Introduction to Proteogenomics

Dr. Sanjeeva Srivastava
Dr. D. R. Mani
Department of Biosciences and Bioengineering
Principal Computational Scientist
Indian Institute of Technology, Bombay
Broad Institute of MIT and Harvard, USA

Lecture – 21
Machine Learning and Clustering

Welcome to MOOC course on Introduction to Proteogenomics. In last couple of lectures with Dr. Mani, you have learnt about different ways of data normalization as well as different type of statistical test employed to look at the data and how to obtain the meaningful information by comparing your controls and treatment and when you can employee the right type of test. Therefore, you have lot of data from the omics experiments, but you cannot obtain the meaningful information until unless you are able to look for the right type of test and know that exactly how you are comparing your data. In this light the machine learning tools can be very helpful.

Today Dr. Mani is going to talk to you about Machine Learning and its application; he will also discuss about Clustering. So, in context of machine learning algorithms, it will be very interesting to note how you can use many of the test, many of the essays which we want to do in much more high through put manner using machine learning. We will also learn different applications of supervised as well as unsupervised learning. Dr. Mani will then talk about different type of clustering such as hierarchical clustering, K-means clustering, fuzzy clustering and consensus clustering and so on.

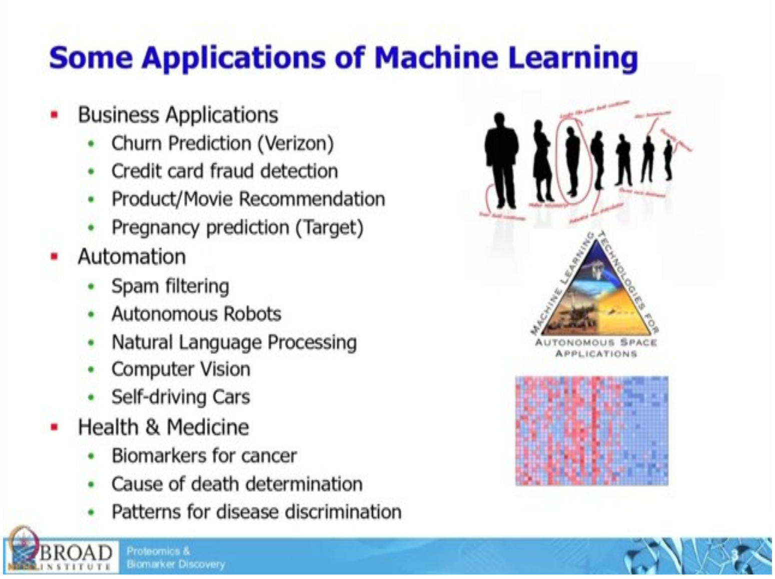
He will also provide you a brief idea about the applications, various type of cluster visualization and principle component analysis. Further he will talk about the differences between classification and regression and then, he will talk about random forest and classification tree and how both of them differ in terms of the over fitting of data. Finally, Dr. Mani will talk about how machine learning can help in omics data analysis and right type of biomarker selection. After this brief introduction about today's lecture and different type of test which Dr. Mani is going to explain, let us welcome Dr. Mani for his today's lecture.

So, the next topic, that I am going to quickly cover Machine Learning. So, machine learning is a sort of an out tree outshoot of research in artificial intelligence. So, in the 60s people started trying to build computers that behave like humans. So, and the area was called artificial intelligence and one of the signs of intelligence is that you can learn. And so, getting computers to learn was part of the endeavor to make them intelligent and so, machine learning is like a area of artificial intelligence.

The kind of disconcerting thing though is that you think something is if a machine did that, that would be considered intelligent and then the machine does it and then you say, oh! no that is quite it, it should, it did not behave intelligently. So, I think playing chess was one of those, you people would say if a computer could beat a human in chess than the computer would be considered intelligent. The computer beat the world chess champion in chess and then people said, well it was just an algorithm that was only for chess; it is not really intelligent.

So, I think the it the goal keep shifting, but at least the research has provided a lot of tools that are currently in a use or becoming increasingly useful.

(Refer Slide Time: 03:59)



Some Applications of Machine Learning

- Business Applications
  - Churn Prediction (Verizon)
  - Credit card fraud detection
  - Product/Movie Recommendation
  - Pregnancy prediction (Target)
- Automation
  - Spam filtering
  - Autonomous Robots
  - Natural Language Processing
  - Computer Vision
  - Self-driving Cars
- Health & Medicine
  - Biomarkers for cancer
  - Cause of death determination
  - Patterns for disease discrimination

MACHINE LEARNING TECHNOLOGIES FOR
AUTONOMOUS SPACE APPLICATIONS
BROAD INSTITUTE
Proteomics & Biomarker Discovery

So, what are the applications and I have some quick applications. So, this churn prediction was something that I worked on in my previous life when I was working for the phone company. So, what you do is, you look at customers calling patterns, their when they pay their bill; lot of details about the customers. This is like cell phone service and then you predict who is going to leave the company in the next 2 or 3 months. And, then you go and make them special offers or give them a free phone or whatever to get them to stay. So, that is what I was doing before I started doing cancer research.

The other thing that is commonly encountered by people is credit card fraud. So, if I come; so, last time I came to India, my wife wanted to buy some sarees and I did not have enough cash. So, I took out my US credit card and handed it the store clerk and the card was denied. Why? Because I leave in the US, they know that and there is a charge coming from India because they know the source of the charge. So, that is very highly unlikely.

So, it will be denied and even in the US if you make a small charge on a petrol pump like you charge like 1 dollar on the petrol pump and then you go to an electronic store and try to buy a TV, it will be denied because that has been the machine learning algorithms have decoded at. That is the pattern that people use when they have a stolen credit card. Because in a gas station in the US it is unattended, there is no person filling your gas. You go to a machine you put your card and then you try to fill it. If it does not work then, you just drive away. If it works, then you really do not need gas. So, you take the credit card and go and try and buy an expensive thing.

So, they have recognized that pattern and using machine learning and they will stop the transaction if that happens. So, these are kind of like interesting applications of machine learning where if you have used Netflix, you will get recommendation for movies. So, Netflix had a competition called the Netflix price. If you could predict who would like a movie better than what they were doing with some percentage, you would get a million dollars as a price. So, there was a group that got a million dollar prize and the paper is published to it is called the Netflix million dollar competition. You can take a look at the paper if you are interested.

This one yeah, I will mention it. I think these are cool applications of machine learning. So, target is a super market in the US. They sell all kinds of stuff. So, they sell soaps, shampoos, cribs, diapers for babies, there you can buy electronics, you can buy a bike. It is like a super store where you can buy pretty much anything. So, they started a data mining program because they realized that when a woman becomes pregnant, they start shopping in a specific place and then when they have their baby they buy a lot of stuff in the same place. They usually do not go to a different place to buy things after they have their baby.

So, they figured; if we can get the women to come here before they become, before they deliver. So, as soon as they become pregnant if we give them a offer so, that they can come to our store and start shopping here. Then once they have the baby, they can they will keep coming here and they are very profitable. So, they hired a few statisticians and said this is all the data figure out who are pregnant women. So, we can send them a discount mailer to get them come into the store. So, that is kind of the power of machine learning and I guess the misuse of machine learning, you can look at it either way.

So, yeah if you look at the model that was done for the pregnancy prediction and see what are the factors that predict when a women is pregnant, it was simple things like they buy shampoo that is not scented, they buy a specific type of cream for their body. So, there were like very few thing like that, that predicted when a women is pregnant. So, all you had to do was capture all their transactions. So, that you knew what they were buying. And, when they switch from a scented shampoo to an unscented shampoo and they start buying this kind of different cream, then you it is very highly likely they are pregnant.

