EE769 Introduction to Machine Learning

End-semester Examination (Suggested Solutions), 24th Apr, 2023

1. For PCA, the covariance matrix of the data is $C = \begin{bmatrix} 20 & 19 & 0 \\ 19 & 20 & 0 \\ 0 & 0 & 20 \end{bmatrix}$ whose singular value (eigen value) decomposition is $C = U\Lambda U^T = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 39 & 0 & 0 \\ 0 & 20 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \\ \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \end{bmatrix}$. Suggest the number of intrinsic dimensions that this data has, and the unit vectors along the second principal direction, with justification. [1+1=2]

Intrinsic dimension is 2, as two eigen values are much larger than the third one. [1] Looking for observation that two are much larger eigenvalues than the third. Second principal direction is [0 0 1]^T

2. To reduce dimension using t-SNE, we minimize the KL divergence between input data conditional density $\mathrm{p_{j|i}} = \frac{\exp\left(-\|\mathbf{x_i}-\mathbf{x_j}\|^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-\|\mathbf{x_i}-\mathbf{x_k}\|^2/2\sigma_i^2\right)}$, $\mathrm{p_{i|i}} = 0$ and output data conditional density $\mathrm{q_{j|i}} = \frac{\left(1+\|\mathbf{y_i}-\mathbf{y_j}\|^2\right)^{-1}}{\sum_{k\neq i}\left(1+\|\mathbf{y_i}-\mathbf{y_k}\|^2\right)^{-1}}$ using gradient descent to learn the output data matrix $\mathbf{Y}$. Suggest a way to choose σ_i^2 for each sample indexed with i, so that it adapts to the varying probability density in its neighborhood. [2]

We can choose the average distance of k-nearest neighbors (e.g. 5 nearest neighbors) as σ_i. This will be large for sparse regions and small for denser regions. Other methods that are directionally correct will be considered.

3. In DB-SCAN, which of the following is a / are likely effect / effects of increasing radius ε? Justify. [2.5]
   a) More points can become outliers [0.5] ***No, because more points will be connected and not remain outliers.***
   b) Clusters can merge [0.5] ***Yes, because more points will be connected, causing clusters to merge***
   c) Clusters can split [0.5] ***No, because more points will be connected, clusters will not split***
   d) More points can become core points [0.5] ***Yes, because more points will get min_pt neighbors***
   e) More points can become connected points [0.5] ***Yes/No, because more points will get min_pt neighbors***

A combined comprehensive explanation for all five parts is fine.

4. In k-means, if we replace the L2 distance in the objective $\min_{\mathbf{c_k}} \sum_{i=1}^{N}\sum_{k=1}^{K} 1_{k=\arg\min_j \|\mathbf{x_i}-\mathbf{c_j}\|_2^2} \|\mathbf{x_i}-\mathbf{c_k}\|_2^2$ with L1 distance, then show that the cluster prototype will change from the centroid (mean) to the mediod (median) of the cluster, and the preferred cluster shape will be a square instead of a circle for 2-dimensional data. [1.5+1 = 2.5] ***For a group of points belonging to a single cluster, if we equate derivative of sum of distance $d(x_i, c_k)$ with respect to centroid c_k to 0, for L2 distance squared, we get $\Sigma i\ (x_i,c_k) = 0$, which gives $c_k = \Sigma i\ (x_i)/N_k$, i.e., centroid. Similarly, if we take L1 distnace, we will get $\underline{\Sigma i\ sign(x_i - c_k) = 0}$, which gives medoid. [1.5] A point can switch clusters if its L1 disance is smaller to the mediod, which is constant at the boundary of a square centered at the mediod.***

5. The following training data arrives at an internal node of a decision tree during the training phase, where x_1 and x_2 are input data dimensions, and t is the target classification label. Suggest a decision criteria for the node (assuming a threshold classifier). [2]

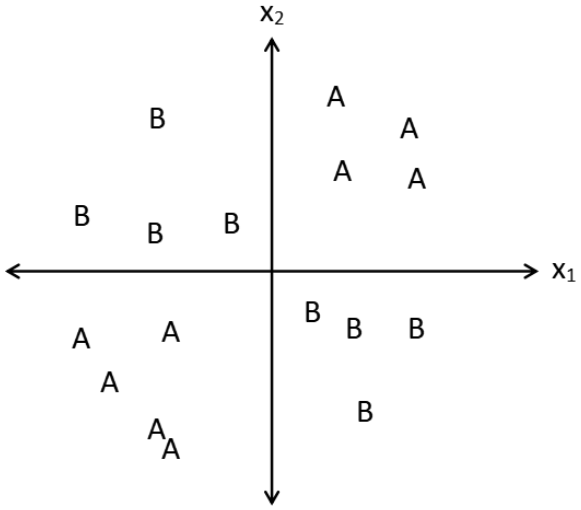
x_1	20	19	18	17	15	14	14	8	13	10	10	10	6	5	5	5	2
x_2	20	6	14	16	16	2	12	19	7	10	16	8	12	18	17	19	1
t	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1

Here, we can see that by either entropy of Gini index, x_1 separates the data better.
x_1 must be mentioned with justification, and some threshold between 7 and 14 must be mentioned.

