COSC 7362 Advanced Machine Learning (*Dr. Eick*)
Some Solution Sketches
for Quiz2
Mo., November 23, 2015

Your Name:
Your Student id:

Problem 1 --- CNN [13]
Problem 2 --- Trajectory Classification [16]
Problem 3 --- ST-DBSCAN [8]
Problem 4 --- Ensemble Learning and AdaBoost [18]

Σ [55]:

Grade:

The exam is "open books" and you have 70 minutes to complete the exam.
The exam will count approx. 19-24% towards the course grade.

1) ImageNet Classification with Deep Convolutional Neural Networks...[13]

a) What are the main characteristics/properties of Convolutional Neural Networks? How does their architecture differ from classical feedforward neural networks? Limit your answer to at most 7 sentences! [6]

Inputs correspond to regions in the visual field [1]
Overlapping pooling [1]
Deep; partially connected layers (followed by fully connected layers) [1.5]
Pooling layer(s) [1]
Employ non-saturated activation functions (kind of f(x)=max(x,0)) for faster convergence[1.5]
Other things: dropout[0.5], architecture is scalable and facilitates GPU processing[0.5], local response normalization.

b) CNN employ overlapping pooling; what does this mean? What are the advantages of this approach? [3]

Inputs are shared; reduction in overfitting

c) CNN's use *dropout* to reduce overfitting in CNNs. How does dropout work? Why does it reduce overfitting? [4]
Neurons participate only with probability 0.5 in forward computing during training, and learnt weights are reduced to half in testing.
Learns more robust, abstract features that are useful in conjunction with many subsets of neurons, and not only a single one.

2) TraClass—Trajectory Classification using Region-based and Trajectory-based Clustering [16]

a) What is the goal of Region-based Clustering? Describe the approach the paper proposes to obtain region-based clusters! Limit your answer to the 7-9 sentences. [6]

Region based clustering discovers regions that have trajectories mostly of one class. After trajectory partitioning is performed, region based clustering is performed as long as homogeneous regions of reasonable size are found. If for a 2-dimensional region space, there are trajectory partitions of major class $\geq \Psi$ (threshold) trajectories but all other class have $\leq \Psi$ trajectories, it is called a region of the major class. The algorithm starts with one big partition enclosing the entire range. Alternatively for X & Y axes, the algorithm selects a partitioning that has maximum code cost $[(L(H)^{max} + L(D|H))]$ and divides it into two parts. This procedure is repeated until there is no decrease in code cost in both X & Y axes. The term $[L(H) + L(D|H)]$ gives tradeoff between the homogeneity & conciseness of regions. After all possible homogeneous regions are discovered, adjacent ones are merged if they share same class.

b) Why is it desirable to have class conscious trajectory partitions? How are class conscious partitions obtained? [4]

Class conscious partitioning makes sure that clusters only contain segment belonging to the same class; that is, of segments of a trajectory partition belong to the same class. An algorithm is employed that splits heterogeneous (sub-)trajectories further into a set of homogeneous ones, by splitting trajectory partitions between segments belonging to different classes.

c) What is the result of the TraClass Clustering process? How are these clustering results used to create features for trajectories that can then be used by a classification algorithm to learn a model? What information do the created features capture? [6]

Set of trajectory and regional clusters.

are defined

Many failed to describe clearly how features, namely how clustering information is used to annotate trajectories; namely, spatial relationships of the trajectory to the obtained clusters are used to create cluster specific features for the trajectory to be annotated, such as how often the trajectory in the test set traverses a particular region-based cluster, or what percentage of a trajectory is in close proximity to a trajectory-based cluster.

3) ST-DBSCAN Article [8]

a) How does ST-DBSCAN generalize DBSCAN—which obtains spatial clusters— to obtain spatial-temporal clusters? [4]

It uses two distances metrics. one for the spatial values, one for non-spatial values. To ensure temporal support, the data is filtered such that retain only temporal neighbors: non-spatial values are observed in consecutive time units. Then density based clustering is applied with the distance metric

b) What role does the parameter Δε play in the ST-DBSCAN algorithm—how does it influence the clustering process and the shapes of the obtained clusters? [4]

Δε is used to stableze adjacent clusters' boundaries. Clusters, due to spatial co-relation, tend to share similar non-spatial values. Thus, the algorithm ensure that, when a datapoint is assigned to a cluster, its value must not be different from the cluster average for non spatial attribute within a threshold of Δε. ✓

Considering Δε, the algorithm is a bit slower, but the obtained clusters will have more uniform non-spatial values. For example a cluster is unlikely to have data with high temperature on one side, and low temperature on the other side

As this requirement imposes a constraint on growing clusters beyond a certain point under certain circumstances, this sometimes leads to breaking up a big cluster into a set of smaller clusters that have better agreement with respect to their non-spatial attributes.

4) Scholarpedia Ensemble Learning Algorithm and Freund/Shapiro Article [18]

a) How is it possible that an ensemble classifier accomplishes an accuracy of 90%, although its base classifiers have an accuracy of about 70%? [4]

see also additional sheet at the end for a different, more statistical answer to this question; other answers received partial credit;

For example, we could have three classifiers that classify ten examples, and classifier C1 miss-classifies example 1, 2, 10, and C2 examples 3, 4, 10 and C3 examples 5, 6, 10; then using majority vote these 3 classifiers will misclassify only example 10, accomplishing an accuracy of 90%. That is using majority vote of classifiers that make different kind of errors leads to an enhancement in accuracy as long as the base classifiers' accuracy is above 50%—if it not above 50% such a majority of "correct" classifiers no longer exists[1].

b) Why do base-classifiers need to have an accuracy of above 50%? [2]

... accuracy decreases in this case; other answer: some theoretical error bounds for AdaBoost are no longer valid

c) What is the key idea for boosting? Systems that employ boosting have been quite successful in the past. Why do you believe is this the case—you can speculate? [4]

...weights of difficult to classify examples are increased, making it attractive for the next classifier to classify those correctly (and others incorrectly), obtaining a classifier that makes different errors compared to its predecessor....

d) What does the parameter β_t measure? What is its influence on the weight update? [2]

no answer given!

e) How does a classifier that has been obtained by running AdaBoost make is classification decision—be precise! [2]

no answer given!

f) What is the key idea of Stacked Generalization? How is training done, when Stacked Generalization is used? [4]

...the Tier2 classifier is trained on a dataset consisting of the labels of the Tier1 classifiers assign to a particular training example, the attributes of the training example, and the class the original example belongs to...

[1] This statement is also a third possible answer to (the next) question b!

Why does it work?

- Suppose there are 25 base classifiers
  - Each classifier has error rate, $\varepsilon = 0.35$
  - Assume classifiers are independent
  - Probability that the ensemble classifier makes a wrong prediction:

$$\sum_{i=13}^{25} \binom{25}{i} \varepsilon^i (1-\varepsilon)^{25-i} = 0.06$$