Intro to Big Data Science — Spring 2023-2024

Name: ____________________ **ID No.:** ________________

Quiz 6 This worksheet MUST be handed in at the end of the class.

1. True or false:

   1) Kernel matrix is symmetric and positive semi-definite.
   2) The one-nearest-neighbor (1NN) classifier may do better than Bayes classifier.
   3) Both PCA and spectral clustering perform eigen-decomposition.
   4) For any two variables x and y having joint distribution $p(x, y)$, we always have $H[x, y] = H[x] + H[y]$ where H is entropy function.

2. Answer the questions in short words:

   1) In one sentence, characterize the differences between linear regression and logistic regression.

   2) In one sentence, characterize the difference between k-Nearest-Neighbors and k-Means.

   3) Assume that we are using a ridge regression with a tuning parameter λ in the penalty term. Sketch a graph showing two curves: training error vs. λ and test error vs. λ.

3. Multiple choice:

   1) Which is incorrect about missing value filling?
      (A) For non-numeric features, one can fill missing values with the mode of the data.
      (B) Filling with means or modes may reduce the variance of data.
      (C) Linear interpolation can be used for missing value filling.
      (D) One cannot use random values for filling.
   2) In the m-th iteration of K-Means, the mass centers are given by $(1, 2)$, $(-1, 3)$, $(6, 0)$. Based on this, which is correct about the assignment of the samples $(2, 4)$ and $(2, 0)$ in the $(m + 1)$-th iteration?
      (A) They are assigned in the same cluster with mass center $(1, 2)$.
      (B) They are assigned in the same cluster with mass center $(-1, 3)$.
      (C) $(2, 4)$ is in the cluster with mass center $(-1, 3)$, while $(2, 0)$ is in the cluster with mass center $(1, 2)$.
      (D) None of the above is correct.
   3) Which is incorrect about support vector machine (SVM)?
      (A) The objective of SVM is to minimize the margin around the separating plane.
      (B) The samples at the boundary of the margin are called support vectors.
      (C) SVM can be solved in dual formulation.
      (D) SVM can also be used to classify samples which are not linearly separable.

4. Consider a multivariate linear model $\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}$ with $\mathbf{y} \in \mathbb{R}^{n\times 1}$, $\mathbf{X} \in \mathbb{R}^{n\times d}$, $\mathbf{w} \in \mathbb{R}^{d\times 1}$, and $\boldsymbol{\epsilon} \in \mathbb{R}^{n\times 1}$, where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$, follows the normal distribution, where $\mathbf{I}_n$ is the $n \times n$ identity matrix. For a given data set $(\mathbf{X}, \mathbf{y})$, we want to use ridge regression with a tuning parameter $\lambda > 0$ to estimate $\mathbf{w}$.

   (a) Please write down the model as an optimization problem. Also show that $\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_d)^{-1}\mathbf{X}^T\mathbf{y}$.

   (b) Is $\hat{\mathbf{w}}$ unbiased, i.e., $\mathrm{E}\hat{\mathbf{w}} = \mathbf{w}$? Prove your result.

5. (a) (PCA) Given 3 data points in 2-d space, (0, 1), (1, 2) and (2, 3),

      i. What is the first principle component? (Hint: remember to do centralization.)

      ii. If we want to project the original data points into 1-d space by principle component you choose, what is the variance of the projected data?

   (b) The goal of Non-negative Matrix Factorization (NMF) is to reduce the dimensionality given non-negativity constraints. That is, we would like to find principle components $\mathbf{u}_1, \ldots, \mathbf{u}_r$, each of which is of dimension $d > r$, such that the d-dimensional data $\mathbf{x} \approx \sum_{i=1}^{r} z_i \mathbf{u}_i$, and all entries in $\mathbf{x}, \mathbf{z}, \mathbf{u}_{1:r}$ are non-negative. NMF tends to find sparse (usually small L_1 norm) basis vectors $\mathbf{u}_i$'s. Below is an example of applying PCA and NMF on a face image. Please point out the basis vectors in the equations and give them correct labels (NMF or PCA).