ENGR-E 511 Fall 2018: Assignment #1

Due on Friday, September 14, 11:59P

Professor Minje Kim

Abhilash Kuhikar (akuhikar@iu.edu)

September 14, 2018

Contents

Problem 1: Picky Eater

The objective is to find the probabilities $P_{blueberry}$, $P_{strawberry}$ and P_{yogurt} using MLE given the class data and using MAP when we have a priori knowledge

(a) MLE for Multinomial Distribution using Lagrange's Multiplier:
Maximizing the likelihood function to get the probabilities:
$\frac{N!}{\prod_{i=1}^{i=K} X_i!} * \prod_{i=1}^{i=K} P_i^{X_i}$

By solving the above equation we get the probabilities as: $\frac{X_k}{N}$
where N is the total sample size and X_k is the size of any particular group.
We have,
$X_{blueberry} = 106$, $X_{strawberry} = 23$ and $X_{yogurt} = 50$
Hence, $P_{blueberry} = 106/179 = 0.592$
$P_{strawberry} = 23/179 = 0.128$
$P_{yogurt} = 50/179 = 0.279$

(b) MAP Estimation using a priori knowledge :
We know that the likelihood is the multinomial distribution. We need to find the conjugate priors for the given a priori distribution to solve MAP estimation.
Dirichlet distribution for finding conjugate priors(as given in the lecture slides):

$\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} * \prod_{k=1}^{K} (P_k)^{\alpha_k - 1}$

We know this:
$MAP = \frac{(MLE)*(a\ priori)}{constant}$

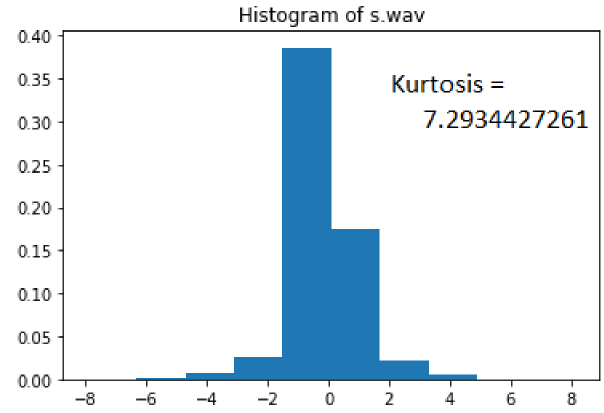
Multiplying the likelihood and prior, we get the function:
P(P1,P2...Pk — $\alpha 1, \alpha 2 ... \alpha_k) = \frac{N!}{\prod_{i=1}^{i=K} X_i!} * \prod_{i=1}^{i=K} P_i^{X_i} * \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} * \prod_{k=1}^{K} (P_k)^{\alpha_k - 1}$

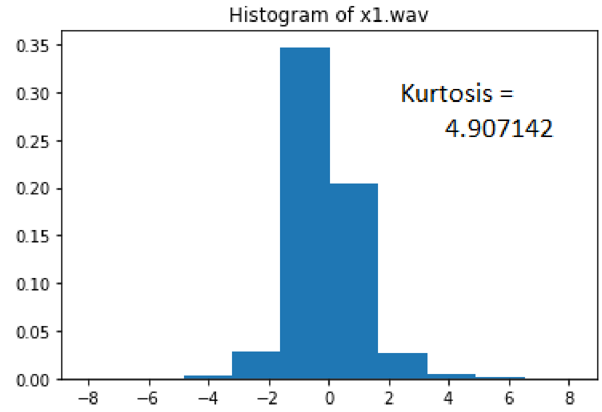
(Ref:https://stats.stackexchange.com/questions/304148/deriving-the-map-estimate-for-multinomial-dirichlet)
Maximizing this function using Lagrange multiplier, we will get the probability estimates as:
$\frac{X_k + \alpha_k - 1}{N + \alpha_0 - K}$

