Introduction to Machine Learning, Spring 2023
Homework 4
(Due Friday, May. 5 at 11:59pm (CST))

April 22, 2023

1. [20 points] Consider a dataset of n observations $\mathbf{X} \in \mathbb{R}^{n \times d}$, and our goal is to project the data onto a subspace having dimensionality p, $p < d$. Prove that PCA based on projected variance maximization is equivalent to PCA based on projected error (Euclidean error) minimization.

2. [30 points] Let's see how well you remember k-means clustering, also known as Lloyd's Algorithm. As usual, the input is a set of n sample points, $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n \in \mathbb{R}^d$ and an integer k. There are no input labels; we want to assign each sample point $\mathbf{X}_j$ a label $y_j \in \{1, 2, \cdots, k\}$, which can be interpreted as "$\mathbf{X}_j$ is assigned to cluster y_j." The means of the clusters are written $\mu_1, \mu_2, \cdots, \mu_k$.

   (a) What is the cost function that k-means clustering tries to minimize? Write it in terms of the $\mathbf{X}_j$'s, the y_j's, and the μ_j's. (Not the formula for the within-cluster variation, please; a formula that uses the means. We are flexible about what notation you use to sum the terms in the cost function; please explain it if it's not obvious.) [8 points]

   (b) Consider the step of the algorithm where the labels y_j are held fixed while the cluster means μ_j are updated. I asserted in lecture that it is easy to show with calculus that if we want to minimize the cost function, we should choose each μ_j to be the mean (centroid) of the sample points assigned to cluster i. Please do that calculus and show that this claim is correct. (Make sure you explain your notation for counting the points in a cluster.) Show your work and don't skip any steps of the derivation. [10 points]

   (c) Consider the step of the algorithm where the cluster means μ_j are held fixed while the labels y_j are updated. Sometimes a sample point $\mathbf{X}_j$ has several cluster means that are equally close. In class I said, "If there's a tie, and one of the choices is for $\mathbf{X}_j$ to stay in the same cluster as the previous iteration, always take that choice." What could conceivably go wrong if you don't follow that advice? [6 points]

   (d) In which of the following cases should you prefer k-nearest neighbors over k-means clustering? For all the four options, you have access to images $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n \in \mathbb{R}^d$ [6 points]

      A: You do not have access to labels. You want to find out if any of the images are very different from the rest, i.e., are outliers.

      B: You have access to labels $y_1, y_2, \cdots, y_n$ telling us whether image i is a cat or a dog. You want to find out whether the distribution of cats is unimodal or bimodal. You already know that the distribution of cats either has either one or two modes, but that's all you know about the distribution.

      C: You have access to labels $y_1, y_2, \cdots, y_n$ telling us whether image i is a cat or a dog. You want to find out whether a new image z is a cat or a dog.

      D: You have access to labels $y_1, y_2, \cdots, y_n$ telling us whether image i is a cat or a dog. Given a new image z, you want to approximate the posterior probability of z being a cat and the posterior probability of z being a dog.