CS 559: Final
Duration: 2:30

Answer all problems. Show all calculations and provide sufficient explanation. If in doubt, explain more.

Short and to-the-point answers are preferred.

NAME:

Problem 1. (30 points) True of False. If you choose false, briefly explain.

- SVMs can be used in an ensemble classifier. (True or False)

- In bagging, we choose random samples from the input without replacement. (True or False)

- A deep neural network with a non-linear activation function in each layer is meaningless. (True or False)

- Linear discriminant functions always lead to non-convex decision regions. (True or False)

- To achieve better performance, we should look at the testing set when training a classifier. (True or False)

- When applying K-Means for clustering, the distance measure should always be differentiable. (True or False)

- Compared to SVM, Adaboost is more sensitive to and thus easier to fail on un-normalized data. (True or False)

- A kernel matrix have to be designed to be invertible. (True or False)

- Using high-degree polynomial features help to reduce overfitting in a SVM classifier. (True or False)

- EM algorithm always finds the global optimal. (True or False)

- Deep neural networks usually require fewer training data as compared to Adaboost. (True or False)

- When training a random forest, the order of training the decision trees does not play a role. (True or False)

- When applying SVM, we have to explicitly specify a kernel function that maps a sample from a lower dimension to a higher dimension. (True or False)

- Deep neural networks rely on setting the weights of the networks mannually.

- In Adaboost, the weak learner that yields a higher error rate results in a higher weight in the combined classifier.

Problem 2. (20 points) In AdaBoost, the weight of sample x_i in round t is updated according to:

$$D_{t+1}(i) = D_t(i) \cdot exp(-\alpha_t y_i h_t(x_i))$$

where $y_i \in \{-1, 1\}$ is the label of sample x_i, $h_t(x_i)$ is a weak learner that returns $+1$ or -1 depending on which class it predicts for x_i and α_t is the (positive) weight of the weak learner.

(a) Explain under which conditions the weight of a sample is increased or decreased. (5 points)

(b) How does this mechanism of updating the weights affect the selection of subsequent weak learners? (5 points)

(c) Initially, all samples have equal weights. In the first iteration, we use a weak learner $h_1(x)$ on the data, update the weights and normalize them to sum to 1. Which of the following weight distributions, both sum up to 1, is NOT valid and why? (10 points)

1. $D_2 = \{0.1667\ \ 0.25\ \ 0.25\ \ 0.1667\ \ 0.1667\}$
2. $D_2 = \{0.1538\ \ 0.2308\ \ 0.2308\ \ 0.1538\ \ 0.2308\}$

Problem 3. (15 points) Given a set of 2D data $X = \{[-4\ \ 1], [-4\ \ -1], [1\ \ 1], [1\ \ -1], [3\ \ 1], [3\ \ -1]\}$, we apply the k-means algorithm to cluster them into two clusters. Starting from two different initializations, we obtain two solutions represented by their centroids:

Sol. 1: $C_1 = [-4\ \ 0]$ and $C_2 = [2\ \ 0]$

Sol. 2: $C_1 = [0\ \ 1]$ and $C_2 = [0\ \ -1]$

How would you select which of the two solutions to use? Which clustering is preferable and why? It is safe to assume that the algorithm has converged in both cases.

Problem 4. (15 points) Consider the k-means clustering algorithm which comprises the following steps:

1. Initialize by selecting k cluster centers arbitrarily and assign each example to the closest center.

2. Compute sample means for each cluster.

3. Reassign all samples to the closest mean.

4. If clusters changed at step 3, go to step 2.

The objective function of the k-means algorithm is the sum of squared errors:

$$J_{SSE} = \sum_{i=1}^{k} \sum_{x_j \in D_i} ||x_j - \mu_i||^2$$

where x are the samples, D_i is the i^{th} group and μ_i is its center.

Explain how the objective function is minimized by step 2 of the k-means algorithm and, then, how it is further minimized by step 3.

Problem 5. (10 points) You have been given 1000 samples from each of two classes, ω_1 and ω_2, in 100-D and you wish to find the number of principal components that should be kept in order to achieve the maximum possible accuracy using a linear SVM after PCA. Consider all parameters of the linear SVM fixed.

If k denotes the number of principal components kept (and 100-k the number of principal components discarded) that result in good classification performance, describe the procedure you would follow in a way that it can be implemented by someone who is not familiar with PCA or the classifier, but can implement all necessary data structures and store intermediate results if necessary.

Problem 6. (10 points) Consider the problem of training a classifier on a training set comprising two **linearly separable** classes. Explain why maximizing the margin using a Support Vector Machine is a better option than

(a) minimizing the number of misclassified samples using a Perceptron (5 points)

(b) training a classifier using Maximum Likelihood Estimation and modeling the classes as multi-variate Gaussian distributions (5 points)