Exam2
COSC 6342 *Machine Learning*
March 26, 2009

Your Name:
Your SSN:

Problem 1 [12]: Decision Trees (and kNN)
Problem 2 [5]: DENCLUE
Problem 3 [12]: Non-Parametric Density Estimation
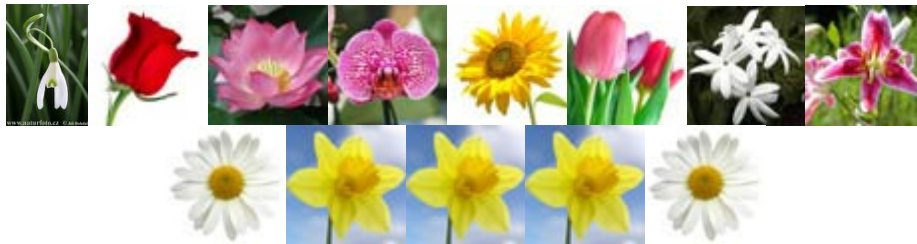Problem 4 [7]: Editing and Condesing
Problem 5 [8]: PCA
Problem 6 [10]: DBSCAN and Density-based Clustering
Problem 7 [9]: EM and Mixtures of Gaussians

Σ [63]:

Grade:

The exam is "open books and notes" and you have 75 minutes to complete the exam. The exam is slightly too long. The exam will count approx. 24% towards the course grade.

1) Decision Trees (and kNN) [12]

a) Compare decision trees with kNN to solve classification problems. What are the main differences between these two approaches? [5]

kNN	Decision Tree
- **Lazy learner**	- **Learn model (tree)**
- **Local model**	- **Global model**
- **Distance based**	- **Based on attribute order**
- Voronoi convex polygon decision boundary	- **Rectangle decision boundaries**
	- Hierarchical learning strategy

b) We would like to predict the gender of a person based on two binary attributes: leg-cover (pants or skirts) and beard (beard or bare-faced). We assume we have a data set of 20000 individuals, 16000 of which are male and 4000 of which are female. 80% of the 16000 males are barefaced. Skirts are present on 50% of the females. All females are bare-faced and no male wears a skirt.

   i)   Compute the information gain of using the attribute leg-cover for predicting gender! Just giving the formula that computes the information gain is fine; you do not need to compute the exact value of the formula! Use H as the entropy function in your formula (e.g. $H(1/3, 2/3)$ is the entropy that 1/3 of the examples belong to class1 and 2/3 of the examples belong to class 2). [2]
   ii)   Computer the information gain of using the attribute beard to predict gender! [2]

   i)   **Gain(D, leg-cover) = H(1/5, 4/5) – (1/10)*H(1, 0) – (9/10)*H(1/9, 8/9)**
   ii)   **Gain(D, beard) = H(1/5, 4/5) – 4/25 H(0,1) – 21/25 H(16/21, 5/21)**

(*) This question doesn't require you to compute the exact value but you have to write the formulas in above forms to get credit.

c) Why do decision tree learning algorithms grow the entire tree and then apply pruning techniques to trim the tree to obtain a tree of smaller size? [3]

   - **It grows the entire tree to fit training data (search as much detail as possible in hypothesis space).**
   - **Then it prunes the tree to avoid over-fitting.**

2) DENCLUE [5]

What role do non-parametric density functions play for the DENCLUE clustering algorithm? Give a description how the DENCLUE algorithm clusters a data set. Limit your answer to the second question to at most 6 sentences.

- **Non-parametric density functions are used to derive a density function for the dataset; maxima of the so defined density functions are called density attractors, separate hills of the density function correspond to different clusters that will be identified by DENCLUE**

- **Description:**
  - **Hill climbing is uses to associate the objects in the dataset with density attractors**
  - **Objects that are associated with the same density attractor will be in the same cluster.**

3) Non-parametric Density Estimation [12]

a) Assume a dataset X={x^t,r^t}consisting of 4 examples (0,1), (1,3), (2,7),(4,1) is given and the bin-width is 2.5: assume that x and x' belong to the same bin if |x-x'|≤2.5.

a1) Compute the values (also give the formula) for the regressogram for inputs 0.2, 1.7, and 4.6 for the mean smoother (see formula 8.19 on page 165 of the textbook) [4].

$\hat{g}(0.2)=$

$\hat{g}(1.7)=$

$\hat{g}(4.2)=$

$$\hat{g}(0.2) = \frac{b(0.2,0)\times1+b(0.2,1)\times3+b(0.2,2)\times7+b(0.2,4)\times1}{b(0.2,0)+b(0.2,1)+b(0.2,2)+b(0.2,4)} = \frac{1\times1+1\times3+1\times7+0\times1}{1+1+1+0} = \frac{11}{3}$$

$$\hat{g}(1.7) = \frac{b(1.7,0)\times1+b(1.7,1)\times3+b(1.7,2)\times7+b(1.7,4)\times1}{b(1.7,0)+b(1.7,1)+b(1.7,2)+b(1.7,4)} = \frac{1\times1+1\times3+1\times7+1\times1}{1+1+1+1} = 3$$

$$\hat{g}(4.2) = \frac{b(4.2,0)\times1+b(4.2,1)\times3+b(4.2,2)\times7+b(4.2,4)\times1}{b(4.2,0)+b(4.2,1)+b(4.2,2)+b(4.2,4)} = \frac{0\times1+0\times3+1\times7+1\times1}{0+0+1+1} = 4$$

Now assume the bin-width is only 1. Recompute the prediction for input 1.7!