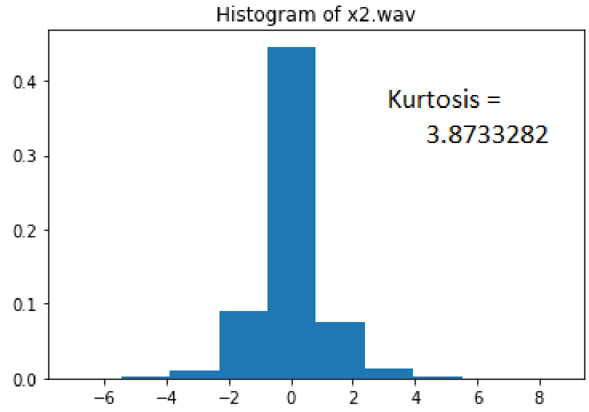
where α_k is the pseudo count for each category and α_0 is the sum of the pseudo counts.
I am assuming the pseudo count to be same as the original count from a priori
By using the formula, we get the MAP probabilities as :
$P_{blueberry} = 0.454$
$P_{strawberry} = 0.181$
$P_{yogurt} = 0.364$

Problem 2: Central Limit Theorem

I have first centered the signals by subtracting their means and dividing by their standard deviation. Let us take a look at all the histograms of the standardized signals along with their Kurtosis values







We can see that the x2.wav seems more Gaussian-like than the other two
We can also see that the histogram of s.wav seems the least Gaussian-like.

Let us take a look at the Kurtosis values of all the signals found by this formula:

$K(x) = E(x^4) - 3.0$

Signal	Kurtosis Value
s.wav	7.2934428
x1.wav	4.907142
x2.wav	3.8733282

We can see that the Kurtosis value for the x2.wav is lower than the x1.wav. That is because the x2.wav contains more sources than x1.wav. According to Central Limit Theorem, the sum of random variables gets closer to a Gaussian distribution. The more the number of random variables, the more the distribution will be Gaussian-like.

Here we can say that as the x2.wav's Kurtosis value is the closest to 0, it is the most Gaussian-like amongst the two. It contains the most number of random variables(i.e sources) and hence is the noisiest amongst x1.wav and x2.wav. Having said that, we can say that x1.wav is closer to the source s.wav because it contains lesser sources.

Problem 3: Lagrange Multiplier

The power iteration gives the highest eigenvector(eigenvalue).
We will set up an optimization problem. We want to find x which will explain the most of the data **A**.
Hence we will say that x will be our highest eigenvector.
If we project our data to our first eigenvector, i.e. $y^T A$, then new coefficients of y should explain most of the data.

Hence the objective function is as follows:

$$arg\ _x\ max(x^T A)(x^T A)^T \tag{1}$$

Here we assume the energy of the dataset after the projection can
be represented by the sum of squares of the coefficients.

(a) **Why is it difficult to solve this equation?**
We can see that the function is the continuously increasing function of x vector. That is it doesn't have a maxima. Since we don't have any constraint on x, x can assume any value and hence the function will always increase with the increasing value of x. It will lead to the trivial solution.
We can say that the increasing value of x can be considered as increasing the length of the vector x and assuming any direction

(b) **Show that eigendecomposition gives you the solution to this optimization problem.**
Let us establish an equality constraint to avoid the trivial solution in the previous case.

$$g(p) = xx^T - 1 \tag{2}$$

Since x is the solution, we can say

$$g(p) = 0 \tag{3}$$

Problem 3: **Lagrange Multiplier** continued on next page...

Let us multiply this equality constraint wwith the Lagrange Multiplier.
Hence the final objective function with Lagrange Multiplier is

$$arg\ _x, \lambda\ max(x^T A)(x^T A)^T + \lambda g(p) \tag{4}$$

Now since we know that the function (4) will have the maxima subject to the constraint, let us try to find the maxima by differentiating (4) with respect to x, λ. We get:

$$\frac{\delta}{\delta x}(x^T A)(x^T A)^T + \lambda g(p) = 0 \tag{5}$$

$$\frac{\delta}{\delta x}(x^T A)(x^T A)^T + \lambda(xx^T - 1) = 0 \tag{6}$$

$$2AA^T x + 2\lambda x = 0 \tag{7}$$

We get:

$$AA^T x = \lambda x \tag{8}$$

Looking at the above equation, we can say that x is the eigenvector for the matrix AA^T. That is x is the vector that explains most of the data AA^T.
Hence the eigendecomposition gives the highest eigenvector which is the solution to this optimization problem

