STA142B Final

Ryan Buchner, David Mao,and Jiaxu He

May 2021

1. (a) Maximize $v^T Av$ subject to the constraint $||v||^2 = 1$. To solve, we can utilize Lagrangian multipliers. Setting the derivative of $L(v, \lambda) = v^T Av - \lambda(\langle v, v\rangle - 1)$ to 0, we find:

$$\begin{bmatrix} \frac{\partial L}{\partial v} \\ \frac{\partial L}{\partial \lambda} \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial v}(v^T Av - \lambda(\langle v, v\rangle - 1)) \\ \frac{\partial}{\partial \lambda}(v^T Av - \lambda(\langle v, v\rangle - 1)) \end{bmatrix} = \begin{bmatrix} 2Av - 2\lambda v \\ \langle v, v\rangle - 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad (1)$$

   The system of equations we are left with is $Av = \lambda v$ subject to $||v||^2 = 1$. This is simply the eigenvector equation, and thus v must be an eigenvector of A. As for which eigenvector, we must reexamine the original equation which we are interested in maximizing.

$$v^T Av = v^T \lambda v = \lambda v^T v = \lambda ||v||^2 = \lambda$$

   Thus, the maximizer is the eigenvector associated with the largest eigenvalue

   (b) The same derivation (using the same equation and constraint) as part (a) is used, except we are now interested in minizing the function $v^T Av$.

$$v^T Av = v^T \lambda v = \lambda v^T v = \lambda ||v||^2 = \lambda$$

   but since we are looking for the minimum, we will take the eigenvector associated with the smallest eigenvalue of A.

2. (a) Since the rows of x are data vectors, then the projection of the data vectors on the vector v is $\frac{Xv}{||v||^2}v$. Assuming v is a unit vector, so $||v||^2 = 1$, then this projection is just $(Xv)v$. The scores for the projection are just found by Xv. The variance of these scores can then be found by $\frac{1}{n}\sum_{i=0}^{n}\langle X_i, v\rangle^2 - (\frac{1}{n}\sum_{i=0}^{n}\langle X_i, v\rangle)^2$ where the X_i are the data vectors (row vectors of X). Note that

$$var(scores) = var(Xv) = \frac{1}{n}\sum_{i=0}^{n}\langle X_i, v\rangle^2 - (\frac{1}{n}\sum_{i=0}^{n}\langle X_i, v\rangle)^2 = \frac{1}{n}||Xv||^2 - \frac{1}{n^2}(\sum_{i=0}^{n}\langle X_i, v\rangle)^2$$

We can solve for $\sum_{i=0}^{n}\langle X_i, v\rangle$

$$\begin{aligned}\sum_{i=0}^{n}\langle X_i, v\rangle &= \sum_{i=0}^{n}(x_{i,1}v_1 + x_{i,2}v_2 \cdots + x_{i,d}v_d) \\ &= \sum_{i=0}^{n} x_{i,1}v_1 + \sum_{i=0}^{n} x_{i,2}v_2 + \cdots + \sum_{i=0}^{n} x_{i,d}v_d \\ &= v_1\sum_{i=0}^{n} x_{i,1} + v_2\sum_{i=0}^{n} x_{i,2} + \cdots + v_d\sum_{i=0}^{n} x_{i,d} \\ &= v_1 0 + v_2 0 + \cdots + v_d 0 \\ &= 0\end{aligned} \tag{2}$$

Since the data is centered, $\sum_{i=0}^{n} X_i = \mathbf{0}$, and thus for any dimension j, $\sum_{i=0}^{n} x_{i,j} = 0$. Now, substituting $\sum_{i=0}^{n}\langle X_i, v\rangle = \mathbf{0}$ into the origional equation, we get

$$\frac{1}{n}\sum_{i=0}^{n}\langle X_i, v\rangle^2 - (\frac{1}{n}\sum_{i=0}^{n}\langle X_i, v\rangle)^2 = \frac{1}{n}||Xv||^2 - \frac{1}{n^2}(\sum_{i=0}^{n}\langle X_i, v\rangle)^2 = \frac{1}{n}||Xv||^2$$

(b) Since we wish to maximize the variance, we are wishing to maximize $\frac{1}{n}||Xv||^2$. We can evalute, giving us

$$\frac{1}{n}||Xv||^2 = \frac{1}{n}(Xv)^T Xv = \frac{1}{n}v^T X^T Xv = v^T\frac{1}{n}X^T Xv = vSv$$

. Since we defined v as a unit vector, we will constrain this maximization by $||v||^2 = 1$, which leads us to the maximization problem in part 1, which the solution to is the eigenvector corresponding to the largest eigenvalue. Thus the direction that maximizes variation is the largest eigenvalue of the sample covariance matrix.