$\hat{g}(1.7)=$

$$\hat{g}(1.7) = \frac{b(1.7,0)\times1+b(1.7,1)\times3+b(1.7,2)\times7+b(1.7,4)\times1}{b(1.7,0)+b(1.7,1)+b(1.7,2)+b(1.7,4)} = \frac{0\times1+1\times3+1\times7+0\times1}{0+1+1+0} = 5$$

a2) In general the function obtained using the above approach has discontinuities. What could be done to obtain a continuous function? [2]
- **Use a influence function for b (i.e a Gaussian)**

b) What is the main difference between the Gaussian Kernel Density function approach as described on page 157 of the textbook and the k-nearest Neighbor Density Estimator that has been described in Section 8.2.3. [3]

- **Gaussian Kernel density approaches uses a fixed, global width while the kNN estimators employs a local, variable width that that corresponds to the k-nearest neighbor distance (if the density is measure in point v $d_k(v)$ determines the width used when measuring the density in v)**

c) What advantages you see in using non-parametric density estimation approach compared to parametric density approaches, such as multivariate Gaussians? [3]
Non-parametric density estimation approach:
- **Doesn't have to make assumption about model.**
- **Can estimate more complex models/shapes than parametric density approaches.**
- **Uses a non-global, Regional/case based approach to measure density**
- **Is easy to parallize on a single instruction multiple data (SMID), each processor stores 1 training function values for that instance.**

4) Editing/Condensing/Toussaint Paper [7]
a) Both editing and condensing a are popular in conjunction with kNN classifiers. What is the goal of dataset editing? What is the goal of dataset condensing? [3]
- **Editing: enhancing accuracy**
- **Condensing: improving speed by reducing the size of the dataset**

b) Give a sketch of an algorithm that uses Voronoi diagrams (or their dual Delaunay graphs) for condensing a classification dataset![4]

- **Voronoi condensing diagram retains the points whose neighbors are of opposite class and removes the points whose neighbors are of the same class. This is done by Delaunay triangulation.**

5) PCA [8]

a) What is the goal of Principal Component Analysis (PCA)? Limit your answer to at most 5 sentences. [4]

- **Its goal is computing most meaningful basis to represent data in a lower dimensional space.**

b) The eigenvectors chosen to form the transformation w^T that reduces dataset dimensionality in PCA have to orthonormal. What does this mean? Why is it desirable that the selected eigenvectors are orthonormal? [4]

- **Othonomal means their covariance is equal to 0.**
  Other answers may receive full credit.

- **If they are dependent (not orthonormal), their contribution to the variance will be less; e.g. two variable with correlation 1 contribute as much to the density as just using one of the two variable.**

6) DBSCAN & Density-based Clustering [10]

a) What the characteristics of objects that are classified as outliers by DBSCAN? [2]

- **They are not in the radius of a core point.**
- **Or they are not density reachable from any core point.**

b) How does DBSCAN form clusters? Limit your answer to at most 5 sentences [3]

- **It specifies two parameters: radius (Eps) and MinPts (number of points).**
- **Then, it finds the *core points*, which have more than MinPts within Eps, the *border points*, which have fewer than MinPts within Eps but is in the neighborhood of a core point, and the *noise points*, which are not a core point or a border point.**
- **Any two points x and y to be *density connected* if there exists a core point z, such that both x and y are density reachable from z.**
- **Next, the method iterates from each core point, and finds all other points *density connected* to it; all such points belong to the same cluster.**

c) DBSCAN does not well to cluster datasets that have clusters of varying densities. What is the explanation for that [2]?

- **Because there is no global parameter setting for radius (Eps) or number of points (MinPts) that captures all the clusters. If we try to caputure high**

density clusters, points in low density area may be considered as outliers. Otherwise, if try to capture low density clusters, high density and low density clusters may be merged together into a single cluster.
- *Other answers might also receive credit!*

d) What is the runtime complexity of DBSCAN?[1]
- **Both $O(n^2)$ and $O(nlogn)$ are correct answers.**

e) Give a sketch of an approach that is capable of clustering datasets that contain clusters with varying degrees of density [2+up to 4 extra points].

- **Start with large MinPts (number of points) and small Eps (radius) to deal with high densities. After finding the high density clusters, marked them out and cluster the remaining data points by decreasing the MinPts and increasing the radius.**
- *Other answers may receive full credit.*

7) EM and Mixture of Gaussians [9]

a) What advantages do you see in using density functions based on mixtures of Gaussian instead of a single Gaussian? [2]

- **It can deal with *multimodal* classes/clusters**

b) Summarize in natural language what computations EM performs during its M-step! [3]

Estimating new parameters (or re-estimate parameters):
- **Membership of one data example to the clusters.**
- **Centroid (or mean) of each cluster.**
- **Covariance of each cluster.**

* For this question, you have to explicit names (meanings) of parameters updated in M-step to receive full credit.

c) EM uses covariance matrices when clustering data. What are the benefits of doing that? Hint: one way to answer this question is to compare EM with a restrictive version of EM that just uses covariance matrices whose diagonal is 1 and whose other entries are 0. [4]

- **It can find ellipsoid clusters (characterized by different entries in the diagonal of the covariance matrix) with arbitrary orientations (characterized by non-0 in off-diagonal entries of the covariance matrix) .**