* Trees
  + How doe it create randomness
  + Why is it random
* #2 want to build a different features they're also built on different parts of the data set OK so that's why they're called random forests so it's corrected their forest but No 2 trees with some you might you might get two of the same 'cause it's random so whenever it's random there's a small probability that two things will be the same it's called the collision.
* And then the other alternative is gradient trees what do you need to know about those what was the defining thing with gradient boosted trees learn from the mistakes of the previous situation itself and corrects it and then fills new tree and corrects
* Classification OK final thing before we move on so accuracy we use to measure classification and as I mentioned we sometimes have problems if I just classify everyone is not a terrorist I'll get a high accuracy so we have alternatives which are called precision and recall we should be able to explain these also also the confusion matrix it has we're only going to do it for binary so it will just have four boxes true positive true negatives false positives false negatives so what's the false positive case someone exam cheating what is a true positive yes I predicted that you cheated and you actually cheated so I should identify you as a cheater yes what is that a false positive I what's a true negative yeah I say you don't cheat and you actually get cheap false negative you are Siri you got it so you need to be able to explain those in terms of but I gave you like we just did so probably give you some tragic thing the confusion matrix use the case maybe I'll say back clovers should you get alone or cor yeah what so the precision gives you the correct positive predictions the recall gives you it detects how many positives out of the so I'm guessing right finish before so break time.
* You should know the basic concept of wanting progression so regression means you have a Y variable that is yes continuous and what you're trying to do is the best model for this data and there is only one best model so you will always be able to only find one model so why is so so what's the basic idea so the basic idea is you have a bunch of data so let's just do it two-dimensional something like this and then you ask the question which line and here in this case it's just that one excellent could be multiple but it makes it very difficult to illustrate so you're trying you're asking the question which line minimizes the error the most so So what is an error so let's say we have this line here this is this is candidate #1 so an error which is also sometimes called a residual is the distance between the clear distance between this point and the prediction so you should picture picture of case where you have some X and then you're asking for this value of X here you have a prediction which is this part here and you have an observation which is this part so you're trying to fit the line that minimizes all of these errors and then the acute student will of course say while this distance here is positive whereas this distance is negative so I might just come up with the errors are 0 on average so in order to come back overcome that problem we actually don't take the arrows we take the square of the errors so you make all these small boxes here this you get the idea the error and you add that that gives you the sum of squared errors and that could be really useful but the problem is the sum of squared errors what will happen if if I get if I have two data sets and one is larger than the other the sun will be bigger wouldn't for one and the other of course the more observations at the more of these boxes are without degree so we don't actually use the sum of squared errors we use the mean of the squared errors that's called the MSE that's the thing we're trying to optimize So what we're doing our optimization problem whether it's gradient descent or whatever we try to to minimize the mean squared error so we tried to make these boxes as small as possible now then picture another line so let's say I now have this line OK the distance between these two so if we call this observation here why so this point on the line here let's call this Wyatt and we call this one just why like regular live without makeup sorry then we can say that this is basically y - y had square some divided by 8 basically no value minus product that will call you yes its real value minus predicted value that's the receipt now there is another element here which we're not going to talk about in discourse which will talk a lot about it which is the one that called the base case'cause if I took all the Y values there's gonna be one more line remember there's another model the base case for regression around we just discussed this two hours ago so the mean is just some constant here that says this one yeah so this also plays a role but we don't need to discuss it but yeah so essentially we have that the point is it actually turns out we have three different errors we have this error we have this error but we also have this one this error here so the distance between the observed and the mean we have the arrow which is the distance between the mean and the predicted and we have the error which is the distance between the predicted Sir those 3 errors and all of them have different meanings and the reason why that R-squared can become negative sometimes is this error here it's too big it's actually bigger than all the other parents as you can see here it should technically be smaller if it's not on the sums of all those squared errors which is just dividing some by it and there's only one there's only one best fit line always To Do List there is the most elegant way which some of you I know having your algebra the normal equations that is the analytical way of doing it that's basically where you take your data matrix X transpose it