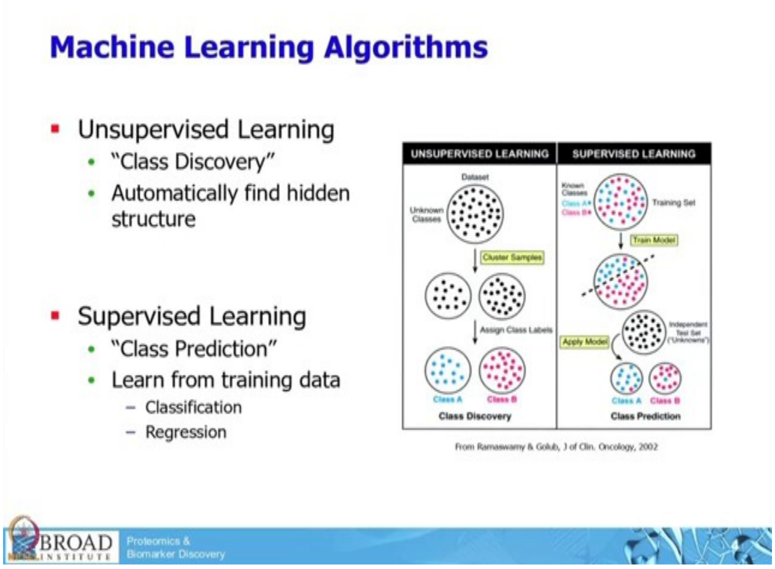
So, there are lot of other applications and off late you might have so, most of your phones have machine learning based voice recognition. So, previous like 5 years ago when I gave a command to my phone, it would not recognize anything because of my accent. Now, it is fine that is because the natural language processing using these newer techniques called deep learning has made the voice, the speech recognition so robust that its now does not care about accent, it can read through accents.

There is also self driving cars; so, in the US there is a car company called Tesla, that has a car that basically it is a auto pilot mode. You put it in that mode; it will if there is a lane marking on the road, it will follow the lane and make sure it keeps enough distance from the car in front of it and just keep going, you do not need to do anything.

So, there was a recent two days ago there was an article in the newspaper where the police saw that there was this car Tesla going on the highway and they noticed that the driver was like half asleep on the steering wheel. And so, they looked a little closer and it looked like he was drunk and he was asleep at the wheel and the car was still going. So, they knew it was a Tesla on autopilot.

So, what they did was they overtook the car, went in front and slowly started slowing down and this car slows down, because it is maintaining a distance from the car in the front. And the cars, police car stopped this car stopped, they went and woke up the guy and gave him a ticket. So, those are all like cool applications that we can talk about, but the main thing we want to do is apply machine learning to what we are doing which is cancer research.

(Refer Slide Time: 10:01)



Machine Learning Algorithms

- Unsupervised Learning
  - "Class Discovery"
  - Automatically find hidden structure
- Supervised Learning
  - "Class Prediction"
  - Learn from training data
    - Classification
    - Regression

From Ramaswamy & Golub, J of Clin. Oncology, 2002

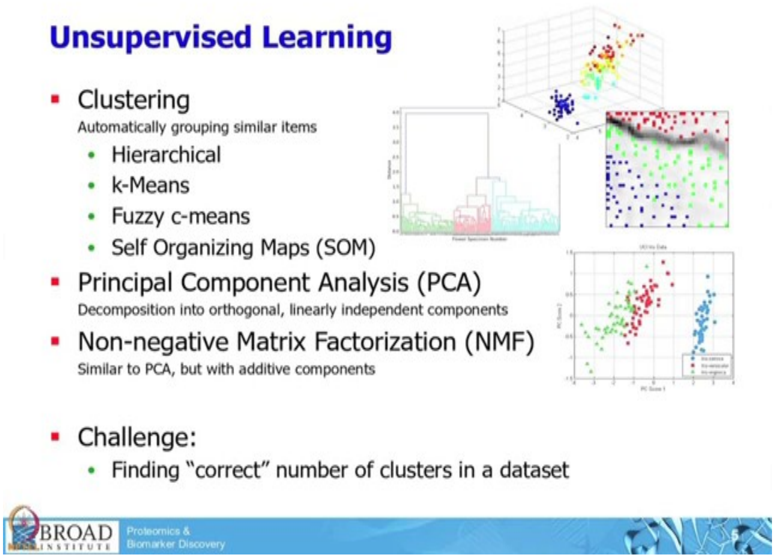
So, machine learning comes in; so, there are lots of variants, but I will kind of simplify this a little bit. So, there is this thing called unsupervised learning where you try to use machine learning to discover things in the data. So, it is called class discovery where you want to automatically find hidden structure.

So, you have breast cancer samples. What are the innate groups in breast cancer? So, you can you want to try to find the classes or the structure in breast cancer or any data using machine learning. So, that part is called unsupervised learning. The other part is called supervised learning where this is class prediction. So, you have seen examples of some type. So, you have examples of people having cancer and then you know their genome profile. You have examples of people who are normal who do not have cancer, you have seen their genome profile.

So, looking at the profiles and the fact that you know they have cancer or not. Can you predict when a new genome profile is given whether that is from a cancer patient or a normal patient? So, that is called class prediction. So, there you are trained using non-classes and then you want to make a prediction on the class when you get new data.

So, this is you learn using what is called training data. Here you basically just look at data and try to find similarities in the data automatically. So, this is like a cartoon from a paper that kind of describes it, I would not go into that, but so, some examples of unsupervised learnings.

(Refer Slide Time: 11:28)



Unsupervised Learning

- Clustering
  Automatically grouping similar items
  - Hierarchical
  - k-Means
  - Fuzzy c-means
  - Self Organizing Maps (SOM)
- Principal Component Analysis (PCA)
  Decomposition into orthogonal, linearly independent components
- Non-negative Matrix Factorization (NMF)
  Similar to PCA, but with additive components
- Challenge:
  - Finding "correct" number of clusters in a dataset

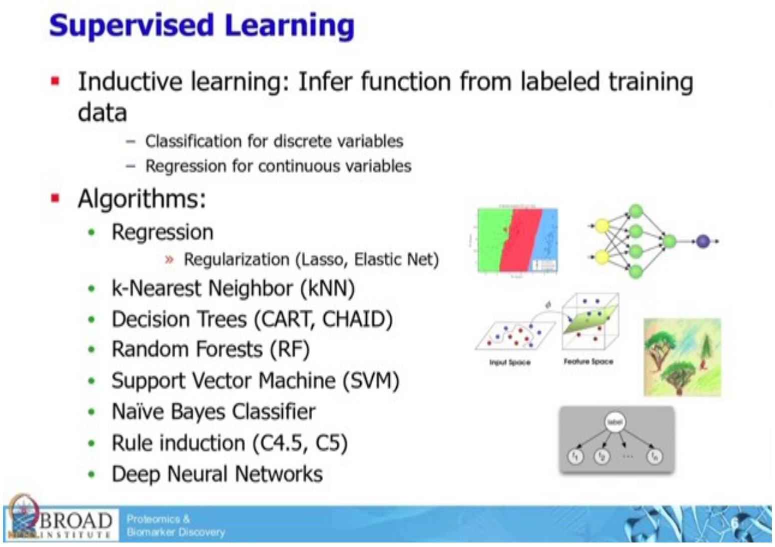
BROAD INSTITUTE Proteomics & Biomarker Discovery

So, clustering is a really classic example of unsupervised learning. So, you want to run clustering, you get some clusters and you hope that those clusters have some biological meaning. So, that is clustering.

The other thing many of might be familiar with is called Principle Component Analysis or PCA. So, if you do PCA; if there are natural groups in your data, it will show you. So, one thing you can do is you can do PCA and then color your samples by the batch. If the one batch and the other batch completely separate in PCA, you have a batch effect that you need to address. But if they are all mingled together, you do not have a batch effect; then you can color by some other information you want to see whether it is naturally available in the data or not. So, like subtype or cancer subtype or cancer versus normal things like those that have very strong markers will result in separation those samples in PCA.

So, PCA is even though it is unsupervised many times you mark the labels using other knowledge to kind of visualize the data. So, PCA is unsupervised method where you primarily use it to visualize data and I can ignore the rest. So, supervised learning is where you look at training examples and then you try to learn. So, you are shown an image and you say this is an image of a cat; you are shown another image you say it is an image of a dog. So, you see millions of these images with labels of cat and dog and then if you get another new image, can you say whether it is a cat or a dog? So, that is like a supervised learning algorithm.

