

**Introduction to Proteogenomics**  
**Dr. Sanjeeva Srivastava**  
**Dr. David Fenyo**  
**Department of Bioscience and Bioengineering**  
**Indian Institute of Technology, Bombay**  
**New York University**

**Lecture - 48**  
**Association/ Marker Selection**

Welcome to MOOC course on Introduction to Proteomics. In today's lecture Dr. David Fenyo will talk about Association and Marker Selection. Marker selection helps to model an easy interpretation taking few features into account. Most of the protein candidates may not directly related to the phenotype, so marker associated selection or mass help us to feature those candidates and build a good predictive model.

Dr. Fenyo will also discuss about how many features need to be considered for building reliable model, and why multiple features makes a model complex, what is the optimal number of features. He will briefly discuss different kinds of methods which are available which could help in the feature selection. He will also highlight what is data snooping. So, let us welcome Dr. David Fenyo for his last lecture.

(Refer Slide Time: 01:33)

Marker Selection

**Few features**

Easy to interpret  
Less likely to overfit  
Lower prediction accuracy

**Many features**

Difficult to interpret  
More likely to overfit  
Higher prediction accuracy



The other thing is marker selection that is we had already mentioned earlier. So, now, we do all these measurements and we you we know that most of the proteins or most of the

transcripts are not going to be related to our phenotype. So, we really, it would be much better to just have build the model using the ones that we know are related, but of course, we do not know which ones to start with. So, we need to; so, there if we look at mark, so by the we do marker selection. So, the having few features its makes the model easier to interpret.

So, one thing that we have talked about building these predictive models and we want to predict something, but if we can also understand that is of course, a much better thing. And often when we build very complex models we do not understand and maybe will not have a chance to understand. And few features; so, it is easier to interpret we can start thinking about biological function and they are also less likely to over fit because few are parameters and, but usually they get a little bit lower prediction or accuracy. So, that is something to balance and that is what we use to then decide how many features.

So, as suppose to if you have many features it is difficult to interpret, we do not know what is going on. And then of course, more likely to overfit because we do have an enormous amount of parameters, but of course, as we add in more and more things we get higher prediction accuracy, but it we are not sure whether that is really the real.

(Refer Slide Time: 03:34)

## Marker Selection

### Filter methods

- Predictive power of the variables in evaluated separately
- High correlation with target variable
- Low correlation between predictors
- The higher the information value, the better is the variable



So, there are a few different ways to do this. One set of methods are called filtering methods. So, in these we look at what is the predictive power of each protein. And then of course, and then they select the one with a the highest predictive power.

And so, we look for which proteins are have high correlation with target variables of whatever we let us say the tumor subtype, we also do not want lot of predictors that are correlated with each other. We want them to be somewhat uncorrelated and also we want them to have lots of information and that is pretty much to same as high correlation with the target value. So, but this is now we look at evaluate each individually.

(Refer Slide Time: 04:39)

## Marker Selection

### Wrapper methods

- Predictive power of the variables in evaluated jointly
- Set of variables that performs the best:
  - Subset selection
  - Forward selection
  - Backward selection



104

And now the other classes these wrapper methods where we look at the predictive power jointly. And so, the idea is that it is they are not independent of each other, so probably the we will get a better result using the very evaluate them jointly. So, they are two ways. So, one is that we start with the let us say the one that has most information than we had the second one, we evaluate them together and then was there with what we mean by evaluating as we check is there any additional point of adding the second one and that it improve our results.

And then, they continue adding until it is not the results do not improve anymore. And of course, this again we have to do this with the in cross validation because otherwise it will over fit. And then, that was forward selection backward selection is that we remove instead. And then, we can also there are people have combined this, so one very popular way of doing this is called recursive feature extraction which is probably the most widely used one.

(Refer Slide Time: 06:11)

## Marker Selection

### Embedded methods

- Methods without explicit feature selection, e.g. regularization using Lasso.



105

And so, one some methods like Lasso; so, Lasso was when be regularized adding in the absolute value of the parameter vector times a constant. There we actually get explicitly feature extraction and the some variables will fall out, will be, I mean some parameters will become 0.

(Refer Slide Time: 06:39)

## Marker Selection

### What is the optimal number of features?

- The ideal model should do justice to both:
  - Good prediction yet not overly complex to interpret



106

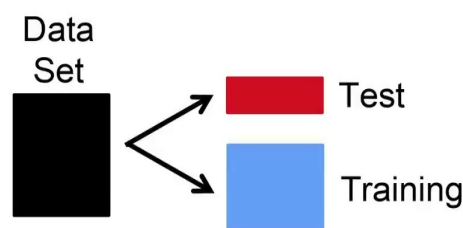
So, then the marker selection, so the question is what is the optimal number of features. So, that is usually we want two things. We want it to be as simple as possible, so we can interpret it and but we want to get good predictions still. So, if you one thing that people talk about is

the curse of their dimensionality which we definitely have all those is that we have the measure of few samples, even a I mean within CPTAC you measure 100 samples that still quite view.