multiplied by X in order to obtain the covariance then you subtract sorry you concatenate it with X transpose why and you reduce it to reduce agile so that's a very simple method which you can use here and it was used for many many years but for large datasets this is computationally infeasible OK so the other way to do it which is gradient descent but again because of larger and larger datasets gradient descent is also too slow will take too long to do gradient descent or regression so the way in which it's actually done today in Python is with something called the singular value decomposition which is a very very smart way of working with data that divides the that separates the data into 3 what should we call it parts picture data all of your data and rearranges in a specific way that makes the computations much much faster because one of those mattress is that you obtain is going to be a diagonal matrix which means it has only values on its diagonal and zeros everywhere else and you know what computers like they like zeros it make computations very fast so this one here is the one which is implemented today just make sure you so the last function becomes this here which I started yeah any questions for penalty how they work no but you need to know that there is a penalty so if I ask what is rich regression so you should say something about it's the same as normal regression feature selection 12 it can effectively do feature selection but it doesn't do it in the same way that lasso does which I think is the next thing maybe I'll skip that yeah so so rich effectively does feature selection lasso actually does feature search because rich can only drag the proficient's close to 0 can't eliminate them whereas less open driving proficient all the way down to 0 meaning that feature is irrelevant So what does that mean it means so if you have a model here so it's just fuck the idea is the bridge is able so you have all these coefficients let me go back to this one here so the task is to figure out what are all these cool fishing so each X is a feature so X is a feature X1 is a feature X2 is a feature XN as a feature in each one of these tests are is the thing you were French predictions so the coefficient here rich is not able to make this 0 but it's able to make it very close to 0 meaning that X1 will effectively have no influence on the prediction hey disabled able to drive this parameter here to 0 actually removing the effect of X1 so there's a difference between whether this is really really tiny or whether this is it's really really tiny X1 still plays a role but very very small whereas if it's 0X1 plays no role so that's why rich effectively feature selects whereas lasso actually now if you scaled your data before doing written lesson you get another piece of because then the coefficient will also tell you how important are the features because if each of the features on the same scale then each coefficient tells you how important each features high coefficients means creator emphasis is given to that feature lower coefficients means lower emphasis is convinced so if the data is scaled rich and lasso also gives you feature importance and that's actually the feature important method in Python uses Richard yeah sorry only use one feature I don't think so no simple linear regression can only take one feature that can use like multiple features classification no there's something called multiple regression where where you're allowed to have multiple features and takes from 1:00 to infinite market OK so that's regression there are all of these different values that you can check for and I mean the ones that you most off to look at is the mean squared error there's also the root mean squared error which is just mean squared error and usually the one we're going to talk about is MSE and r ^2 now the problem with MSE is you can't does not like a general measuring stick it if an embassy is 3554 then not look like there's no like objective measurement what is a good embassy there's only intersubjective measurements so if you do one model that doesn't NEC that's lower than the other one then the lower 1 is better but you can't there's no generalized metric whereas R-squared you do actually have like a general stop that if you have an R square that's greater than open 6 really good that's a good model which is greater than open 8 exceptionally good and R-squared tells you something about the explained variance which there's also something called the explained variance and that's what you just need to know R-squared is the explained variance there are small differences so if I if I give you if I ask you so let's say you have an R-squared of 0.67 what does that mean what's the meaning of that number it means that the model which you've created can explain 68 + 67% of all the variants in the data set and if variance means information an R-squared of 0.67 effectively tells you that your model would you built can account for 67% of all the information in your data set which is pretty good pretty good so that's what that's what that means you should also know what the number means so it's just the explained variance that means the 23% plus 10 is unexplained so yeah so I mentioned this so emotionality reduction we already talked about feature selection model regularization the last way in which you can do dimensionality reduction is not existing features in smaller number principle component analysis or dimensionality reduction data compression so this is data compression so PCA basically what happens here is you look at some data set and you try to map it we tried to create a new coordinate system so you have this regular coordinate system which you know you used at X here and you apply here what PC does is it creates data or sorry creates a new coordinate system for the data first access will be the same as the axis where the data has the most variance because