \leq 2 \\ 0 & \text{if } (x_i, y_i) \text{ is classified correctly and } |g_{w,b}(x_i)| > 2 \end{cases}$$

   (a) Explain $\Lambda(g_{w,b}, (x_i, y_i))$ as a function of $z_i = y_i g_{w,b}(x_i)$.

   (b) Design a loss function $\ell_\Lambda(z)$ as proxy for $\Lambda(z)$ that is (i) convex, (ii) has a derivative defined for all z, and (iii) for all values of z satisfies $\ell_\Lambda(z) \geq \Lambda(z)$. This part is asking for the equation $\ell_\Lambda(z)$.

   (c) Plot $\Lambda(z)$ and $\ell_\Lambda(z)$ with z on the x-axis, and the value of those loss functions represented on the y-axis. Both functions should be on the same plot.

4. **[15 points]**

   (a) Construct and report a set of labeled points (X, y) in $\mathbb{R}^2$ that is *not* linearly separable (provide a plot).

   (b) Explain what will happen if you run the perceptron algorithm for a linear classifier on this data set? (don't allow a fixed upper bound on T the number of steps)