(Refer Slide Time: 13:05)



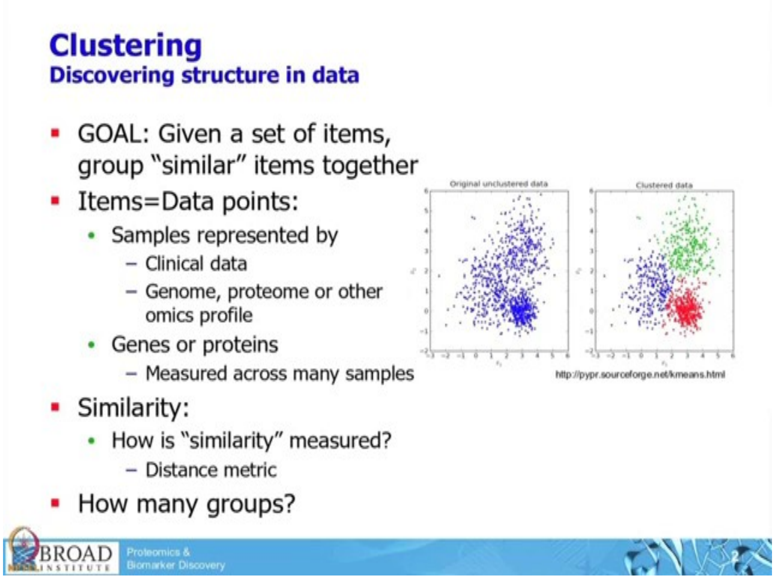
Supervised Learning

- Inductive learning: Infer function from labeled training data
  - Classification for discrete variables
  - Regression for continuous variables
- Algorithms:
  - Regression
    - » Regularization (Lasso, Elastic Net)
  - k-Nearest Neighbor (kNN)
  - Decision Trees (CART, CHAID)
  - Random Forests (RF)
  - Support Vector Machine (SVM)
  - Naïve Bayes Classifier
  - Rule induction (C4.5, C5)
  - Deep Neural Networks

So, here there are lot of algorithms you can use and there is another distinction between regression and classifications. So, if you have to predict a continuous value so, how long are you going to live after you are diagnosed with cancer? So, that is a prediction, survival prediction is a continuous prediction. So, that would be regression. But if it is grouping so, which group do you fall into? Do you fall into like the basal cancer or luminal breast cancer? So, that would be classification.

So, unsupervised learning I mentioned is primarily clustering PCA techniques like that. In clustering the goal is to you are given a set of items; so, your samples or your images or speech from a set of people and you want to group similar things together. So, you somehow have to mathematically define what similar is and then things that are similar, you group together to form a cluster.

(Refer Slide Time: 14:06)



Clustering
Discovering structure in data

- GOAL: Given a set of items, group "similar" items together
- Items=Data points:
  - Samples represented by
    - Clinical data
    - Genome, proteome or other omics profile
  - Genes or proteins
    - Measured across many samples
- Similarity:
  - How is "similarity" measured?
    - Distance metric
- How many groups?

http://pypr.sourceforge.net/kmeans.html

BROAD INSTITUTE Proteomics & Biomarker Discovery

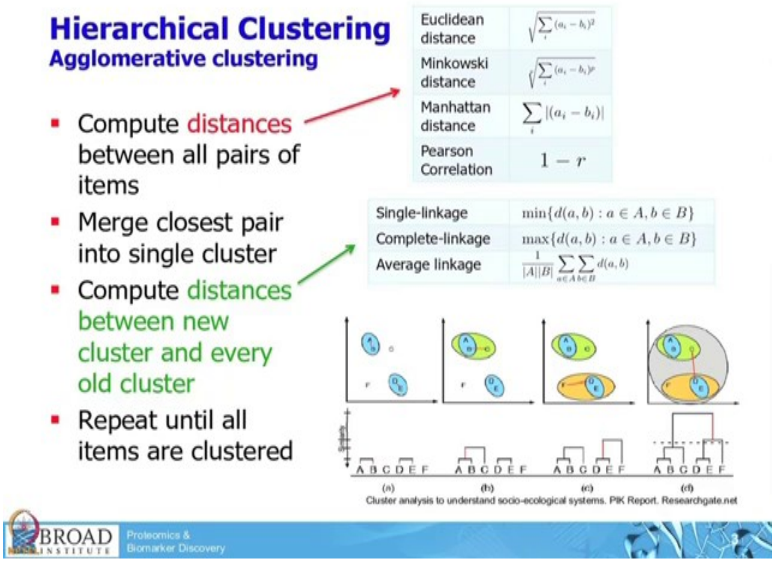
And so, items are basically data points. So, there have listing of what they could be, they could be clinical data when you are looking at samples, they could be genes or proteins. So, which genes are similarly changing across all your patients that you can cluster genes by similarity; you can cluster proteins by similarity.

So, any data point can be an item and similarity is usually measured using a distance metric. So, I think I have few examples, but basically what you do is you can do your Pearson correlation or any kind of correlation is the similarity matric. There are other like Euclidean distances, similarity metric in geometric space for example. If in 2D or 3D, it is easy to visualize. You have some points scattered around which are your closest points and there you are using Euclidean distance.

So, there are many different metrics for measuring similarity and when you want to do clustering which matric you pick would depend on the kind of data you have and what you want to accomplish with the clustering. And, the big question in clustering that is that for which there is still no definitive answer is how many groups are there. So, many times when you run clustering, you have to say I want 5 clusters and then it will come up with group of 5 things that are all similar. But you said 5, how many are there naturally in my data?

So, I have 100 breast cancer samples, are there only 5 50 subtypes or are there 7 or is there only 2? How do I find the natural number of groups in my data? So, that is a much more complex problem; there are some solutions to it, but nothing is ideal you have to try a lot of it before you can figure out what is happening.

(Refer Slide Time: 15:54)



Hierarchical Clustering

Agglomerative clustering

- Compute distances between all pairs of items
- Merge closest pair into single cluster
- Compute distances between new cluster and every old cluster
- Repeat until all items are clustered

Euclidean distance	$\sqrt{\sum_i (a_i - b_i)^2}$
Minkowski distance	$\sqrt[p]{\sum_i (a_i - b_i)^p}$
Manhattan distance	$\sum_i \lvert(a_i - b_i)\rvert$
Pearson Correlation	$1 - r$

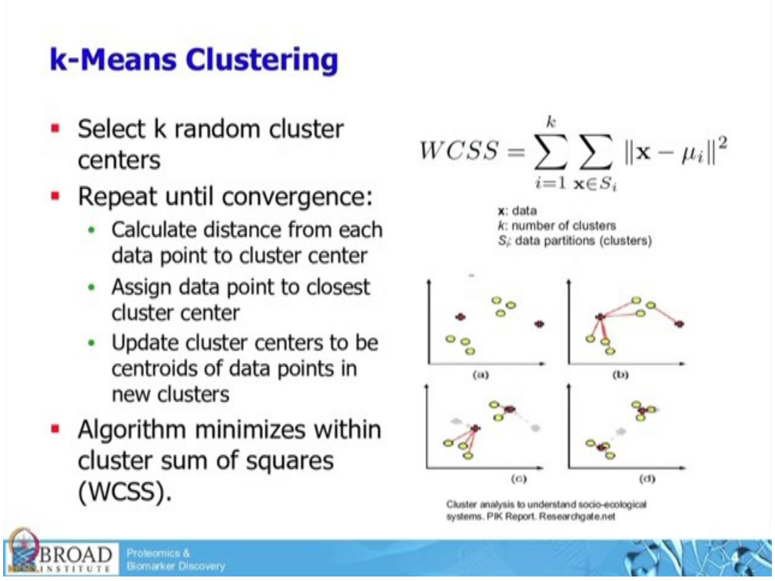
Single-linkage	$\min\{d(a,b) : a \in A, b \in B\}$
Complete-linkage	$\max\{d(a,b) : a \in A, b \in B\}$
Average linkage	$\frac{1}{\lvert A\rvert \lvert B\rvert} \sum_{a \in A} \sum_{b \in B} d(a,b)$

Cluster analysis to understand socio-ecological systems. PIK Report. Researchgate.net

So, this is another thing for which you need to hire somebody like me. So, it is all job security I think that is why people do not a. So, hierarchical clustering is one. So, when you are measuring things that are similar, you can do it in two ways. You can say I have all my samples for each sample, I am going to look at everything and then find the things that are closest and then keep merging things that are close together. So, that is called bottom up or agglomerative clustering. The other one is you look at all your samples and say I want to divide based on similarities. So, you start with everything and you start dividing your set of samples. So, that is top down clustering.

So, hierarchical clustering is agglomerative. So, it starts with all your samples, it measures distances between all pairs of samples and then starts grouping things that are similar together and you usually get like a dendrogram that shows which samples are similar and how far apart they are. Many times hierarchical clustering is used to kind of group samples and genes when you are visualizing data using a heat map.