that's the most information and the second axis will be the dimension where the data has the second most variance and the third axis will be the dimension where the data has the third most variance and so on and so forth open till you reach end so a principle component analysis if you if you if you start out with N features you will still end up with an principle component OK it doesn't actually reduce the dimensions it gives you new dimensions but it gives you the same amount but the difference is now very very big so component analysis what it does is it takes all the original features it takes height weight and nationality and hair colour and all different things yeah puts it in a big bag mathematical bag shakes it and takes out new features and those features are called principle components and they are structured in such a way that the most information is placed in the first feature which is called PC one the second most information is located in PC 2 and so on and so forth so why is it compression what why why do we call the compression well the compressor part is that we don't take all N features we only take a subset of the of those OK so if we took all in principle components we wouldn't have compressed anything who would have the same amount the same data size you see no the compression occurs when we take a smaller subset and in your case you took two for the plotting and that's usually what we do we also want to plot stuff so we just take two principle now it's eight you were able to print 2 principle components or three print how much variance they explain there's like a rule of thumb that you should choose the amount of principle components that accumulates more than 70% of appearance so if PC 1 so you check so there's a hyperparameter you could put in which is ratio of variance explained let's say you put in open seven and it returns PC1PC2PC3PC4 that means that the four first principle components together account for 70% of all the information in the data so that means that say that there's 104 features the last 100 features only account for 30% so you basically reduce the file size I love yeah by 96% So what is the difference between components and the original features while the most important thing is if you did a correlation matrix of any data you will always have some correlation always some of them would be highly correlated or very little course but none of them will be 0 if you take a correlation matrix of all the principle components you will get identity measures only be ones on the main diagonal and you will have zeros everywhere else OK because all of the principle components are uncorrelated so that that's what it does it takes all the features gives you similar features but now all the features are completely uncoordinated meaning they don't have the same excellent explain different things you see that's what correlation means it means something has high correlation it kind of explained the same thing number of bedrooms number of rooms highly correlated but if you put those two through a principle component they won't be correlated anymore so why is that well that's because PCA and PC 2 and all the pieces are what we call orthogonal serve dot product is equal to 0 OK So what you what you actually do is you create you take your data matrix and create what is called from orthogonal basis so you basically create a new set of columns that are completely uncorrelated not stop called an orthogonal basis principle component so it's the most uninteresting correlation matrix 'cause it's just ones on the diagonal and zeros everywhere else OK so that's the first thing now the second thing is if you do it correctly and you scale your data and so on you can also print out something called an eigenvalue this is also a very important metric the idle volume is associated so each of these PC's is going to be a column so column number one will car is PC1 and that will have associated a very special number called and eigenvalues value tells you it's kind of like how much or how good is this variable or this principle component it tells you how many of the original features can it replace so if you do a principle component and you print out the eigenvalues and the first eigenvalue is 11 let's say that that means that the principle component one has the same explanatory power as eleven of the original features so it can actually replace 11 of the original features it has the same power as eleven of the original features so the eigenvalues will be in descending order because PC 2 will have an eigenvalue that smaller 'cause it explains less and PC3 smaller pieces were below so if I use the eigenvalue approach what should I cut off beat if an eigenvalue means same explanatory power as an original feature yeah 70% of the number what is greater than one yes because if it's greater than one it has more explanatory power that one that's smaller than so all the eigenvalues will be a number between infinite at 0 yeah some of them will be less than one all the lower lower level pieces stop music those pieces explain less than original feature therefore you could also choose the number of principle components by retaining all of those that have eigenvalues greater than one yes which also means that the sum of the eigenvalues is equal to N number of features Yep exactly this is much much cooler than I make it it's very very brilliant it's very very you know I'm not a religious person but sometimes you think somethings going on here that you just can't explain it 'cause it's really weird that you can do all this stuff so yeah look for eigenvalues greater than one or you can use the method with the variance 70% so there are different ways and has turned out all these pieces are obtained using something called projection so this is how it's done so you censor the data that means you subtract the mean from all the features will call this the data matrix