### ML\_week8\_lec02-20241101

说话人1 00:00  
At all, is that the problem was leave the city to zero, and they have the following derivative. This is a principal one. Then new we transpose we want derivatives that we will give you newly one. And if we, if we ignore this, again, we arrive at the again when the equation, but here we have this extra term, uv one. How do we do about it? In a central fashion to prove that also validity as in the independence we get tomorrow? We won on both sides. And this will give you two. We went ask them away, - 2 number. We want to transpose we. Last new. We want to transpose we want. V one transpose v one will give you one, and v one transpose v will give you zero. For manager assumption. And therefore, the certain becomes new. The same number becomes zero. And what's v one transpose? Sigma. And here, v one transpose sigma is equal to sigma transpose v one transpose. Right? The sigma is symmetric matrix. This is equal signal with one transfer. Why signal is equal to is a symmetric matrix? Remember, the definition of sigma imperial grams.

Therefore, sigma transpose is equal to one over and signal time x transpose x transpose. This is the one over, again, x transpose x transpose, which will give the x transpose x what's sigma we want? We want is I have a better term. Therefore, sigma vy is equal to sigma one, v one transpose. V one transpose is equal to zero. Therefore, the first term is also equal to zero. We have only the third term. New is equal to zero. New is good, or we can get a little bit better term, then we will arrive at the eigenvalue equation again.

Then the optimal icon matter is icon matter associated with the second, not just like that, because we have already achieved today. The animator is an artist as we want. So it's easy to say that we start, as I remember, response a signal to that matter. In general, talk a direction of errors. We are oki came by not given by the key. I can first correspond to the key 92nd matters of this imperial the mass matrix. In assignment four, i'm going to ask you to prove this general case, so which is similar to I enter the second one. And the pc can also be directed by picking the principal managers on minimum of approximation error arising from projecting data onto the case dimensional subspace spot levels.

Here shows the example when k is equal to one. I think the first principal vector by minimizing the observation, they can understand this.

Again, by you mean, this figure, this is basic term. It's about datas, because someone excited. This is c and here we project the exam to be and this point is far as the projected data. And this, if this is a vector and this vector is the michigan, we have already proved on this lens is we transpose with xi it points to the same direction. We this is exam, except r is equal to v transpose exam as assigned magnitude. And v is a direction.

Therefore, this of submission area with area inventor is simply xi of manners. We transpose xi so that's is that. And this is rejecting it, and they are sold to each other. Therefore, due to the possible zero, right? Due to the possible zero, abc a squared plus b squared, c squared. Here, c squared is equal to it looks more squares of the exam. Excise fixed. Excess is official given to you, therefore, cd space.

And in the previous case, we want to maximize ac six is corresponds to minimizing d that's this one. Right? So that's two links to be pc one from the how's the manner perspective, money from the minimize the approximation error perspective. Where is the present on this side? They're only have a dimension. I get your document. You say they inform of the first present our data subserve, the mean of the data set. The first pc is the direction that explains the most valuable data. The second pc is the direction perpendicular, or also to be one that explains the most remaining gaps. The surface, basically the direction perpendicular to both divide the two. That explains the most steps.

Dimension reaction is, we say, the formal arguments. So did the preposition accomplish the mean? It will be replace s with these d mean version, then to compute the given the pre center data matrix x we come to the sample to learners matrix, one overhead x transpose x then we compute the kdi vectors between the key of sigma. We are realizing that much better.

Then I can make it together in between and comes key matrix, where each column of I each column, I will be response to ai then we now have it in the projection to the k dimensional space, the particular matrix into round case of space maps and variance that completely in this matrix product.

Here, x is n by n these n by k therefore, these n by k originally, each data is represented by n n dimensional space.

After this project into the each of the end is theta one is represented by the key features. Therefore, we refuse the damage analogy from n to take prime, make use of this projection, and to reconstruct x we only have the project, indeed, right?

And leave the basis vectors. The only way to construct it is to do this work is to do this visible multiplication, where z is n by kb is n by kv transpose is k by n then we have the matrix of n by n which matches the dimension. I do this measure product and modification. We reconstruct, we protect from key dimensional space in dimensions space.

And then you can compute the reconstruction here in matrix form. Since since this is equal to x minus, what's x hat? That's the we transpose. What's the xv we need to minimize this subject to that. Each column raise a solid paper, and it has to be this one. This is the object of the magic store. And here is a for example, on block data. This is the data in the original space. This is the first principle component. This is the data in the project space. Here we have the first principle component saving principal component, which are ashamed to be possible to each other. Right? And then we really draw the data in the project space. And in this case, since we don't do damage and reduction, but when and the project of two dimensions is for months to rotate coordinate axis, are there any examples on every state?