(Refer Slide Time: 17:00)



k-Means Clustering

- Select k random cluster centers
- Repeat until convergence:
  - Calculate distance from each data point to cluster center
  - Assign data point to closest cluster center
  - Update cluster centers to be centroids of data points in new clusters
- Algorithm minimizes within cluster sum of squares (WCSS).

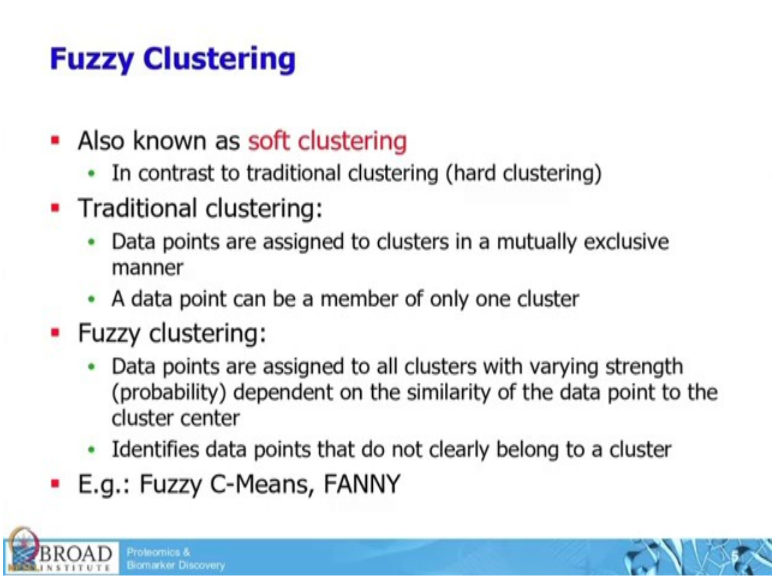
$$WCSS = \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \mu_i\|^2$$

$\mathbf{x}$: data
k: number of clusters
S_i: data partitions (clusters)

Cluster analysis to understand socio-ecological systems. PIK Report. Researchgate.net

k-Means clustering basically starts with everything and tries to find for a given a group, it tries to find things that are close by and then it keeps iterating the process till the whole clustering has stabilized and then you find some number of clusters. Usually in most clustering algorithms a point either belongs to a cluster or not so, the membership in a cluster is mutually exclusive. If you are member of cluster 1, you cannot be a member of cluster 2 also because you take that point and put it in the cluster. Many times you may not want to do that especially if you have many similar things and you are not sure where it falls in a point falls into.

(Refer Slide Time: 17:44)



Fuzzy Clustering

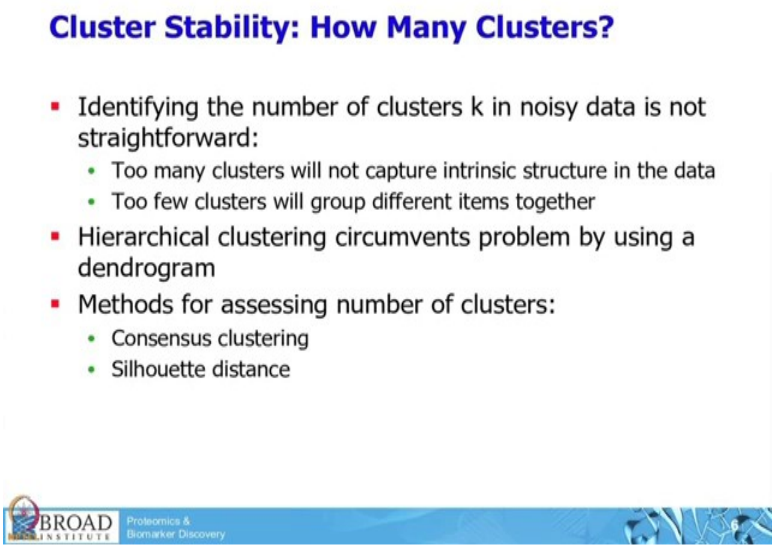
- Also known as soft clustering
  - In contrast to traditional clustering (hard clustering)
- Traditional clustering:
  - Data points are assigned to clusters in a mutually exclusive manner
  - A data point can be a member of only one cluster
- Fuzzy clustering:
  - Data points are assigned to all clusters with varying strength (probability) dependent on the similarity of the data point to the cluster center
  - Identifies data points that do not clearly belong to a cluster
- E.g.: Fuzzy C-Means, FANNY

You might want to do fuzzy clustering where you say my point belongs to cluster 1 with probability 0.7 and it belongs to cluster 2 with probability 0.1 and to cluster 3 with probability 0.2. So, it will all add up to a probability of 1, but it will give you the proportion that the algorithm thinks that a sample falls into a each of the clusters. So, fuzzy clustering; so, it might so happen that you have 3 groups in your samples, but then there are some samples that are like do not belong at all.

So, when you do clustering the by the nature of the algorithm, it will force fit those samples into one of the groups which may not be the right thing to do. So, in those cases what you would do with fuzzy clustering is you would get low probabilities for all 3 clusters or approximately similar probability for all 3 clusters and then, you would say I do not know where to place these points. So, I am going remove those. So, if you remove those then you will have very clean clusters and then you can go on try to biologically determine what each of the clusters are without the other samples adding to the noise of the biological analysis. So, in some situations, it may be useful to do fuzzy clustering.

So, in determining how many clusters one of the things people do is to check how stable a clustering is. So, let us say you come up with some clustering. You perturb the data in some way, you remove a couple of samples or you repeat a couple of samples and then you redo the clustering, how different is the clustering. So, you redo this 1000 times and then you measure how many times do a pair of points always fall into the same cluster so, that is called cluster stability.

(Refer Slide Time: 19:25)

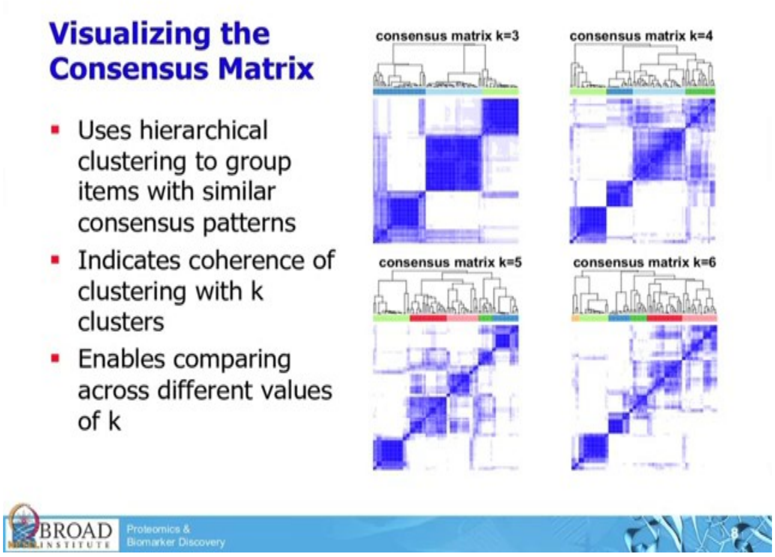


Cluster Stability: How Many Clusters?

- Identifying the number of clusters k in noisy data is not straightforward:
  - Too many clusters will not capture intrinsic structure in the data
  - Too few clusters will group different items together
- Hierarchical clustering circumvents problem by using a dendrogram
- Methods for assessing number of clusters:
  - Consensus clustering
  - Silhouette distance

And there is a method called consensus clustering that kind of uses this to come up with the correct number of clusters to look at. So, what it does is, you start with number of clusters predefined, you start with 2 clusters, 3 clusters, 4 clusters and so on. For each number of clusters that you have specified, you go and redo the clustering a 1000 times and you create what is called a consensus matrix.

(Refer Slide Time: 19:56)



Visualizing the Consensus Matrix

- Uses hierarchical clustering to group items with similar consensus patterns
- Indicates coherence of clustering with k clusters
- Enables comparing across different values of k

So, this is say these are samples these are also the same samples in the same order. And what this is saying is if I take this first sample here, how many times does it fall in the same cluster as this other sample? So, if their number the proportion is high, then you have a darker blue dot; if the proportion is low, then you have a lighter blue dot and if they never occur together, you have white.