X so that's the first step you take all of your data and you should practice each observation from the mean of that cult step number one step #2 covariance matrix which is just X transpose times X it's called the covariance matrix it tells you how each column varies with every other colour step #3 you find what is called the eigenvectors and eigenvalues this is called the eigendecomposition this then you sort the eigenvalues from largest to smallest because the ones with the highest will be the most important and the one with the smallest will be the least important so PC one will be the one with the highest eigenvalue and then five calculate the eigenvectors composition and of course much of this would make a bit more sense if hot there and finally you take your original data matrix the 1 which made-up here and you multiply all these this the vector or be sorry this be matrix for the X is and that will give you all the principle codes So what you now do is this year will be a matrix that is as big as X he will be as big as X but you only choose the first 2-3 or four or five or six columns whichever you want because it will be ordered such that the most important information is in PC one second most computed 2 and so on that's how you do file you rearrange the data and then you select the most important elements to represent Justin Bieber whatever sports yes desktop I hope you do 'cause it's it's it's pretty smart this makes sense especially so we don't do this if we stopped here send need you to know that the components are uncorrelated you don't need to use the word nice to use the word orthogonal dated to a concept which is probably have had since high school the dot product you probably never thought about this adult product is important what most of the stuff would you do in this course deals with the duck promise between two columns if the duck product is 02 columns are uncorrelated the dot product is a measure of how similar two things are OK so you could you could make me into a matrix and you could make yeah you insert matrix dot +2 and will get a measure of how similar are the most similar two things can be is orthogonal so that's this here so that's represented here two things that go in opposite directions are not dissimilar they are still on the same vector they just go in different directions so the most Dick the biggest difference between two things is this year you orthogonal so similar are two things and it turns out that the eigenvalue is the same as the variance open feature so yeah I won't ask you about this but yeah it makes sense did you go through testing yeah conceptually but it's very very complex mathematically very calm I just say a little bit about conceptually so PCA I've done it on the M list data with the classification of zeros and six is this is how PCA does is all testing testing is brilliant but it can only be used for visualization that you can't you can't get out the PC's like you can in the PC and use them so there are no values here this is only visualization just so you know so it's only used for visualization the difference between them is PCA tries to find global structure testing tries to find local structure so it's what what it stands for is see distributed stochastic neighborhood embedding so T distributed means it uses a conservative form of the normal distribution stochastic means it deals with probability any we can talk about that simple neighborhood in bed So what does it mean well think about it see all of you coming from all your different countries yeah Indian all these different places so you know sometimes when people come from different places they come to a new country like Denmark yes now in some cases you can spread out live in different parts of Denmark and you kind of retain was called the global structure of them and then it's very diverse insert or sometimes with people move from India geria that kind of go to a new place and they live in the very you know in the same distance to where they used to live to understand the analogy here So what this means is the neighborhood where you came from you kind of want to recreate that in the gnu space where you go to so the reason why this is called neighborhood embedding is you go from a higher dimensional space like these end features to two dimensions yeah and when you move from I100 dimensions to Denmark you want to retain the same neighborhood we want to live in the same space proximity to whoever you were living with in the in the 100 dimensional space OK so it's neighborhood preserving so so PCA would create this one and it will create this one correct it will never see that there's actually a neighborhood here it will never see that there's actually a neighborhood here PCA won't be able to do that 'cause it doesn't look at local structure it only looks at global structure so it was only be able to identify this and this similar to what I did here Easter so that's what it can do where testing is able to look at local structure so testing preserves local structure testing would be able to look at this and say they actually belong together well remove these into two-dimensional space they need to be grouped together you can't spread them out into new dimensions because these ones here APC 8 my might move into PC 7 and one moves into PC 10 no testing says they have to be in the same cluster same grouping so PCA tries to preserve global shape of the data whereas testing preserves local stream so if you're close to something in higher dimensional space you should also be close to it in lower dimensional space plus the difference between Yep it's able to identify them yes yes Netflix number so unsupervised learning have you gone through these 3K means DB scan and hierarchal clustering I'm gonna ask you about K means so what's K how the number of clusters so in that way it's not completely unsupervised visit hey is the number of clusters in