CS 3190: Foundations of Data Analysis **Due: Friday 12.06 at 11:59pm**

Homework 5: Clustering and Classification

Instructions: Your answers are due at 11:59pm on the due date. 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: $S = \{s_1 = (0,0), s_2 = (3,4), s_3 = (-3,2)\} \subset \mathbb{R}^2$. We will consider the following 5 data points $X = \{x_1 = (1,3), x_2 = (-2,1), x_3 = (10,6), x_4 = (6,-3), x_5 = (-1,1)\}$.

   For each of the following points compute the closest site (under Euclidean distance):

   (a) $\phi_S(x_1) =$

   (b) $\phi_S(x_2) =$

   (c) $\phi_S(x_3) =$

   (d) $\phi_S(x_4) =$

   (e) $\phi_S(x_5) =$

   Now consider that we have 3 Gaussian distributions defined with each site s_j as a center μ_j. The corresponding standard deviations are $\sigma_1 = 2.0$, $\sigma_2 = 4.0$ and $\sigma_3 = 5$, and we assume they are univariate so the covariance matrices are $\Sigma_j = \begin{bmatrix} \sigma_j & 0 \\ 0 & \sigma_j \end{bmatrix}$.

   (f) 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

   (g) $w_1(x_1), w_2(x_1), w_3(x_1) =$

   (h) $w_1(x_2), w_2(x_2), w_3(x_2) =$

   (i) $w_1(x_3), w_2(x_3), w_3(x_3) =$

   (j) $w_1(x_4), w_2(x_4), w_3(x_4) =$

   (k) $w_1(x_5), w_2(x_5), w_3(x_5) =$

2. **[10 points]** Construct a data set X with 4 points in $\mathbb{R}^2$ and a set S of $k = 2$ sites so that Lloyds algorithm will have converged, but there is another set S', of size $k = 2$, so that $\textsf{cost}(X, S') < \textsf{cost}(X, S)$. Explain why S' is better than S, but that Lloyds algorithm will not move from S.

3. **[25 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 1/2 for any point within a distance of 2 of the classification boundary – even if it has the correct sign. So the cost is

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

   (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 $\ell_\Lambda(z) \geq \Lambda(z)$.

4. **[25 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)

   (c) Describe another algorithm discussed in the class (Chapters 9.1 - 9.3) which would provides a acceptable linear classifier for set of points.

Extra Questions

5. **[10 points]** Consider the quadratic (polynomial of degree 2) regression on a data set (X, y) where each of n data points (x_i, y_i) has $x_i \in \mathbb{R}^2$ and $y_i \in \mathbb{R}$. To simplify notation, let each $x_i = (a_i, b_i)$.

   (a) Expand $x_i = (a_i, b_i)$ and write the model $M_\alpha(x_i)$ as a single dot product of the form

$$M_\alpha(x_i) = \langle \alpha, (?, ?, \ldots, ?)\rangle$$

   where α is a vector, and you need to fill in the appropriate ?s.

   (b) Write the batch (of size n) gradient $\nabla f(\alpha)$ for this problem, where

$$f(\alpha) = \sum_{i=1}^{n} (M_\alpha(x_i) - y_i)^2.$$

   Your expression for $\nabla f(\alpha)$ should use the term $(M_\alpha(x_i) - y_i)$ as part of its solution.

6. **[15 points]** Consider a matrix $A \in \mathbb{R}^{n \times d}$ for $n > d$, and its SVD is $\mathsf{svd}(A) = [U, S, V^T]$. Let the left singular vectors be $u_1, u_2, \ldots, u_n$, the right singular vectors $v_1, v_2, \ldots, v_d$, and the singular values $\sigma_1, \sigma_2, \ldots, \sigma_d$. Let A_k be the best rank-k approximation of A (we'll consider $k = 2$ and $k = 3$).

   (a) Using only the singular values (and mathematical operators), write

      i. $\|A_2\|_2^2 =$

      ii. $\|A_3\|_F^2 =$

   (b) Using only the elements of the SVD (i.e., the expression should not include A), write

      i. $A_3 =$

      ii. $A_3 - A_2 =$

   (c) Consider a point $x \in \mathbb{R}^d$. Using only x and the elements of the SVD, write an expression for $\pi_{A_3}(x)$; that is x projected onto the 3-dimensional subspace spanned by A_3.