CSCI 590: Machine Learning

Homework 1

Due Date: 2/8/2018

Please submit your homework to TA Wen-Hao Chiang, email: chiangwe@umail.iu.edu.

Question 1. (10 points) Let $x_1, \ldots, x_n$ be n independent identically distributed (i.i.d.) random variables with mean μ and variance σ^2. Let y_n be a new random variable defined as the weighted average of x_i's as follows.

$$y_n = \sum_i^n w_i x_i$$

a. Find the mean and variance of y_n as a function of w_i.
b. Repeat part a but this time assume that x_i's are identically distributed but they are only correlated rather than independent.

Question 2. (10 points) Let $\boldsymbol{x}$ be a multivariate random variable distributed according to $N(\boldsymbol{\mu}, \Sigma)$. Prove that

$$E(\boldsymbol{x}) = \mu$$
$$E((\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^T) = \Sigma$$

Question 3. (5 points) Show that the number of ways n samples can be assigned to K classes is equal to

$$\frac{n!}{m_1! m_2! \ldots m_k!}$$

Question 4. (5 points) Let $\boldsymbol{\mu}$ be distributed according to a symmetric Dirichlet distribution with a parameter α. Find $E(\boldsymbol{\mu})$ and $Var(\boldsymbol{\mu})$.

Question 5. (30 points) In this exercise you will write a script that will generate samples from the Multinomial-Dirichlet data model.

- Use a symmetric Dirichlet distribution to generate 5 topics for a vocabulary containing 50 words.

$$\boldsymbol{\mu}_i \sim Dir(0.5) \quad i = \{1, \ldots, 5\},$$

- For each $\boldsymbol{\mu}_i$ generate 10 documents with 200 words each from a Multinomial distribution (total 50 documents).
- For each group of 10 documents plot the histograms of the frequency of words. Find the maximum likelihood estimates of μ_i.
- Consider documents generated from the same topic μ_i as belonging to the same category. Implement a naïve Bayes classifier, train it on half of the cases (25 documents) and test it on the remaining half. What is the accuracy of your classifier?
- Repeat the data generation and classification (parts a, b, and d) hundred times and compute the mean and the standard deviation of your accuracy. Is your classifier better than a random classifier?

Question 6. (40 points) In this exercise you will deal with documents with mixtures of topics (unlike exercise 5 where you had documents with one topic). Each mixture of topics will define a class.

- Use the five topics $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_5$ you generated in Question 5a and use the following $\boldsymbol{\beta}$ values to generate 10 documents with 200 words each for each class (each $\boldsymbol{\beta}$ defines a unique category)
  - $\boldsymbol{\beta}_1 = [0.50.30.10.050.05]$
  - $\boldsymbol{\beta}_2 = [0.20.60.050.050.1]$
  - $\boldsymbol{\beta}_3 = [0.10.10.30.30.2]$
- Use the naïve Bayes classifier implemented in 5.d, train it on half of the cases (15 documents) and test it on the remaining half. What is the accuracy of your classifier? Repeat this process (data generation and classification) a hundred times and compute the mean and the standard deviation of your accuracy. How does your accuracy compare to your accuracy from exercise 5. What has changed and why?