6. Which of the following are true for a random forest classifier? Justify your answer:
   a) Second tree can only be trained after the first tree is finished training, and so forth. [1] ***No, because RF is an ensemble, each of the tree is trained independently.***
   b) Input data needs to be normalized so that it is zero mean and has unit variance. [1] ***No, because threshold classifiers within the trees and the nodes only depend on the order of the data.***

7. Suggest a way for weighted (unequal) voting of individual (presumably independent) regressors based on their training or validation performance. [1.5] ***We can weigh each regressor with its correlation coefficient of the predicted and true output and take weighted average. Other directionally correct schemes are also fine.***

8. Training data for a binary classification problem is shown in the figure to the right, where the approximate sample location in 2-d space is marked with its label A or B. Based on this figure, answer the following questions:



   a) Can the classification problem for the training data shown in the figure on the right be learned using a linear classifier? If so, then give the expression for the decision boundary. If not, then suggest a third feature that is dependent on the first two features that can be introduced to solve this problem using a linear boundary. [1.5] ***No, this is not linearly separable. Try $x_3 = x_1 . x_2$***
   b) Suggest weights and biases of a neural network with two hidden layers, having four neurons in the first hidden layer and two in the second hidden layer to classify this data (almost) correctly. Hint: Use the neural network of t he form $y = f(w_{10}.f(w_8.f(w_4.x_2) + w_7.f(w_3.x_1) + b_2) + w_9.f(w_6.f(w_2.x_2) + w_5.f(w_1.x_1) + b_1) + b_3)$, where f is the sigmoid activation function. Here, the first hidden layer separates data along the axes, second hidden layer selects quadrants, output layer combines 1st OR 3rd quadrant. [3]

$\mathbf{y = f(k.f(k.f(-k.x_2) + k.f(-k.x_1) - 1.5k) + k.f(k.f(k.x_2) + k.f(k.x_1) - 1.5k) - 0.5k);\ k >> 0}$

Other neural networks are also possible. Looking for quadrant detection and OR logic etc.

9. In a soft SVM with RBF kernel $k(x_i, x_j) = \exp\left(-a\|x_i - x_j\|^2\right)$ will increasing the hyperparameter a increase or decrease regularization? Justify with conceptual diagrams of 2-D space. [1.5] ***If we increase a, we will make the kernel width smaller, leading to more jagged decision boundaries, which leads to lower regularization.***

10. For a convolutional neural network, an image feature map of size 100x100x10 (10 being number of channels) is input into a convolutional layer of a neural network with filters (kernels) of size 3x3, whose output has 20 channels. This is followed by a ReLU activation, and a 2x2 max pool layer. What are the number of weights and biases of the following layers: [2]
    a) Convolutional layer ***10x3x3x20 + 20 = 1800 + 20 = 1820 [1]***
    b) ReLU activation ***0 [0.5]***
    c) Max pooling ***0 [0.5]***

11. How is a discrete variable representing four classes {cat, dog, rabbit, parrot} can be coded when it is:
    a) An input to a model? [1] ***$[0\ 0\ 0]^T$, $[0\ 0\ 1]^T$, $[0\ 1\ 0]^T$, $[1\ 0\ 0]^T$***
    b) An output of a model? [1] ***$[0\ 0\ 0\ 1]^T$, $[0\ 0\ 1\ 0]^T$, $[0\ 1\ 0\ 0]^T$, $[1\ 0\ 0\ 0]^T$***

12. Metrics for binary classification include accuracy, F1-score, and AUC.
    a) Which of these is sensitive to the decision threshold (which is usually taken to be 0.5 for a probability-based classifier, or 0 for a linear expression)? Explain. [1.5] ***Accuracy and F1 score require a threshold***
    b) For p(class=0) = 0.999, which of these will be very high for a classifier that always gives class=0 as its output? Justify. [1] ***Accuracy will be 99.9%, (F1 score will be based on both precision and recall; AUC will consider all thresholds)***

13. Assume that the total cost of treating a cancer patient is 2, erroneously treating a healthy person is 5, not treating a cancer patient is 10, and not treating a healthy person is 0. Determine the total cost assuming 10000 people were tested using a machine learning classifier of which 100 truly had cancer, true positive rate was 90%, and false positive rate was 1%. [2] ***P=100, N=9900, TP=90, FN=10, FP=99, TN=9801. Cost = 0xTN+2xTP+5xFP+10xFN = 775.***

14. We want to compare two methods of determining feature importance. In method A, the feature whose importance is being determined is eliminated and the model is trained on one less feature. In method B, the feature is not eliminated, but its value is permuted randomly across samples.
    a) How do we determine feature importance in each case? [1] ***We look at reduction in accuracy***
    b) State any one advantage of either method over the other with justification. [1] ***Permuting ensures that there are no effects with respect to number or variance of variables, and hence a better measure than elimination. Other logically argued advantages can also be considered***