So, for a ideal clustering, you want a solid set of blocks that span the diagonal. So, that is saying that the clustering is very stable; things that are together are always together. So, if you look at this is from the breast cancer paper, this is with 3 clusters 4, 5 and 6. So, you can see the cleanest consensus matrixes is for when you have 3 clusters. So, that is why we decided the proteomics data shows 3 clusters in the breast cancer data. So, with consensus clustering there are more matrix, you can calculate to kind of formally determine how many clusters you have.

So, you do clustering from 2 to 10 clusters and then there are other matrix you can calculate to figure out what is the optimal number of clusters other than just like visually looking at the consensus matrix. So, recently Karsten and I were trying to figure out an automated algorithm to find out the right number of clusters and Karsten found a paper where they have 32 matrix, they calculate and they show you all those matrix and in 100 percent of the time, half of those matrix do not agree. So, I half the matrix will survive and all. Some percentage will say 2 clusters, some will say 4 clusters and there is no way to decide.

So, even if you a lot of matrix, it is not clear what the correct number of clusters are. And so, many times you would have to do it visually or you do it visually and then you see if it makes biological sense by doing pathway analysis of which pathways are enriched in cluster 1, which are enriched in cluster 2 and 3. And, you map this to information, you know about cluster of a cancer sub types and say ok; this seems to make more sense.

So, in other words assigning number of clusters is more of an art than a science even though there are some statistical tools available. So, for clustering also, you would need like a tons of samples to do any clustering; otherwise you would basically be looking at a very a clusters that are not very robust. So, you do clustering only when you are doing like a relatively larger study. If you have like 10 samples where you are looking at some IP or something like that you would not do any clustering. But, in a discovery study usually where you have 10s to 100 of samples, you would definitely look at something like this.

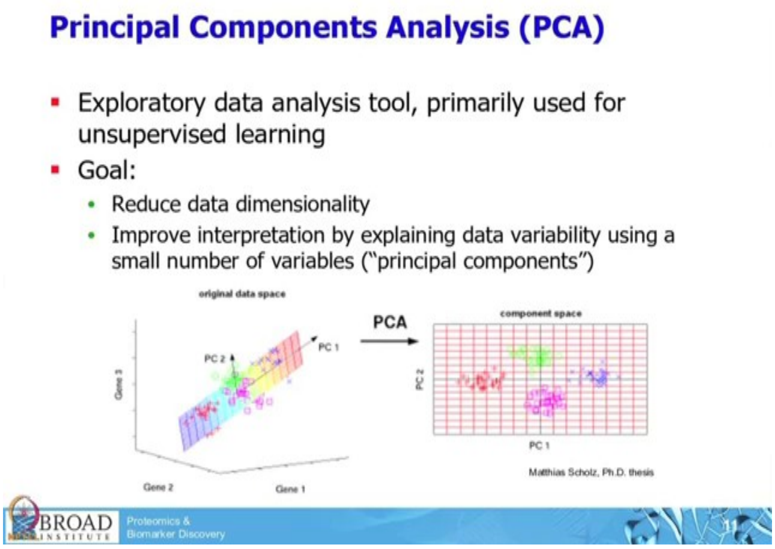
So, if you look at proteogenomics papers, pretty much all of them have like close to 100 or more samples and all of them do clustering. Even if you look at like a genomics and a kind of RNA papers that came out 15-20 years ago, I think they were the first paper that applied a hierarchical clustering to affymetrix microarray data I think in 97 or 98, I think have like 30 or 35 samples. So, you need at least that many to have a reasonable clustering.

When you get through a clustering, you do not know what they are you just the algorithm see the algorithm does not look at what the types are, it just looks at the proteomics data nothing else and it says based on the similarity of the proteomics data, I think these things fall together and there are three groups. It is up to you to go and look at those samples and say are they basal samples, are they luminal samples, what kind of samples are they.

So, that is the post clustering analysis that you have to earn your own. So, in the paper we looked at it and we did a pathway analysis and this and that and we try to say that one was a basil like cluster, other had mostly luminal samples and the third one was I think we called it a stroma enriched cluster where there was more of a stromal signature in the samples. So, all those are analysis you do after the clustering is done. During clustering the algorithm does not know about any of that. So, I think there have been newer studies where the consensus matrix says 3 I think, but when we look at 6 clusters its more biologically informative that is possible. So, again like I said this is more black art than a strong science so.

Student: Yeah.

(Refer Slide Time: 24:23)



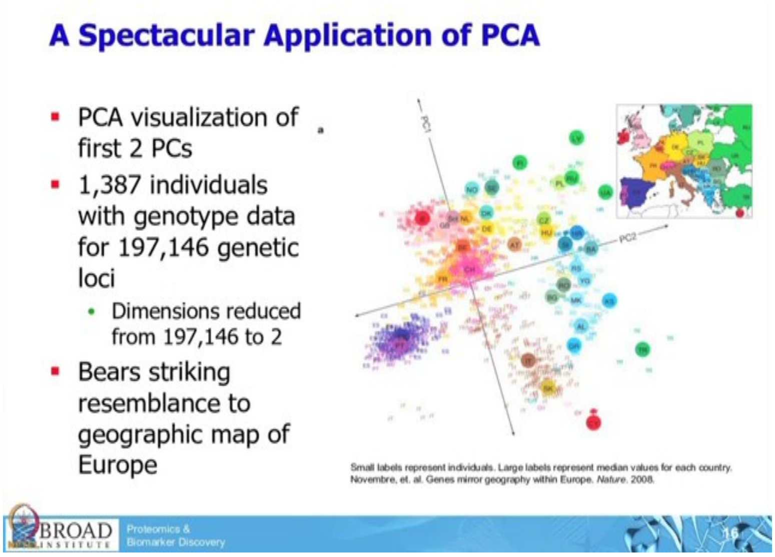
Principal Components Analysis (PCA)

- Exploratory data analysis tool, primarily used for unsupervised learning
- Goal:
  - Reduce data dimensionality
  - Improve interpretation by explaining data variability using a small number of variables ("principal components")

Matthias Scholz, Ph.D. thesis

Principle component analysis I briefly mentioned, it is primarily a visualization mechanism; people use to look at data and the analysis is done without looking at any a labels or groups. But, then once the PCA has been done you color your samples using groups or whatever to see how the separation is. So, it is more a visualization mechanism.

(Refer Slide Time: 24:47)



A Spectacular Application of PCA

- PCA visualization of first 2 PCs
- 1,387 individuals with genotype data for 197,146 genetic loci
  - Dimensions reduced from 197,146 to 2
- Bears striking resemblance to geographic map of Europe

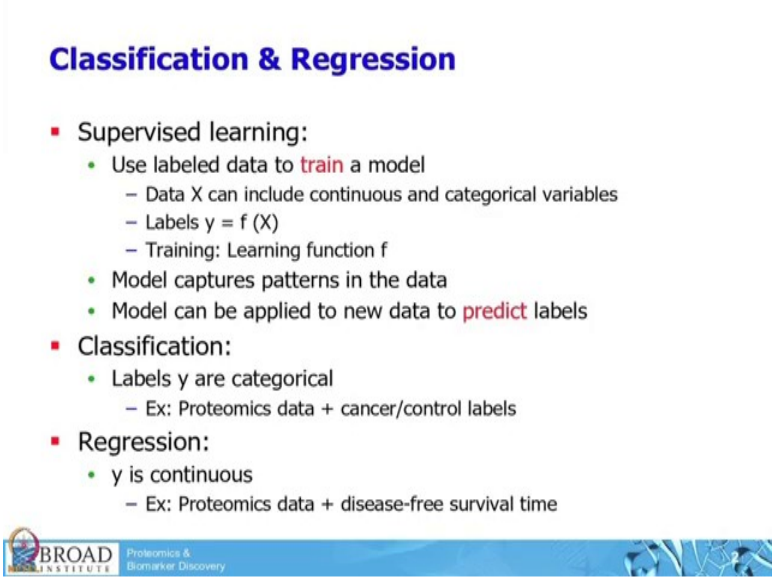
Small labels represent individuals. Large labels represent median values for each country.
Novembre, et. al. Genes mirror geography within Europe. *Nature*. 2008.

So, here is a really good example of how PCA can help in visualizing things. So, this is a PCA visualization of about 1400 people. We are looking at the genotype of that may people and for each person they have measured approximately 200 1000 genomic loci. So, they have characterized the loci for that many locations for about 1400 people and then they have done PCA. So, the original number of dimensions is the number of measurements you made and so that is about 200 1000 and that is now collapsed into two dimensions. So, when you do PCA, you get these principle components which are a kind of dimensions in which there is maximum change in your data.