Now? We can protect two d twenty, forty, two d this is the first principle couple of second principal couple. And they show the data points in the projected space, how to choose the number of principal components. Basically, two methods to choose this case. The first is to increase, preserve some percentage of the values in the data. For example, to do is 95 % the sickness, whatever works for better task. I will show the best learning task as occasional reaction. And in this case, that mention and reduction is a pre processing step for this upstatement is not a president ask. That's in someone have written digit data. So it has this number of images and ten categories. It's ten class classification program. And each image is 8 × 8, and we can flag it into a 64 dimensional way. There is some representative examples with split the data into training and testing set. We run pc on training set, apply the identify physics vector. This matrix be on the 27th and apply test.

So as always, the test data is always not used for testing. It will not be seen by the machine learning articles during the whole training. He runs the top 25 pieces room. And here you see a little bit dramatic structures of this was teaching about pieces. And we can visualize the coefficient of the first principal components coefficient. This term. First is the seven pc and you see a little bit grouping of different digits, image of percentages as you could write in cars boundaries. So you can overlap among those ten categories and matrix and expended balance each. Each pc extend the percentage of the original data. This is called extend balance. Pieces are already sorted by expert matters from high schools. Believe the number of pieces to get a certain percentage of experience. Man explain that as a typically 99 %, how to compute this issue as a percentage?

The percentage is the right? We have an integration. We have only the number is total among the grams. Explain what outcome to total balance. Do you remember? We have sample. We have sample of balance, right? If we insert the ultimate basis vector, honestly, I can make a response would not take matter. The subtle variance becomes the, I can imagine this equation pizza, the optimal manners asking manners is equal to that measure. Therefore, the total amount of balance to be explained is simply the sum of all I can measures. This is total balance. And to configuration, we need to identify k of sigma, k this decay. Right? It increased from 1234, and this ratio increases until it reaches this usual manufacturer on that. This is how to compute the percentage of explain balance is the ratio of the I can minus sort of I have to select when total the sum of the total patent matters use result on the final task.

In the case of classification to select the number of components that is to treat those as I have a plan. No need to rerun pc for each number of components. We run pc once for all, and we just selected in our subset p cs as come on. This is because patient entry is stable after using 20 pieces. So the multinationals in performance in informing 20 pieces by the people.

So another and pc versus svd so we have sold out the minimum obedience, nonlinear dimensionality reduction problem to be solved using the wrong case with the originally here. This is, wow. Right? The optimal man is the svd it's svd which is you is the conference of overall columns. You transpose use I s title matrix, va transpose, v is equal to I and this matrix, these could you pass? Gives the open, rounded representation of x perspective for innocent limitation. Right? You times ksk times k will give you n times k each data one is represented in ak dimension vector or internal axis. N by n reduce n to k if we let k is equal to, we don't do the dimension.

We are simply do simpler, valid inflation. X is equal to usv transpose, then x transpose x is equal to. Here, x transpose is equal to you as a transpose rtv we. S transpose u transpose s is the that momentum. That is b as you transpose. Here is bsu transpose, you speak. And since u transpose u is equal to identity, due to those another view, this gives the x transpose x is equal to b as squares va counter, which one?

And then question two, they have zero interest observations. This means that the red, singular effort, x that is free, and exactly, I can be interested in transportations. So those are the same thing. We can also sit up the eigenvalue of x transpose x not x squares are the squares of the diagonal element to s I can make terms of the response with x is squares of the singular manners of the data metric sevens.

This means that the key, largest singular values and the key design, the values correspond to c a basis vector. That is v v vk red, singular vectors add values. According to pca the protection operator is z is equal to xb z is uxvx is equal to usv transpose. So we have the same way, right? We transpose its identity, and they see the us finally, no time if the composition based upon the key principle vectors, u transpose uvc by ntv transpose space by ntv remains the same, remains true. Therefore, the projection of the is equal to x times v by pca and z is equal to us bias. Svd will still be identical and the same. Those manipulation shoes are pca and transpose x and svd on x identify the exactly the same subspace resulting. Is that the same projection of the data in the substance? Right? Maybe in the final time, you will compare, you will have this concept question.

What's on the Similarity and difference between pc and skt you need to answer that pc on this is equivalent to svd of x simple. As a result, generically, in your dimension and reduction simultaneously minimized for business loans, both in the construction area. That's a mask methods to be and balance in the narrow subspace. And both svd and pc will provide a safety creation of this generic and near damage and reduction. A solo basis for exact the same optimal subspace. That is viv two, the uk which is the same red, singular vector of x as well as the eigen nature of x transpose. When does pc feel? So? Pc is the primary motivation behind pc is to de correlated, de create a data set. I need to remove the second order dependencies. If higher order dependencies exist between the features in the data, pca mean insufficient, that would be revealing all started with it. And here is a simple example. So this is tracking or person fears will. And this motion can become completely characterized by the face of the will not see.

