Math Foundations of ML, Fall 2017

Homework #10

Due Monday December 4, at the beginning of class

As stated in the syllabus, unauthorized use of previous semester course materials is strictly prohibited in this course.

1. Using you class notes, prepare a 1-2 paragraph summary of what we talked about in class in the last week. I do not want just a bulleted list of topics, I want you to use complete sentences and establish context (Why is what we have learned relevant? How does it connect with other things you have learned here or in other classes?). The more insight you give, the better.

2. Please fill out the CIOS course evaluation survey (you should have a link to it in your email). Turn in a short acknowledgement that you have done so.

3. Suppose the random variables (X, Y), $X \in \mathbb{R}^2$, $Y \in \{1, 2\}$, have joint distribution given by

$$\mathrm{P}(Y = 1) = \mathrm{P}(Y = 2) = 1/2, \quad f_X(\boldsymbol{x}|Y = y) = \frac{1}{2\pi\sqrt{\det(\boldsymbol{\Sigma}_k)}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)\right),$$

where

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{\mu}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{\Sigma}_1 = \begin{bmatrix} 3 & -6 \\ -6 & 18 \end{bmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Draw the regions $\Gamma_1(h^*)$ and $\Gamma_2(h^*)$ that correspond to the Bayes classifier. (You can feel free to use MATLAB or Python for this.)

4. The file `hw10p4_data` contains two arrays: `X1` and `X2`. These are samples from an unknown distribution, where `X1` has been assigned "class 1", and `X2` has been assigned "class 2". Implement the nearest neighbor algorithm, and sketch the decision regions Γ_1 and Γ_2 that it defines.

5. In actuality, the data in the last problem was generate using the model from problem 3. Estimate the generalization error $R(h)$ for both the Bayes classifier (problem 3) and the nearest-neighbor rule (problem 4), and compare the two. This will require the generation of many Gaussian random vectors with specified covariance matrices, which you certainly know how to do.

6. The file `hw10p6_data` contains an array `X` whose columns should be interpreted as data points in $\mathbb{R}^2$. Implement the EM algorithm, and use it to train a Gaussian mixture model with 5 components. Initialize the algorithm with densities of the form Normal$(\boldsymbol{m}_k, \gamma_k\mathbf{I})$ for reasonable choices of $\boldsymbol{m}_k$ and γ_k that you surmise simply by inspecting a scatter plot by eye. Turn in your code, and contour plots of each of your 5 mixture components overlayed on a scatter plot of the data.