So, the first principle component is the dimension in which there is maximum variation, the second one is the second most variation and so forth. So, when you look at the dimension two dimensions with that highest variation in your data and you plot them, you get a set of points. And what they have done here? So, the set of points are obtained without looking at any labels, they just look at the genetic loci that have been characterized for all the samples and each dot is a sample.

So, the issue with normalization, I think what he raised some types of clustering actually require you to normalize your data more strictly, you want all your proteins on the same scale. So, in that case in addition to normalizing your samples, you may need to normalize your proteins also otherwise your distance calculations will be biased by proteins that have high ratios or high intensities. So, again there you have to be a little more careful on what kind of input the algorithm needs and how the algorithm works.

(Refer Slide Time: 26:45)



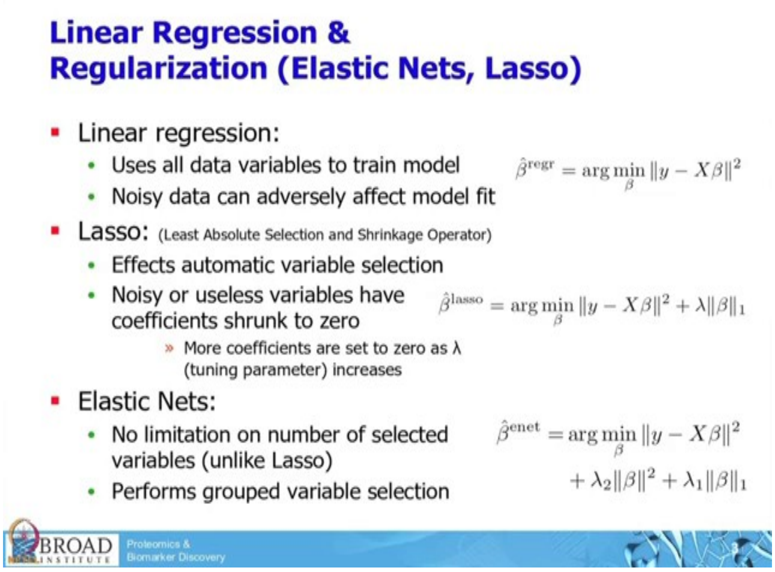
Classification & Regression

- Supervised learning:
  - Use labeled data to train a model
    - Data X can include continuous and categorical variables
    - Labels y = f (X)
    - Training: Learning function f
  - Model captures patterns in the data
  - Model can be applied to new data to predict labels
- Classification:
  - Labels y are categorical
    - Ex: Proteomics data + cancer/control labels
- Regression:
  - y is continuous
    - Ex: Proteomics data + disease-free survival time

So, one thing about supervised learning is if you had three cancer types; let say you had basal breast cancer, luminal A and luminal B breast cancer in your training data and then you give it a sample that is normal. So now, it has never seen a sample that is normal and it knows only about three classes; basal, luminal A and luminal B. So, it is going to try to force fit your sample into one of those three groups. So, there is you the though in most machine learning supervised models, there is no way to say this does not belong in these classes.

There is some other class, I do not know about there are some algorithms that do that, but they are very complex and not easy to use and not commonly available. So, most algorithms will not be able to say that this is a group, they have not seen before. All they are going to do is take it and fit it into one of the groups that they have already seen. So, in designing your training and setting up your analysis, you want to keep that in mind. So, you want to use all types of labels in your training so, that it can predict all types that you would encounter later on.

(Refer Slide Time: 27:59)



Linear Regression & Regularization (Elastic Nets, Lasso)

- Linear regression:
  - Uses all data variables to train model
  - Noisy data can adversely affect model fit

$$\hat{\beta}^{\text{regr}} = \arg\min_{\beta} \|y - X\beta\|^2$$

- Lasso: (Least Absolute Selection and Shrinkage Operator)
  - Effects automatic variable selection
  - Noisy or useless variables have coefficients shrunk to zero
    - » More coefficients are set to zero as λ (tuning parameter) increases

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \|y - X\beta\|^2 + \lambda\|\beta\|_1$$

- Elastic Nets:
  - No limitation on number of selected variables (unlike Lasso)
  - Performs grouped variable selection

$$\hat{\beta}^{\text{enet}} = \arg\min_{\beta} \|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1$$

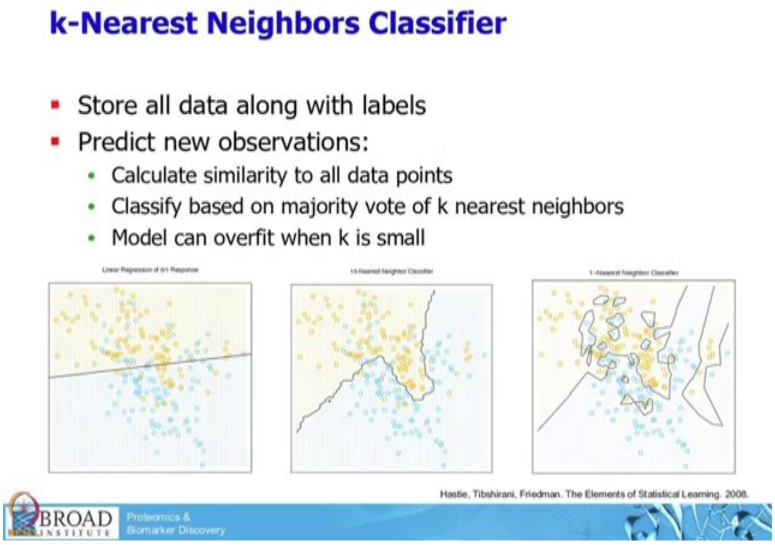
So, there are many ways for doing regression which is like predicting a continuous value. I will not go into the details, but the most simple one is linear regression which is like fitting a straight line. So, I am sure most of you have done this in high school or you have x and y values, you want to find the line of best fit. So, you can do that in a more dimensions than just two and that is linear regression.

Linear regression is very noisy when you have lots and lots of proteins and not enough examples for each of those and so, people have tried to make it more robust by using what is called regularization. So, Lasso and Elastic nets are different ways to make linear regression more robust so, that it can handle noise better. So, suppose you have 10,000 proteins, you have observed and you try to predict cancer you predict survival which is the real value.

So, for a patient you want to say how long they are going to live given that this is the profile you have in the proteome. So, not all 10,000 genes or proteins are useful in actually making the prediction, a lot of them are just noise and they do not change or not they are not related to making the prediction about survival. They are only a small subset that actually have any relation to survival so, but you do not know which ones they are. So, you want to try to kind of build your model using things that are relevant, but leave out the others.

So, that is called feature selection and some of these other more robust methods. So, in linear regression, your you have to give all the 10,000 and it will try to build a model and it will be very noisy because only let us say, a 100 out of the 10,000 actually have any information about predicting survival. But with these other methods, it will automatically say a kind of filter out things that do not have any information and try to narrow down the number of features it uses to the most useful in order to make the prediction. So, they end being a more robust models when you are looking at real data.

(Refer Slide Time: 30:08)



k-Nearest Neighbors Classifier

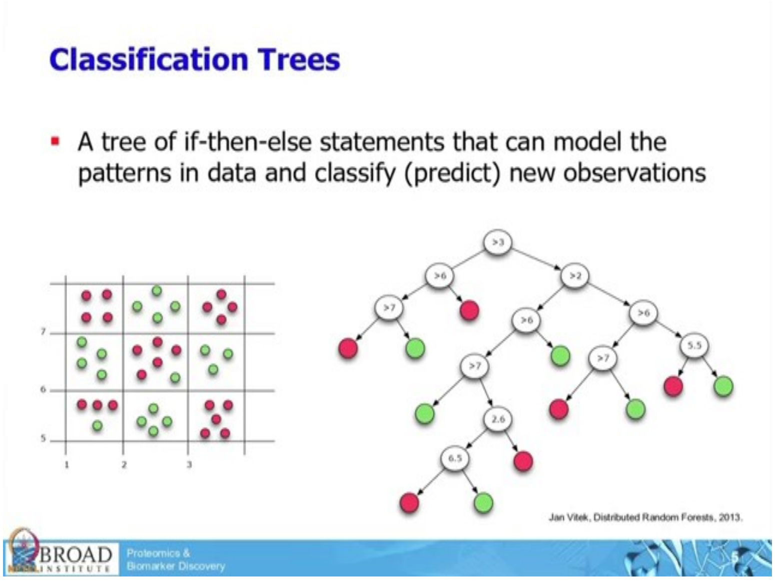
- Store all data along with labels
- Predict new observations:
  - Calculate similarity to all data points
  - Classify based on majority vote of k nearest neighbors
  - Model can overfit when k is small

Hastie, Tibshirani, Friedman. The Elements of Statistical Learning. 2008.