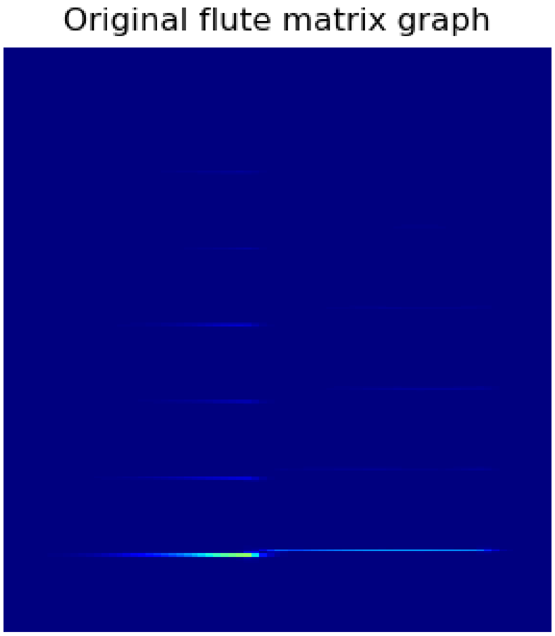
(c) **Justify why the largest eigenvector corresponds to the solution.**
The x eigenvector found in the previous problem is the solution for the optimization problem. That means the eigenvector x explains most of the data for the matrix AA^T.
Hence it is the highest eigenvector for AA^T
In power iteration method, for finding the second eigenvector, we subtract the first eigenvector to get the residual matrix.

$$AA^T x = AA^T x - V_1 S_1 U_1^T \tag{9}$$

Then we say that the second eigenvector best explains the data for this residual matrix.
But from the solution from (8), we can say that the eigenvector, x, found explains most of the data for AA^T it is the highest eigenvector corresponding to the solution.

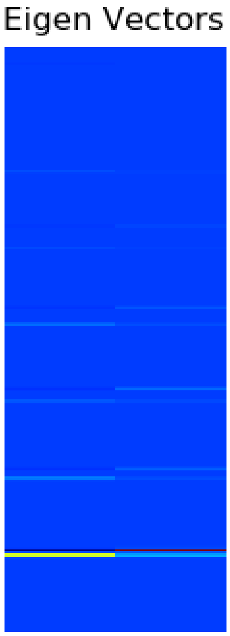
Problem 4: Power Iteration

(a) Graph of a original flute matrix

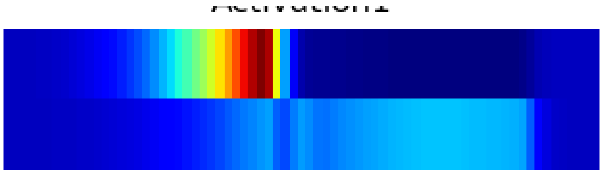


(b) I have computed these eigenvectors for the covariance matrix of dim(513,513) using power iteration

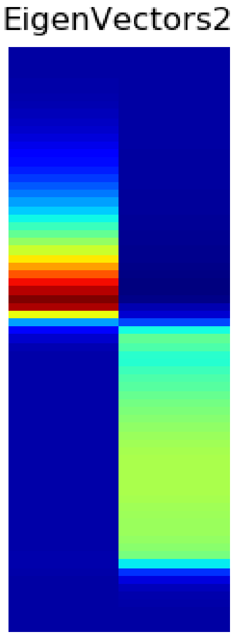
Graph of eigenvectors of covariance matrix of 513x513. This is also called representative spectra for the two notes



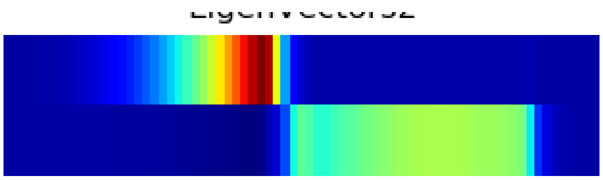
(c) How did we calculate activations?
$temporalActivation = V^T X_{flut}$
Here X_{flut} is the centered original data matrix and V is the vector of eigenvectors. The dimension of the activation found here is (2, 73)
Graph of activations



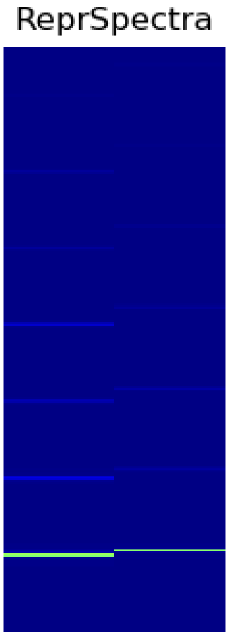
(d) Finding the eigenvectors by another way.
Here we take the covariance of the transpose of original X_{flut} matrix. The covariance matrix will be of dim (73, 73). We again find the eigenvectors using the power iteration method.
If we take a close look at this, we can see that the eigenvectors found in this case corresponds to the activation found in part (c). Here is the graph of eigenvectors found in this case:



We can actually plot the transpose of this matrix and see for ourselves that the previous activation and the eigenvectors found in this case are similar.

