Assignment 1: Linear Algebra and Probability

Submission: Tueday August 16th
Maximum 2 students per group

Prof. Fabio A. González
Machine Learning - 2011-II
Maestría en Ing. de Sistemas y Computación

---

1. Let $D = \{d_1, \ldots, d_n\}$ be a set of documents and $T = \{t_1, \ldots, t_m\}$ a set of terms (words). Let $TD = (TD_{i,j})_{i=1\ldots m, j=1\ldots n}$ be a matrix such that $TD_{i,j}$ corresponds to the number of times the term t_i appears in the document d_j. Assume a process where a document d_j is randomly choosen with uniform probability and then a term t_i, present in d_j, is randomly choosen with a probability proportional to the frequency of t_i in d_j.

   (a) How do you transform the matrix TD to obtain a matrix $P(T, D)$, such that $P(T, D)_{i,j} = P(t_i, d_j)$ (the joint probability of term t_i and document d_j).

   (b) How do you obtain the $P(T|D)$ matrix?

   (c) How do you obtain the $P(D|T)$ matrix?

2. Let l_i be the lenght, number of characters, of term t_i, and let $L = (l_1, \ldots, l_m)$ be a column vector.

   (a) Calculate $E[l]$, the expected value of the random variable l corresponding to the length of a randomly choosen term.

   (b) Calculate $\text{Var}(l)$, the variance of l.

   (c) Show that $\text{Var}(al + b) = a^2\text{Var}(l)$, where a and b are arbitrary constants.

3. Find an expression for $P(T|D)$ that exclusively uses $P(D|T)$, $P(T, D)$ and, optionally, constant matrices/vectors.

4. Find an expression for the matrix $COV = (\text{Cov}(t_i, t_j))_{i,j=1\ldots m}$, the covariance of the random variables corresponding to terms.

5. (optional) What is the meaning of the Eigenvector of COV correspondign to the largest Eigenvalue?

Note: In all the cases use standard matrix and scalar operations: transposition, matrix multiplication (*), matrix elementwise operations (+, -, .*, ./), matrix-scalar operations, etc.