Europe end up with so you might tell it I want 5 clusters and give you 5 clusters no matter if it's right or wrong so it's not really completely unsupervised so basically it partitions the data into K clusters based on centroids so it kind of looks at the centroid so how it works is it initializes randomly or the old algorithm initializes randomly and then it kind of moves the clusters centres in one iteration so to make it better and better and better this very good and when does it stop when is it and when does it not continue anymore exactly when it becomes stable and no observation is moved into a new cluster exactly so that's K means clustering DB scan is density based so you have this kind of looks at where how dense is the maps it into a N dimensional space looks like where where how this are the different places and it makes clusters out of those that's used a lot in like in the universe because there's a lot of empty space in the universe yeah so you would want to detect where things are cluster density hierarchal clustering has 2 forms I don't know which one you learn there's one called agglomerative and one called divisive first one kind of the same so I'll come up tip everyone is its own cluster we keep adding them together until we start at or we end up at 1 supercluster yeah and the truth lies somewhere in between Siri and we have these dendrograms that can tell us where it is very good but divisive 1 is the opposite everyone starts at one cluster and instead of looking at what unites us we look at separates spit up the clusters so it's just two ways of doing the same thing more like family we should look at similarities between us whereas the divisive 1 is I'm different let's go into different clusters but they do the same thing it's just whether you have a bottom up or top down yeah I promise you the distance metrics these are the ones when she I usually talk about there's the euclidian one which you should probably know by now the Manhattan one yeah where you always walk perpendicular to the axis always four what's a better word for perpendicular or orthogonal yes mikowski here you can see that minkowski is exactly the same as these two because this P parameter that I was talking about is just so if it's euclidean this is 2 the square root there's technically 2 here technically we just don't right there's technically script so if P is equal to two you have euclidean if P is equal to 1 what's X - y to the power of 1 that's just this yeah and then you can have 3 and so on so yeah then you have some other ones I'm guessing you didn't they're actually pretty cool especially the humming this is what what you use if you for instance I was talking about similarity between us we could also look at the DNA sequence my DNA in your DNA and in order to assess how similar two things are he used to edit distance that's if you ever feel like true crime and stuff like that all these DNA analysis they use this edit instance to figure out how similar one DNA string is to another that's where they say this person this DNA belongs to this person he like one out of a million or something like what they basically mean is that measured the distance between this DNA string and this DNA string and the distance between them is very very small and the distance metric would be used is called edit distance it basically what you basically do is you have 3 operations so you have two strings OK so let's say you have a string called hello and you have a string called goodbye OK so you want to know how many operations does it take me to translate hello into goodbye if I could only do insert delete and replace and the less operations I use the more similar to things are yeah the more operations I need to do more dissimilar that's called edit OK what I do want to talk about is activation functions so these aren't just chosen arbitrarily I've tried to make a table here which tells you when to use different this one is marked out because I wasn't sure whether you've talked about them in this class but you will definitely see leaky relu and animal in mount to course try to look at them when you're doing binary classification you should always have the sigmoid in your output layer OK always when you're doing multi class classification so that means that when the when the when the label for instance if you're trying to predict numbers 012345678910 nine sorry not said you should use softmax softmax has what we call soft voting we talked about this yeah so it calculates 10 probabilities if it's sent classes and it just selects the one with the largest probability and predicts it must be this yeah so the output of the softmax is a vector with the number of probabilities as you have classes and it just selects our Max of this so it takes the largest probability and predicts that class so in that way it's very similar to the sigmoid 'cause that does the same except here we only have two probabilities P and 1 minus P2P3 before but all of Papa till the number of classes but each of them will some or all of them will sum to one OK so remember that now I've written here that I I believe the best results come when you use relu for all your hidden layers so use radio in your hidden layers never in your output so you should either use sigmoid softmax or linear in your output layer linear if you're using regression or if it's some aggression problem softmax if it's multi class classification and sigmoid if it's binary classification and everything in between relative now multi label classification I don't know if any of you are doing that but that's what you want to classify multiple things at the same time so for instance if you have an image of a cat and a dog that's multi label classification top 2 outputs and that will enable it to give you a probability or each one of them will this be associated with probability the probability that this is a cat and the probability to not