Now. We see how is highly non linearly depend on this coordinate axis, expensive, higher line two, the order. Therefore, of ipc will give you the first principle of this direction, the second principle on this type building, which is meaningless in reviewing the structure, which is what we did. And since we said only curious first order statistic, that is the mean, we need to be mean. And the second order statistic that is covariance matrix, right? Byebye. The only curse of arts, the same ones, personal signals that is, therefore, in the only it will work perfectly, give us it, which is sample or generated by a probability, is looking that is completely characterized by this one first order and second order statistics, which is also is really becoming company, represent me better.

And the raspberries. Therefore, pc will work perfectly complicated. And pc requires each component to be perpendicular to each other. If someone asks you why the p cs results from pc are also to each other, so you should answer. It's the assumption made by pc the number should also looked up. But this assumption is overly stringent. So the data may be arranged a long or solar axis, such as this is as the optimal direction, shall be this. Right? I'm also looked at, but he said only give you a social business. So in order to generalize pca we have another important argument that not has already been cut from schools in order to make more for deep learning entries. Not. I say it, I means independent the same, say, independent independent company. That is so part of the independent company that is essential pc it has some edge competition and it works perfectly for this type of thing, which won't have this particular function.

Now, let's move to the nature 8 . 2 kernel principle, common analysis. What if the data leaves on non flat surface that fix? Pc cannot capture the curvature of the data. And purple points will be close to each other. After this protection, red points will be close to each other on project as well. Here is more clearly visualization. We will have the eyes of counter piece of projection forms on the same dash, nine object or sympathy coefficient. And you see, what time do you see on this nonlinear structure? Will won't preserve geometric structure, have a much more specific how to project to a non inner surface. As always, you can treat mapping, right? Will apply a high dimension figure transform of the data, exactly our website, or take higher dimension data to an inner surface on surface, that therefore be one pc on this transformed data matrix. And to this linear eyes, we have linearly eyes of control by purchase this to the original space. It comes from this market signal theorem as previous classification with random networks.

And therefore, we have a facial map and pass svd given a data set and manner. Our feature mapping from n to l where l is generally larger than any we all kind of falling svd based on those, will buy svd on the transformed these matrix, get usv we return the project datas matrix as new concepts.

And here, there is a deep source behind this kind of operations in order to reduce the dimension. You need to first expand on that and arrange on the end. In order to reduce the dimensionality k we need to first expand the dimension to l so it's not the dimensionality. And the future man capacity, given a data set, a feature map, and we open the following pc elements, we compute the covariance matrix in the future space, and where mu is the mean of the future.

And after from this covariance matrix to the computer kd and vectors, where each way now becomes ill by one vector, because the growth is much higher in space.

As we stand the kind of efforts together to form me called my special. Where are these like it? And then we project the matrix to run case of specific amounts and barrels that compute the matrix for that. And here is the megmu represented by l and here we have l by k which will finally give you in banking. What else for tonight possible city. So as in classification is become very expensive to use the explicit feature mapping in the basic svd based on the person, no need to avoid to explicit to the big man. But in pca based on them, we are able to take advantage of the purity and to progress it.

And until now, we have learned at least three parts of this program, right? The first one is, I swear, from travel to deal the deal form, has something to do with your product and therefore realize it. The second is greater regression. Remember, the cross form solution for ordering experience, x transpose x inverse, x transpose one. This has nothing to do. This has nothing to do with minimum products. Then we make use of the asymmetric identity. I want you to put me assignment three. Remember, this will give you x transpose, xx transpose. This is the inner product. Make realize it. And the third time they are supported regression from due from crime on the floor to deal with business is in the former inner product, realize it. Center. Here, we want to review the through the manipulation that depends on in your product and realize it. And here this is currently saying, given feature mapping, we compute the covariance matrix in the new feature space.

Here we make assumption that so it's pretty center in the new transport space, but this is not generally true. Here we take it for granted, then we will describe how isn't it without explicit computer transformation. This is the current metrics. And I can identity conversation of these metrics given is given by this is sigma vk is equal to lambda, akk is equal to one up there. Right? So that's, I I wish here, sigma, we pop this into this corner, which will give you this component.

It's more hard to set up if it can be expressed by which the combination of transform they don't wants. Where are the rates? Sequel this? Right? This is to move number eight to here, number eight, and move one over. And now the key to here. And now the key, this is you, jk it's a simple. And then we copy these two equations to the next time. Vk is the in learning combination, weighted completion, or data points in transform space, where wjk is one over mrk via exchange transportation. That is this kernel, pc principal component is an internal combination of high dimensional. And wjk are the ways to be determined, which depends on making. Once we determine w we can compute the case, this formula, as well as this formula. We left multiply phi xi to cosine. We have here. This equation is a result from this equation. We left multiply five xi transports vk which will give you j from one to mwjk five xi transpose times five xj this is equal to this, comes from this equation where we have phi xj transpose the uk is equal to wj am lambda k here, we simply replace the index from j to map.