So, there are lot of classification algorithms for predicting groups, I will not go into each one of those. So, k nearest neighbor is basically you pick a sample and you make your prediction based on looking at the samples that are closest to you. So, you have seen cancer and normal samples, you get a new sample; you calculate the distance from this sample to all the samples you know and then you see which are my three closest samples.

And if my three closest samples are a cancer, then I say this is a cancer. If they are a normal, I say if it is a normal; if they are half and half I make a prediction based on majority with some probability saying, I think this is the cancer sample with probability 0.6 something like that.

(Refer Slide Time: 30:47)



Classification Trees

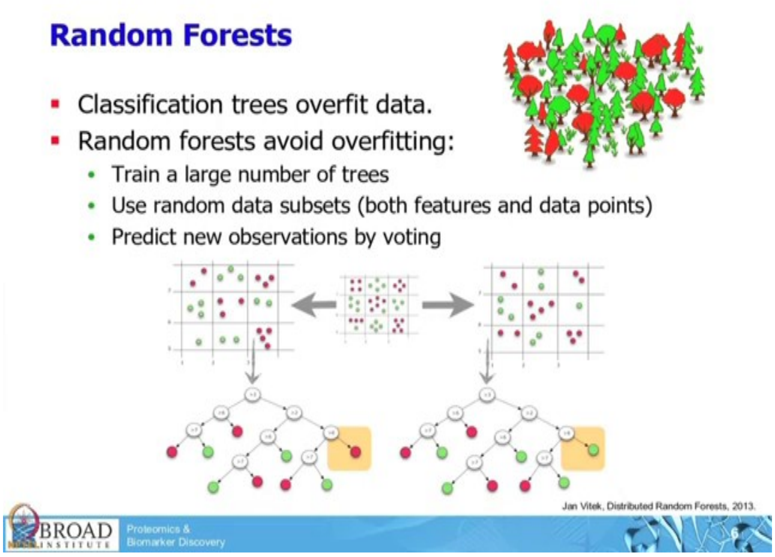
- A tree of if-then-else statements that can model the patterns in data and classify (predict) new observations

Jan Vitek, Distributed Random Forests, 2013.

BROAD INSTITUTE Proteomics & Biomarker Discovery

Classification trees result in models that you can understand because they say if this protein is less than some value and this other protein is greater than some value, then it is a cancer. So, it is a set of if then statements that will tell you which group to put it in. So, they are very interpretable and people like it because they can see the model and understand what it is doing; the other ones many times you cannot.

(Refer Slide Time: 31:14)



Random Forests

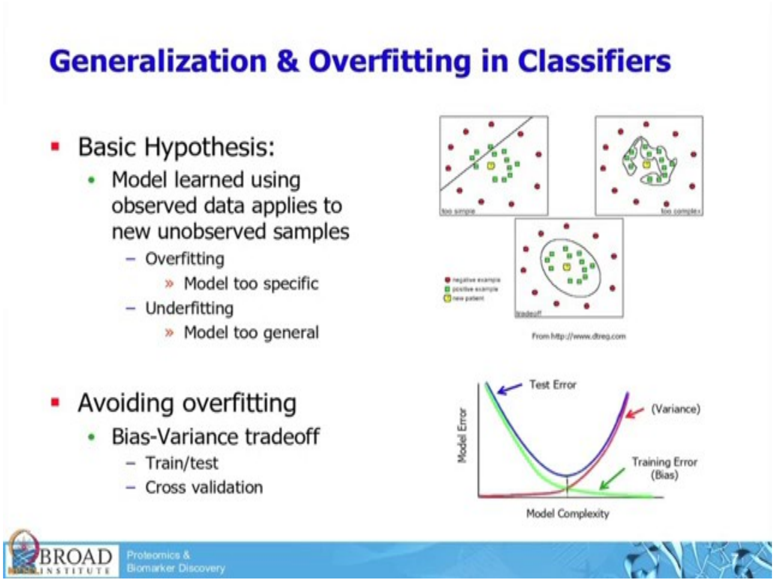
- Classification trees overfit data.
- Random forests avoid overfitting:
  - Train a large number of trees
  - Use random data subsets (both features and data points)
  - Predict new observations by voting

Jan Vitek, Distributed Random Forests, 2013.

BROAD INSTITUTE Proteomics & Biomarker Discovery

A random forests are a collection of trees, you all know that forest is a collection of trees, but these are random in the sense that they are built by making random subsamples of your data. So, the result is a lot more robust than just a tree. So, if it is built just one tree and there are quirks in your data that do not usually occur in your population, it will fit the model to the quirks and not be generalizable. So, when you get a new sample, it will make a wrong prediction. So, in order to avoid that in random forests, you build a lot of trees with random changes so, that the collection of trees is more robust than just a single tree. And when you make a prediction, you use all the trees to make the prediction.

(Refer Slide Time: 31:58)



Generalization & Overfitting in Classifiers

- Basic Hypothesis:
  - Model learned using observed data applies to new unobserved samples
    - Overfitting
      - » Model too specific
    - Underfitting
      - » Model too general
- Avoiding overfitting
  - Bias-Variance tradeoff
    - Train/test
    - Cross validation

From http://www.dtreg.com

BROAD INSTITUTE Proteomics & Biomarker Discovery

So, I have mentioned that models could be too specific or not specific enough. So, the concept is called generalization. So, the basic hypothesis is that you learn a model using data and then your model is at the right level so, that it is not too specific and not too general. And how you get there is basically by trying to avoid over fitting and picking a model that has enough parameters to kind of fit the data you have. So, that is a lot of hand waving, but here is an example.

So, let us say you want to separate the green dots from the red dots. If you use a model that is too simple that does not have enough parameters to tweak. So, the linear regression a straight line has only thing, you can tweak which is the slope. So, you can tweak the slope till you are blow in your face, but you can never separate the red dots from the green dots.

So, here the model is over simplified, you have a model that is too simple to fit the data you have. If you had points that lied that were only in the top and the bottom, the red points where all here and the green points where all here; a linear line could separate those. But here your data is complex enough that you cannot use the line to separate them. So, what you can do? I can say ok, I am going to have lot of parameters I can change. So, I can draw an arbitrary line around my points so, that is this one.

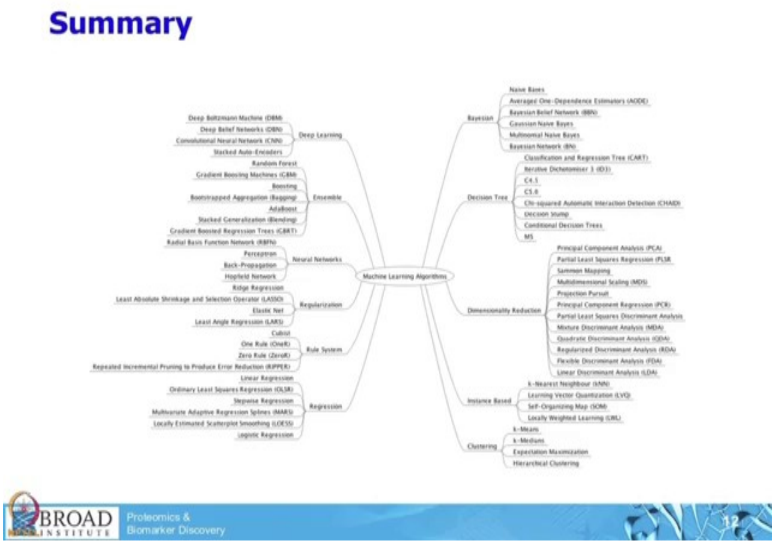
So, now it is so, specific to the green that if you get a new green that was right here next to the yellow, it is going to be outside the green points and so, it will make a wrong prediction. So, this model is over fit. It is so, specific to the small example you saw that it did not capture what was happening in reality. This is a more correct model; this is basically like a ellipse and so, this has enough parameters to overcome the restriction of this, but not too many so, that you do not over fit.

