Video 3:The Mathematics Behind Principal Component Analysis

wbg231

December 2022

introduction

- vedio link
- today we are talking about the math behind pca

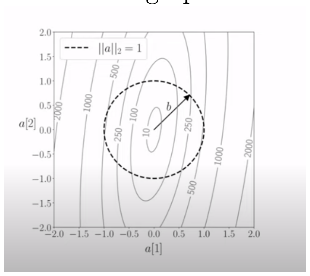
pca

- we are going to focus on PCA of a dataset (but we showed that the same thing holds for a random vector)
- the steps of pca for a given dataset X are
  1. compute sample covariance matrix Σ_X
  2. do the eigen decomposition of Σ_X to get Principal directions $u_1 ... u_d$
  3. center the data and compute Components directions by projecting our data onto each Principal direction that is $w_j[i] = u_j^t ct(x_i), \quad \forall i \in [1, n], j \in [1, d]$
  4. where $ct(x_i) = x_i - M(x)$
- this allows us to find the directions of maximal variance, as well as the Components of our data that capture the maximal variance
- he then goes through an example, which we did in last Video so i am not going to write it down again

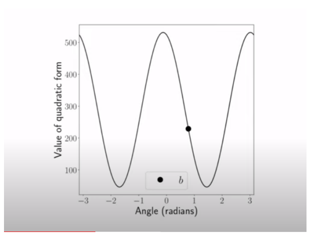
maximizing variance

- so recall given a dataset we can find the variance of our dataset in a in a certain direction $b \in \mathbb{R}^d$ as the variance of our dataset projected onto that direction a
- so our dataset X projected onto direction a is $X_a = \{a^t x_1 ... x^t x_n\}$

- and thus the variance of our dataset in that direction is given by $var(X_a) = a^t \Sigma_X a$ which we can call $q(a)$
- we can look at this function over the set $A = \{a \in \mathbb{R}^d : ||a|| = 1\}$ that is $q(A)$
- which when graphed looks like this along the contour lines of some dataset



- then we can look at the values $Q(A)$ it's self in



- so how can we maximize this function representing variance in every direction?

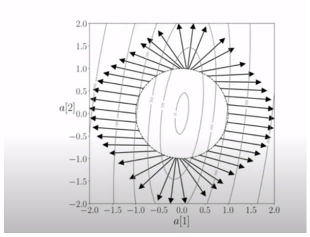
will there be a max

- first we want to show there must be a maximal value for this function
- first off we know $q(a) = a^t \Sigma_X a$ is continuous
- and we know the set we are looking at (A) the unit sphere is closed and bounded thus there must be a max by the extreme value theorem
- t

what will this max look like?

- to look at the max of $q(a) = a^t \Sigma_X a$ we want to reason about it's gradient
- we know that Σ_X is symmetric thus $\nabla_a q(a) = 2\Sigma_X a$

- that looks like this



- the gradient encodes the directional derivative of $q(a)$
- the directional derivative of $q(a)$ in the direction of some unit vector b is given by $q'(b) = lim_{\epsilon \to 0} \frac{q(b+\epsilon h)-q(b)}{\epsilon} = (\nabla q(b))^t h$ so it is the gradient of the quadratic form and h
- so in other words $q'(b) > 0 \Rightarrow q(b + \epsilon h) > q(b)$ for some $\epsilon > 0$
- at the max u_1 we can not have $(\nabla q(u_1))^t h \geq 0$ for any $u_1 + \epsilon h$ in the constrained set (So in this case on the unit sphere)
- so we can not stay on the constrained set if we move in the constrained set

tangent hyperplane

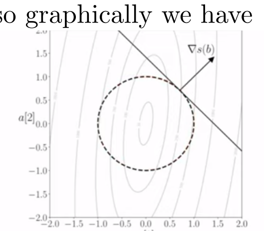
- the unit sphere is a level surface of the function $s(a) : a^t a$
- so the unit sphere is the set given by $A := \{a : s(a) = 1\}$
- we know a vector $y \in \mathbb{R}^d$ is in the tangent plane of A at be if

  $$\nabla s(a)^t (y - b) = 0$$

  that is saying that the vector y-b must be orthogonal to our gradient at b
- this is important because we are in this case because as we are moving we are almost staying at the same value of $s(a)$ in this case on the unit circle
- so we can then write if y-b is small ie y and b are close

  $$s(y) \approx s(b) + \nabla(y - b)^t$$

- that is we are almost staying on the circle
- so graphically we have

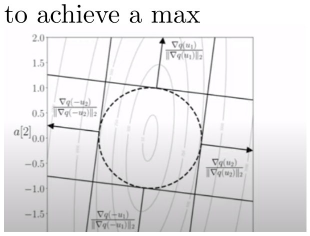


maximizing q

- so given what we showed above when can a point maximize q(a)
- - so for h such that $b + \epsilon h$ is on our tangent plane and we have

$$\nabla(b)^T g = q'(b) > 0 \Rightarrow q(b + \epsilon h) > q(b)$$

   - for some b in the tangent plane, then we can use a taylor approximation to find a a $y \approx (b + \epsilon h)$ where y is on the unit cricle
   - this means we can move within our constrained set (ie within the circle) and increase our function value (so that point can not be a max of q(a))
- so when will a point be a max of $q(a)$ there should be no h such that $b+\epsilon h$ is in our level set an $\nabla(b)^t h = q'(b) > 0$
- so in other words we need the gradient of our level set to be orthogonal to our hyperplane and thus colinear with the gradient of $s(b)$
- so in our case the gradient of $q(s)$ $\nabla q(s)$ must be orthogonal to the circle to achieve a max



- this is equivlent to having a point u_1 such that $\nabla q(u_1) \parallel \nabla s(u_1)$ ie $q'(u_1) = 0$
- so this tells us at the max/min of $q(a)$ there must be some scaler λ such that $\nabla q(u) = \lambda s(u) \iff q(u) \parallel \nabla s(u)$
- we know that $g(a) = a^t \Sigma_X a$ and thus $\nabla q(u) = 2\Sigma_X u$
- and that $s(a) = a^t a$ and thus $\nabla s(u) = 2u$
- so we want $2\Sigma_x u = 2\lambda u \Rightarrow \Sigma_x u = \lambda u$ in other words u must be an eigenvector of Σx
- and clearly that is maximized at the largest eigenvector u_1 corresponding to the largest eigenvalue λ_1
- this establishes the spectral theorem, which we used in the lsat Video