Midterm: Learning Theory II (580.692)

Instructor: René Vidal, Office: 308B Clark Hall, E-mail: *rvidal@cis.jhu.edu*

November 7, 2006

HONOR SYSTEM: This examination is strictly individual. You are not allowed to talk, discuss, exchange solutions, etc., with other fellow students. Furthermore, you are not allowed to use the book or your class notes. You may only ask questions to the class instructor. Any violation of the honor system, or any of the ethic regulations, will be immediately reported according to JHU regulations.

NAME: ____________________________________ **Signature:**________________________

1. **(20 points) Dimensionality Reduction**

Let $\psi : \mathbb{R}^d \rightarrow \mathbb{R}^D$, where $d \ll D$, be a differentiable function from $\mathbb{R}^d$ to $\mathbb{R}^D$, and $k : \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}$ be a positive dedinite kernel on $\mathbb{R}^D$. Let $\boldsymbol{X} = \{\boldsymbol{x}_i \in \mathbb{R}^D\}_{i=1}^N$ be a set of points drawn according to the model

$$\boldsymbol{x} = \psi(\boldsymbol{z}) + \boldsymbol{w}, \tag{1}$$

where $\boldsymbol{w} \in \mathbb{R}^D$ is a small amount of noise. Assume that the map ψ, the points $\{\boldsymbol{z}_i \in \mathbb{R}^d\}_{i=1}^N$ and the noise are unknown, but the dimensions d and D, and the kernel k are known. You would like to find a low-dimensional representation $\boldsymbol{Y} = \{\boldsymbol{y}_i \in \mathbb{R}^d\}_{i=1,N}$ for the data X, and you consider either PCA, KPCA or LLE.

(a) Given d, under what conditions in $\boldsymbol{X}$ would you use PCA?

(b) Given d and k, under what conditions in $\boldsymbol{X}$ would you use KPCA?

(c) Given d, under what conditions in $\boldsymbol{X}$ would you use LLE ?

(d) Imagine you would like to use model selection to decide whether to use PCA or LLE. Write down the cost functions that each one of the two methods minimizes for fitting the data. Are these functions comparable? Use these functions to obtain a model selection criterion for choosing which of the two methods to use.

2. **(20 Points) Manifold Clustering**

   Locally Linear Embedding (LLE) is an algorithm for obtaining a low-dimensional representation $\{\boldsymbol{y}_i \in \mathbb{R}^d\}$ of a given set of points $\{\boldsymbol{x}_i \in \mathbb{R}^D\}_{i=1}^N$ lying in a d-dimensional manifold $\mathcal{M} \subset \mathbb{R}^D$. Given a dimension d and a number of nearest neighbors k, the algorithm proceeds as follows

   - Find a set of weights W_{ij} that minimize the cost function $\sum\limits_{i=1}^N ||\boldsymbol{x}_i - \sum\limits_{j=1}^N W_{ij}\boldsymbol{x}_j||^2$ subject to the constraints $\sum\limits_{j=1}^N W_{ij} = 1$ and $W_{ij} = 0$ if $\boldsymbol{x}_j$ is not one of the k-nearest neighbors of $\boldsymbol{x}_i$.
   - Find a low-dimensional representation $\{\boldsymbol{y}_i\}_{i=1}^N$ that minimizes the cost function $\sum\limits_{i=1}^N ||\boldsymbol{y}_i - \sum_{j=1}^N W_{ij}\boldsymbol{y}_j||^2$ subject to the constraints $\sum\limits_{i=1}^N \boldsymbol{y}_i = 0$ and $\frac{1}{N}\sum\limits_{i=1}^N \boldsymbol{y}_i\boldsymbol{y}_i^T = I$.

   Notice that both steps involve solving a linear system.

   In this problem, you will generalize LLE to data $\{\boldsymbol{x}_i \in \mathbb{R}^D\}_{i=1}^N$ drawn from a union of n d-dimensional manifolds in $\mathbb{R}^D$, $\{\mathcal{M}_m \subset \mathbb{R}^D\}_{m=1}^n$. More specifically,

   (a) Write down a cost function similar to that of K-means for clustering the data according to the n manifolds.

   (b) Write down a set of constraints that need to be satisfied in order to make the problem well-posed.

   (c) Derive an algorithm for minimizing the cost function in (a) subject to the constraints in (b). Your method should alternate between the computation of an LLE model for each group, and the segmentation of the data according to the n models.

   (d) Does the algorithm converge? Why? Doest is converge to the global optimum? Why? Does it converge in a finite number of iterations? Why?

3. **(30 points) Segmentation of Subspaces of Different Dimensions.**

Consider now a collection of points $\{\boldsymbol{x}_i \in \mathbb{R}^3\}_{i=1}^P$ lying in 3 subspaces of $\mathbb{R}^3$

$$\begin{aligned} S_1 &= \{\boldsymbol{x} : x_3 = 0\} \\ S_2 &= \{\boldsymbol{x} : x_1 = 0 \wedge x_2 + x_3 = 0\} \\ S_3 &= \{\boldsymbol{x} : x_1 = 0 \wedge x_2 - x_3 = 0\} \end{aligned}$$

(a) **(10 points)** Show that the data can be fit with a set of m homogeneous polynomials of degree $n = 2$ in 3 variables. Determine the value of m. Write down the m polynomials explicitly. What is the minimum number of points P and how should such points be distributed in S_1, S_2 and S_3 so that the m polynomials can be uniquely determined? Show how to determine m and the polynomials from data. Compute the gradient of each one of the m polynomials at a data point $\boldsymbol{y}_1 \in S_1$, $\boldsymbol{y}_2 \in S_2$ and $\boldsymbol{y}_3 \in S_3$. Is it possible to segment the data into the three subspaces using these gradients? If yes, say how. If not, say what segmentation can be obtained from the gradients.

(b) **(10 points)** Answer questions in (a) with $n = 3$.

(c) **(10 points)**Answer all questions in (a) with $n = 4$. If your answer to the last question in (a) is yes, then explain why the data can be segmented correctly into the three subspaces, even though the degree of the polynomials is greater than number of subspaces.