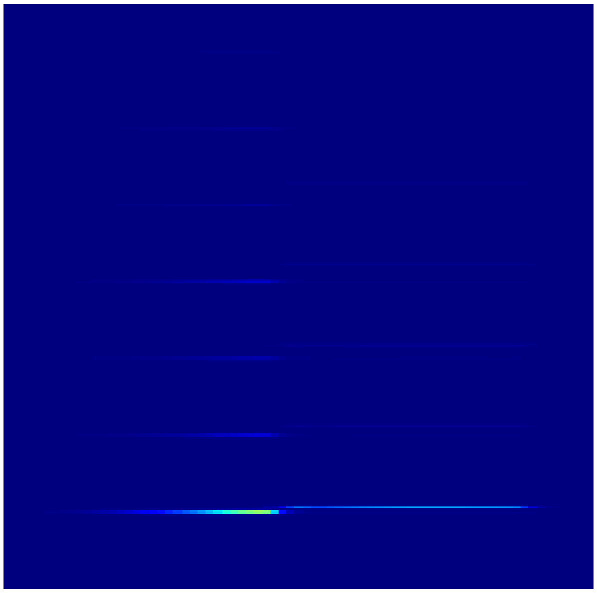


(e) How to get the representative spectra?
First we will find the activation in the second case from the eigenvectors found in part (d).
$temporalActivation = V_2^T X_{flut}^T$
Here X_{flut} is the centered original data matrix and V_2 is the vector of eigenvectors found in (d). The dimension of the activation found here is (2, 513)
Now, we know that this activation corresponds to the representative spectra in the first case. So this is how we find the representative spectra using the activation in the second case.
Graph of representative spectra (found by plotting the transpose of the activation2):



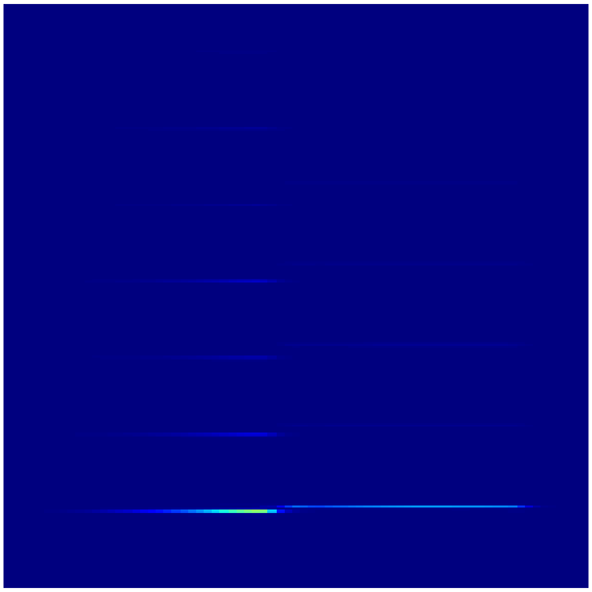
(f) Reconstruction of the data matrix from the eigenvectors and the activations
The data matrix is reconstructed using this formula:
$XFlute_{recov} = V Activation$
where V is the vector of eigenvectors and $Activation$ is the activation found by that particular eigenvector. So we have two recovered data from two cases.
$XFlute_{recov1} = V_1.Activation_1$
$XFlute_{recov2} = V_2.Activation_2$
Reconstructed data from first case:

Recovery1



Reconstructed data from second case:

Recovery2



(g) Comparing errors of two methods using euclidean norm and justify why and which is better I have calculated the error in terms of the L2 norm.
The calculations are as follows:

Reconstruction error of first case: 44.2817882344513
Reconstruction error of second case: 43.347256740573656
We observe that the reconstruction error is less in the second case. Hence, doing eigendecomposition on the transposed version of the original data matrix is preferred.
We can say that in first case, the covariance matrix was of dim(513,513). The two eigenvectors found in this case were of dimensions(513,1). These were the basis vectors which best explains the data. However the dimensions were reduced from 513 features to two features(since only two basis vectors). Hence the data loss is more in this case.
In second case, the dimensions were reduced from 73 features to two features. In other words, 73 features were explained in two-vector space. Hence the data loss was lesser in this case.

References