CSCI 59000: Machine Learning
Assignment 1
Assigned: Jan 26, 2016
Due: Feb 9, 2016

Please submit your homework to TA
Halid Yerebakan
Email: hzyereba@umail.iu.edu

1. (10 points) Let $x_1,\ldots,x_n$ be n jointly distributed independent identically distributed (i.i.d.) random variables with mean μ and variance σ^2. We define the weighted average as follows.

$$y_n = \sum_{i=1}^{n} w_i x_i$$

$$\sum_{i=1}^{n} w_i = 1$$

a. Find the variance of y_n as a function of w_i.
b. Now assume that $x_1,\ldots,x_n$ are identically distributed with mean μ and variance σ^2, but they are only uncorrelated rather than independent. Find the variance of y_n.

2. (10 points) Let $x_1,\ldots,x_n$ be n jointly distributed independent identically distributed (i.i.d.) random variables from U(-a/2,a/2). That is,

$$p(x) = \begin{cases} 0 & x < -\frac{a}{2} \\ \frac{1}{a} & -\frac{a}{2} \le x < \frac{a}{2} \\ 0 & x \ge \frac{a}{2} \end{cases}$$

Write down a formula for an ML estimate of a.

3. (10 points) Let x be a random variable distributed according to $N(\mu, \sigma^2)$. Prove that $E(x)=\mu$ and $E((x- E(x))^2)= \sigma^2$.

4. (30 points) In this exercise you will write a script that will generate samples from the Multinomial-Dirichlet data model.

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

$$\mu_i \sim Dir(0.5) \text{ for i=}\{1,\ldots,5\},$$

b. For each μ_i generate 10 documents with 200 words each from a Multinomial distribution (total 50 documents).

c. For each group of 10 documents plot the histograms of the frequency of words. Find the maximum likelihood estimates of μ_i.

d. 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?

e. 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?

5. (40) In this exercise you will deal with documents with mixtures of topics (unlike exercise 4 where you had documents with one topic). Each mixture of topics will define a class.

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

   $$\mu_i \sim Dir(0.25) \qquad \text{for i=\{1,…,5\},}$$

   Use the following β values to generate 10 documents with 200 words each for each class (each β defines a unique category)

   Class 1: [0.5 0.5 0 0 0]
   Class 2: [0.5 0 0.25 0.25 0];
   Class 3: [0 0.5 0 0.25 0.25];
   Class 4: [0 0.25 0.5 0.25 0];
   Class 5: [0.25 0.25 0 0 0.5];

   Use the naïve Bayes classifier implemented in 4.d, train it on half of the cases (25 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 4. What has changed and why?