Therefore, phi xi transpose vk is equal to m lambda, kwik that's it. Therefore, we have a number out. This is number q four. Therefore, we catalyze it, replace this with a curl function. This is data points, the inner products between to the organization and transpose advanced kng therefore, here we have the summation. This becomes kij and wjkjk and copy this. On this. You get this equation. It costs on the generation. So each step are just simple domain, a simple after an equation. If we consider I where is isqsq if we consider I is equal to one to m the above scared equations become the ice component of the falling venture equation, kwk and ww where ws w one, kwwk right? Here is recently taking the case, kiss rule by kiss rule, the ice rule, the ice rule of this metric kilometers key.

Therefore, by looking at this equation again, this is eigenvalue equation. Wk is a case. I can manual of dk the kernel matrix. And then number k is the eigenvalue of k which is proportional to the eigenvalue of the covariance matrix in the future space. Originally, this number k is here, is added value response to the sigma. Sigma is in the new transformation space, new feature space. Therefore, she says those are proportional to each other those numbers. Therefore, pc example, a rare symmetry is in a transformer space, is equivalent to pc on cable as he was. I will maybe make it clear in the following durations. Therefore, for your data ., the keys pc can be opened by project five of a star onto the keys, item value of weight, right? This is the original argument later on, both of sigma. And we do this in the program. We magnify to make it. This is fine, is a linear combination of high dimensional vectors. That's a linear combination. This can be completed as a curve function, and have already identified this through solving this item medications, then valuable to compute this case, principal component.

Therefore, here is the kernel pcr of them, given a data set m x is m by ni don't have a kernel function, k the kernel pc can be computed as follows. The first compute the kernel matrix, which is kernel function to xij this is north center, because even if the data is precinct in the oriental space, after perform the strict matter, the data set may not be present, in case, had mentioned in space.

And then how do we go about it? We can use this to present the data set in the new future space without explicitly computing the feature transform. By relating ktok prime to k with zero center data points in the future space. This is the current matrix of the center data ports in the future transform. And how to arrive at this, I believe the details, the measure of these are the final piece of the current principle performance must be sent to data in the customer space without instant computer transformation.

Then we can choose the k matters w one and wkok prime along with im minus and number one and number k we compute the case principal components of the project data using this form. And each w one, wk is an dimensional space. Therefore, the sum is over n this is the or, in principle, coefficients or decrease. In principle, here's the example using polynomial cologne. Now, purple points are far, further apart in the project space. This pc response to locations for number eight occur here. This is zero and a number of curve. This is distance of our curve. The second pc corresponds to the distance from the center. Here. This defines the center. The points on the dash line will have the same principal coefficient in the same principle directions. Here, these two figures, we are drawing data point on in the original space.

Here we are drawing that points in the project space and use clearly see the first pc Is the location around the data curve. The second PC is the distance from the center. And you see the same as symmetric, a potent semantic structure. The principal component analysis with raf kernel, the coefficient is a distance to clusters. Here we have two examples, three clusters. The first PC can split the data into three clusters as follows. A kind of data points perform the coefficient efficient. Here, this first pastor and second pastor is the second is so why are there for response to ask them? Because II heard also exams the in creative business, right? You can stop, right? And in a very much single favor time as k BS is also relies on the Beijing distance.

That's a high level relationship. Why not be a very corresponds to masters? When you perform damage itm here is the some of these images, 8 by 8 flag to 64 dimensional vector. And we use current PC coefficient as a representation. We can turn a logistic regression faster. We can try different number of components. Again, once before, we can only perform kpc plus, and then you can select a different subset of principal coefficients. Here is a classification result. Kpc improves the performance compared to PC is not just regression or pixels with no discretion. The current PC uses the currently perform PC height under space coefficients are based on a number. Rejecting the data, perhaps projection is based on current functions selected. So using up your proper, or if you think has been in clusters, uncouraged, it can provide effective previous step for last message, as well as leaders as different regressions.

However, as this all rural as that personal master that computational proceed, I expensive because the size of the matrix to be composed is on the order of n square, and is a number, of example, in 27. If you make sure 8 . 3, we have decent step by step implementation for principal component analysis and curl principal analysis. If you don't understand the math behind for the principle of person, there's no worry. So check this step by step, then go back to the relations. Right? In karaoke city. We do implement this one took place into the data without these big, competent transformation in order to compute this parallel metrics, in part. That's all for this lecture, guess next week.