account from this is a dog probability is not so here you can use it on multi label never really seen any projects with multiple and here you actually have the sigmoid function looks like this this is the S Max you don't need to know that so you need to know for instance if you do let me give you an example here something like this ship you have some sequential model here here you do radio relos off box so I could ask you why are you using softmax here you should be able to tell me that's be cause that the problem I have is a multi class classification problem OK in that way so that's how I'll go in here and begin to ask you what do these things mean that kind of written here with all these different things I will when you go through the code I think you should you should know what all the different things are because I will ask I will ask you to explain yes what does this mean why do you have that must be a reason even if it did give you it asked please explain fax size 2 yeah or please explain what this means please so do that if you choose to use it get it to explain all the elements like I did here I didn't make this so I got it to do all of this and then I did this from the book but here's another one with the with the sorry the CNN same here So what is the Max pooling and all these different things so please no what your code does that's very important and usually the code is pretty simple I mean in terms of really be more complex OK cute So what will happen I'll be there probably together with or Frederick one of those two I'll be your examiner you come in and there are going to be I don't know 24 cards on the table you choose one of them it's going to have a number between 1:00 and 6:00 that corresponds to your your assignments and whatever 1 you use we begin discussing that there is no presentation the first thing that happens after you've drawn a car is I ask you a question or I say something please show me something please explain the problem so you don't start talking I asked the question is the first thing that OK I will ask you unless the exam goes very badly I will not begin asking you about the curriculum in general I will try to stick to your what you've written and the theory that's inside of your sign so if you let's say you chose our regression test this one we won't be talking about all different types of classification will try and stick to regression and the same with your final project will try to speak to but of course if you know nothing about your project you don't know anything about the assignments then I'll begin to see can I can I find something somewhere if I stop the exam earlier than you expected it's probably because I've already decided that it's the top grade and I don't need to use more time on giving you the possibility to get up first grade I understand kind of like pulling you out before you can so that that can also help or if I begin to the very distant interested that's also because you're doing well and I don't need to follow it that so around 9:00-ish minutes will move over and begin talking about your final then you're gonna go out we're gonna talk in terms of please have all of your code and I think it's also written everything should be execute before you upload I'll be nice 'cause think about it how many projects are how many students drop to execute project so please have outputs so I don't need to execute it or run it again and also before you enter just open all six hundreds in a BS code or whatever they're using and have them ready so you don't have to spend time yeah about the all the distance I'm not gonna ask you specific details about the metrics other than for instance here I'll ask you about the euclidean Manhattan and then I might ask you if everything is going well do you know about the minkowski distance and how is that related to euclidean and Manhattan and if you can then say well it's just a generalization or the Manhattan is just a special case of very good you don't need to account for the formats so yeah thank you to be able to explain especially the 1st 2:00 and also how do they work so so you should know that euclidean distance is the smallest distance in in space is between two points straight line is also sometimes called whereas this one here you always travel orthogonal to the axis Yep we need to have them ready probably before the exam Oh yeah before the exam I'll be honest with you I will most likely not look at them before maybe between like 26th of December depends on if I'm going on a holiday or not we're talking about so the earliest will be around the 27th or else it will be yeah couple of days before is it open so you can change them if you want so no I need the actual IPY NB51 for each hundred in the vice flow that the that the secretary has made the wise will you upload the summaries of each of them it will be 1 file each summary is one page which means page one is assignment one final project has 2 pages of summary that will be pages 7 and 8 so a total of 8 pages in one PDF is uploaded to wise will not the notebooks that's your ticket for the exam that's the one I'll be looking into as soon as I've seen you uploaded it to wise flow I'll never look at those summaries again after 20'cause I need to see the same as you the first one I don't care about those summers OK Oh yeah if you if you're able to do it if you can change it you can change yeah I'm not sure you can change them if you've uploaded yeah but even if I open it I don't think you could resubmit I don't think so yeah but I mean if you had like a crappy first design which it probably happened it doesn't really matter I mean hopefully you will be able to identify first assignment was not very good because what have you will have machine learning two weeks or something else I don't expect I mean if I had given you that assignment today I would have expected it to be better so it doesn't matter Yep play music