3. Let $X_1 \ldots X_n$ be m dimensional data vectors. KDE is a method to estimate the density function by:

$$\hat{f}_{n,h}(x) = \frac{1}{nh^m}\sum_{k=1}^{n} K(||\frac{X_k - x}{h}||^2)$$

For some positive kernel function K, and $h > 0$ denoting the bandwidth. What this equation is for every point x_i in our dataset, a density function based on the kernel is constructed around x_i, and each one of these densities are then added up then averaged. The resulting density function will have peaks were a large number of points are centered and valleys were there are not a lot of points. h controls the bandwidth of each individual kernel function. A small h results in a narrow bandwidth, and more peaks in the resulting density function, while a large h results in a wide bandwidth, resulting in a smooth density function.

The Mean Shift Algorithm is built off of KDE. The idea behind this algorithm is instead of creating an estimation of the density function, we instead only map a data point in our set to its nearest peak if we had estimated the density function using KDE. In essence, if we choose certain parameters to be the same (which we will discuss later), then we are working with the same estimated density function, only for Mean Shift we do not explicitly calculate this. Because of this close relationship, we can see that many aspects of the Mean Shift Algorithm are related to each other.

First, we observe the update step for the mean shift algorithm. Given a starting point $x_{(i)}$, the next point $x_{(i+1)}$ would be equal to:

$$x_{(i+1)} = \sum_{k=1}^{n} w_k(x_{(i)}) X_k$$

where each $w_k(x_{(i)}) = \frac{\phi(||\frac{x-X_k}{h}||^2)}{\sum_{k=1}^{n} \phi(||\frac{x-X_k}{h}||^2)}$, and phi is a positive function. The difference between $x_{(i+1)}$ and $x_{(i)}$ is called the mean shift vector, which is equal to:

$$\sum_{k=1}^{n} w_k(x_{(i)}) X_k - x_{(i+1)}$$

We will find that if we choose for KDE $h = h$ and phi to be the derivative of our kernel K, the mean shift vector will be equal to the gradient of the density function estimated using KDE at x_i. This is actually the intuition of Mean Shift. Given a peak in the estimated density function using KDE, if we restrict this function to only the peak and its surrounding slopes (i.e the places where if a data point were to exist there, then we want to map it to that peak), then the gradient of this function is always going up hill, and will eventually reach the peak.

$$\nabla \hat{f}_{n,h}(x) = \frac{2}{nh^{m+2}} \sum_{k=1}^{n} \phi(||\frac{X_k - x}{h}||^2)(x - X_k)$$

Simplifying, we get:

$$\nabla \hat{f}_{n,h}(x) * \frac{nh^{m+2}}{2\sum_{k=1}^{n} \phi(||\frac{X_i - x}{h}||^2)} = x - \sum_{k=1}^{n} \frac{\phi(||\frac{x-X_k}{h}||^2) X_k}{\sum_{k=1}^{n} \phi(||\frac{x-X_k}{h}||^2)}$$

We know all ϕ are negative, so the above simplifies into:

$$\nabla \hat{f}_{n,h}(x) * |\frac{nh^{m+2}}{2\sum_{k=1}^{n} \phi(||\frac{X_i - x}{h}||^2)}| = \sum_{k=1}^{n} w_k(x) X_k - x$$

Therefor the gradient of the KDE density function at a certain point is also the mean shift vector.

In conclusion, if we choose $h = h$ and phi to be the derivative of K, then:

1. The mean shift vector direction from a point is the same direction of the gradient of the density function estimated by KDE
2. The mean shift path is approximating the path along the density function estimated by KDE from a data point to its closest peak.
3. The endpoints of the mean shift path are just the peaks of density function estimated by KDE.