So, the whole point in machine learning is to come up with models that are generalizable to new data without overfitting data that you have in your data set. Your data set is an example with noise and things that are that could may be only a specific to the data set. If you look hard enough you can find lot of patterns, but only some of those patterns hold with everybody in the population and some are just specific to the quirk that you have these 15 or 100 samples in your data set and to avoid that you want a model that is generalizable.

So, to do that when you are training what you do is you take your data set. So, you have 100 samples you take a random subset of 25 samples and keep it aside. You use the remaining 75 samples chosen at random to build your model and then every so, often you use the test data set to say how am I doing; is my test data set getting good accuracy or is it poor accuracy?

So, what would happen is as your model becomes more and more complex or you train your model more and more, your training error is going to drop because it is doing this. In cross validation what you do is, you take your data set, you split it into 5 pieces, you learn on 4 pieces and predict on the fifth one. And then you keep cycling through. So, you are using all your data to learn and you are not keeping anything aside only for testing because you do not have enough data, but you are still getting an assessment of training and test error.

(Refer Slide Time: 35:45)



Summary

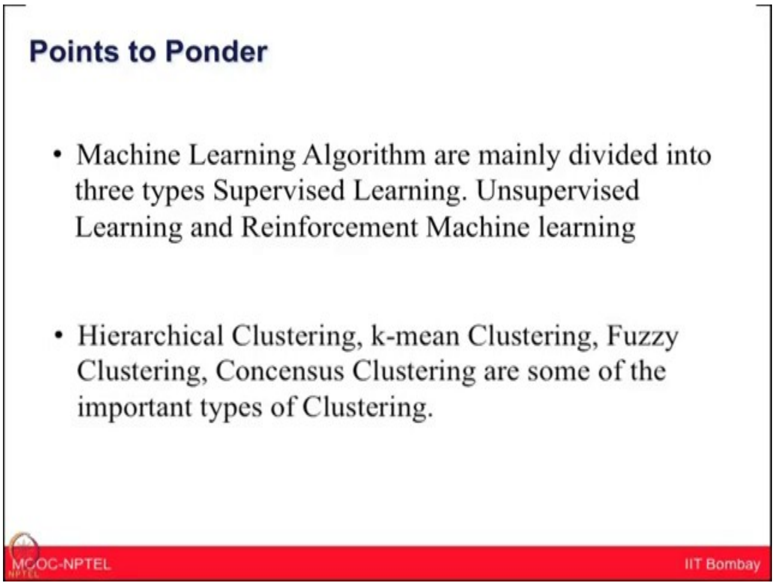
Machine Learning Algorithms

- Deep Learning
  - Deep Boltzmann Machine (DBM)
  - Deep Belief Networks (DBN)
  - Convolutional Neural Network (CNN)
  - Stacked Auto-Encoders
- Ensemble
  - Random Forest
  - Gradient Boosting Machines (GBM)
  - Boosting
  - Bootstrapped Aggregation (Bagging)
  - AdaBoost
  - Stacked Generalization (Blending)
  - Gradient Boosted Regression Trees (GBRT)
- Neural Networks
  - Radial Basis Function Network (RBFN)
  - Perceptron
  - Back-Propagation
  - Hopfield Network
- Regularization
  - Ridge Regression
  - Least Absolute Shrinkage and Selection Operator (LASSO)
  - Elastic Net
  - Least Angle Regression (LARS)
- Rule System
  - Cubist
  - One Rule (OneR)
  - Zero Rule (ZeroR)
  - Repeated Incremental Pruning to Produce Error Reduction (RIPPER)
- Regression
  - Linear Regression
  - Ordinary Least Squares Regression (OLSR)
  - Stepwise Regression
  - Multivariate Adaptive Regression Splines (MARS)
  - Locally Estimated Scatterplot Smoothing (LOESS)
  - Logistic Regression
- Bayesian
  - Naive Bayes
  - Averaged One-Dependence Estimators (AODE)
  - Bayesian Belief Network (BBN)
  - Gaussian Naive Bayes
  - Multinomial Naive Bayes
  - Bayesian Network (BN)
- Decision Tree
  - Classification and Regression Tree (CART)
  - Iterative Dichotomiser 3 (ID3)
  - C4.5
  - C5.0
  - Chi-squared Automatic Interaction Detection (CHAID)
  - Decision Stump
  - Conditional Decision Trees
  - M5
- Dimensionality Reduction
  - Principal Component Analysis (PCA)
  - Partial Least Squares Regression (PLSR
  - Sammon Mapping
  - Multidimensional Scaling (MDS)
  - Projection Pursuit
  - Principal Component Regression (PCR)
  - Partial Least Squares Discriminant Analysis
  - Mixture Discriminant Analysis (MDA)
  - Quadratic Discriminant Analysis (QDA)
  - Regularized Discriminant Analysis (RDA)
  - Flexible Discriminant Analysis (FDA)
  - Linear Discriminant Analysis (LDA)
- Instance Based
  - k-Nearest Neighbour (kNN)
  - Learning Vector Quantization (LVQ)
  - Self-Organizing Map (SOM)
  - Locally Weighted Learning (LWL)
- Clustering
  - k-Means
  - k-Medians
  - Expectation Maximization
  - Hierarchical Clustering

BROAD INSTITUTE Proteomics & Biomarker Discovery 12

This last slide is a summary of supervised and unsupervised machine learning methods; I am sure you cannot read it, but you can take a look at it on your computers. It is a reasonable collection of algorithms, each one has its quirks, each one is more appropriate in some situations than others and you can kind of take a look at it.

(Refer Slide Time: 36:01)

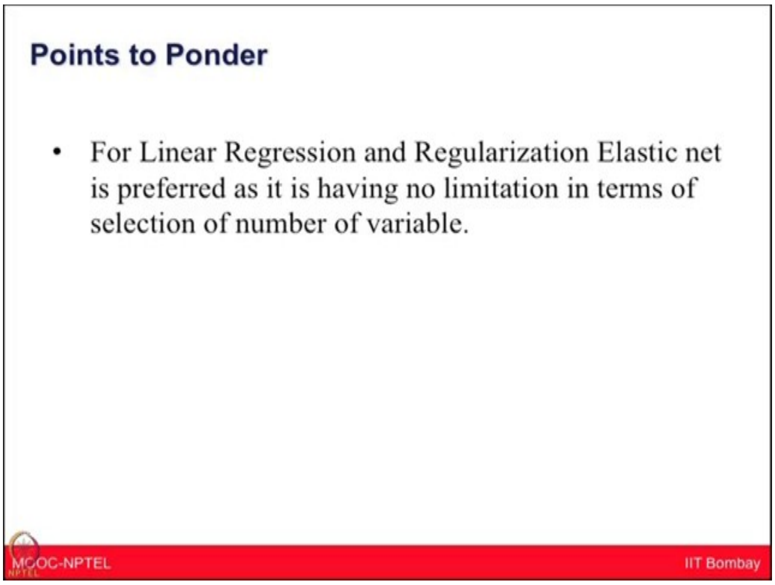


Points to Ponder

- Machine Learning Algorithm are mainly divided into three types Supervised Learning. Unsupervised Learning and Reinforcement Machine learning
- Hierarchical Clustering, k-mean Clustering, Fuzzy Clustering, Concensus Clustering are some of the important types of Clustering.

MOOC-NPTEL IIT Bombay

(Refer Slide Time: 36:15)



Points to Ponder

- For Linear Regression and Regularization Elastic net is preferred as it is having no limitation in terms of selection of number of variable.

MOOC-NPTEL

IIT Bombay

So, today we have learnt that how machine learning applications and omics data analysis when taken into account can provide you the good biomarker candidates. Of course, selecting right biomarker depends on what type of question you are looking for, what type of samples you have analyzed, what was the number of samples and which type of test you have used. After many of these careful considerations only, you might be able to obtain a right candidate biomarker from your entire analysis.

In today's lecture, we have also heard that organization data into clusters shows the internal structure of the data. In linear regression and regularization session, we also learnt about Lasso and elastic nets. The elastic net is preferred as it is having no limitations in terms of selection of the number of variables. Finally, we understood that we should avoid over fitting of data. In the flow of the lectures Dr. Mani will talk to you about hypothesis testing in next lecture.

Thank you.