CSCI 5521: Introduction to Machine Learning (Fall 2017)[1]

Homework 1

Due date: Monday, Oct 2nd, 11:55pm

1. (**30 points**) Find the Maximum Likelihood Estimation (MLE) of θ in the following probabilistic density functions. In each case, consider a random sample of size n. Show your calculation:

   (a) $f(x|\theta) = \frac{x}{\theta^2} \exp\{\frac{-x^2}{2\theta^2}\}, x \geq 0$

   (b) $f(x|\alpha, \theta) = \alpha\theta^{-\alpha}x^{\alpha-1} \exp\{-(\frac{x}{\theta})^\alpha\}, x \geq 0, \alpha > 0, \theta > 0$

   (c) $f(x|\theta) = \frac{1}{\theta}, 0 \leq x \leq \theta, \theta > 0$ (Hint: You can draw the likelihood function)

2. (**30 points**) Question 4.6 from the Alpaydin's textbook 3rd Edition. Please answer the question on your report and submit your source code in the following MATLAB template file that we provided:

   - `main_question2.m`: Template file for question 2.

3. (**40 points**) In this programming exercise you will implement a multivariate Gaussian classifier, with two different assumptions:

   - Assume S_1 and S_2 are learned independently (learned from the data from each class).
   - Assume $S_1 = S_2$ (learned from the data from both classes).

   What is the discriminant function in each case? Show in your report and briefly explain.

   Your program should fit two Gaussian distributions to the 2-class training data in `training_data.txt` to learn m_1, m_2, S_1 and S_2. Then, you use this model to classify the test data in `test_data.txt` by computing the log odds $\log \frac{P(C1|x)}{P(C2|x)}$ with $P(C_1) = 0.2$ and $P(C_2) = 0.8$.

   What is the error rate obtained for the test set in each assumption? Briefly explain.

   We provided a MATLAB template code, which you are required to use. **Please make sure to follow exactly the same input/output parameters provided in each function. Failing to do so may result in points lost.** Complete de following files:

[1] Instructor: Catherine Qi Zhao (qzhao@umn.edu). TAs: Raphael Petegrosso (peteg001@umn.edu); Xinyan Li (lixx1166@umn.edu).

- `ReadData.m`: Reads content from `training_data.txt` and `test_data.txt` and return as matrices.
- `CalculateMeanIndepCov.m`: Calculates m_1, m_2, S_1, S_2 for the training dataset, assuming independent covariance for each class.
- `CalculateMeanSameCov.m`: Calculates m_1, m_2, S for the training dataset, assuming same covariance for both classes.
- `CalculateGaussianDiscr.m`: Finds the discriminants g_1 and g_2 for the test set utilizing the parameters learned in each case.
- `CalculateErrorRate.m`: Calculates the error rate for the test set according to the $g1$ and $g2$ discriminants obtained in each case.
- `main_question3.m`: Main file that calls the appropriate functions to classify the test set and calculate the error rate for each assumption.

More details about the inputs/outputs can be found in each source file.

You are not allowed to use MATLAB built-in functions to calculate the Gaussian probability density function, such as `mvnpdf`. You are allowed, however, to use auxiliar functions such as `mean`, `cov` and `det`. You can create additional functions if you wish.

Submission

- **Submit a zip file containing:**
  1. hw1_sol.pdf: a document containing all your answers.
  2. All the source files required to run your code (except the data).
- All material must be submitted electronically via Moodle.