But, for each sample we measure tens of thousands of I mean 10,000 proteins maybe 30,000 transcriptions are another 30,000 phosphorylation sets. So, we have much more measurements on each variable than we have samples. And, this makes things very hard first of all not to overfit, but also often when we find signatures they are not unique. But, we have a large we could there are many signatures that are equally would make equally good predictive models, ok.

(Refer Slide Time: 08:11)

## Choosing Hyperparameters



Examples of hyperparameters:

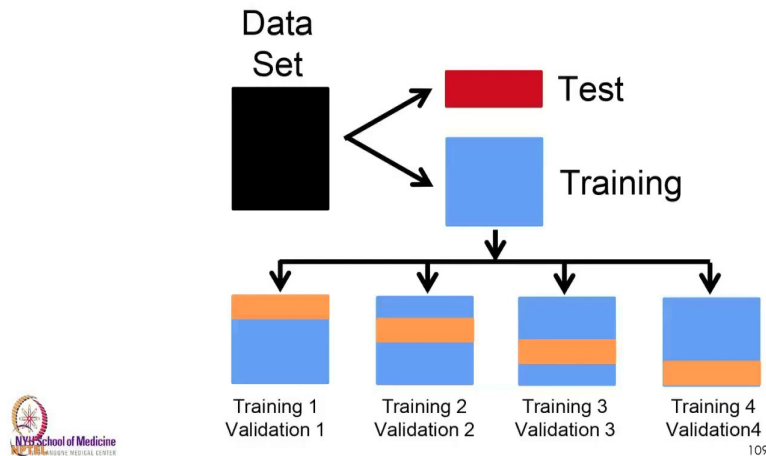
- Learning rate
- Regularization parameter
- Machine learning method



So now, finally, cross validation, so we have all these hyper parameters that we need to decide on, but what we say we do have our data set that we divide into training test. But, we need to decide on these variables, and we not allowed to use the test set to decide on the hyper parameters.

(Refer Slide Time: 08:35)

## Cross-Validation: Choosing Hyperparameters



So, what we do is we further divide the training set, but we divide it many times. So, here we have taken the blue regional training set as the training that we actually do the training and then we use this yellow as validation. For validation meaning that we define for example, the learning rate, the regularization rate and so on, and then we do this many times for different subsets.

So, for example, 5 fold cross validation or 10 fold cross validation are commonly used. And, we can even since we do have limited data sets and we often also do that we do another's they do not do this division of training test, but to what is called nested cross validation. So, we do a cross validation down here, but then we do a similar cross validation up here on top of it. So, that is and this is very important to do well, but a lot of the software packages do have this built in.

(Refer Slide Time: 10:04)

## Sampling Bias



110

So, another few things I wanted to just mention I will only take a few more minutes and so, one is sampling bias. So, that is, so the really the machine learning method will only give us what we trained it for. And so, this is a classical example of in the this election the Truman won, but the polls, the polling companies, the way they did the polls is that they called people on the home phone.

And I forget this was in the late 40s, so only rich people had phones. So, and they preferably bought mostly vote to republican and they, so the polls got it completely wrong. So, here is a newspaper that actually printed it in advance because they were so sure that the Truman would lose.

(Refer Slide Time: 11:09)

## Sampling Bias

	Involving people	Involving specimens
<b>Experimental study (for example, randomized controlled trial)</b>		
Design	Randomize allocation to compared groups at baseline	Arrange for uniform (and, if possible, blinded) collection, handling and analysis of specimens
Conduct	Measure and report baseline characteristics of groups	Check to see whether uniform handling occurred and whether blinding was successful
Interpretation	If groups are unequal, discuss direction, magnitude and potential impact of bias	If groups are unequal, discuss direction, magnitude and potential impact of bias
<b>Observational study</b>		
Design	Avoid heterogeneity in selection; or stratify subjects in a way that minimizes differences between groups	Find specimen groups that have minimal differences; or, where possible (and it is usually not), arrange for uniform and blinded collection, handling and analysis of specimens
Conduct	Measure and report baseline characteristics of groups	Measure and report details of how specimens in each group were collected, handled and analyzed
Interpretation	Discuss possible biases and their direction, magnitude and potential impact	Discuss possible biases and their direction, magnitude and potential impact
Example	Subjects in one group are old and have multiple illnesses; subjects in the comparison group are young and healthy	Collection: blood specimens for the cancer group, from clinic number 1, sit for 6 hours before being separated and frozen; specimens for the non-cancer group, from clinic number 2, are immediately separated and frozen Handling: cancer specimens have been thawed and refrozen five times; the non-cancer specimens only once Analysis: cancer and non-cancer groups are analysed on different days; if the machine 'wanders' over time, 'signal' may inadvertently become introduced into the data



DF Ransohoff, "Bias as a threat to the validity of cancer molecular-marker research", Nat Rev Cancer 5 (2005) 142-9.

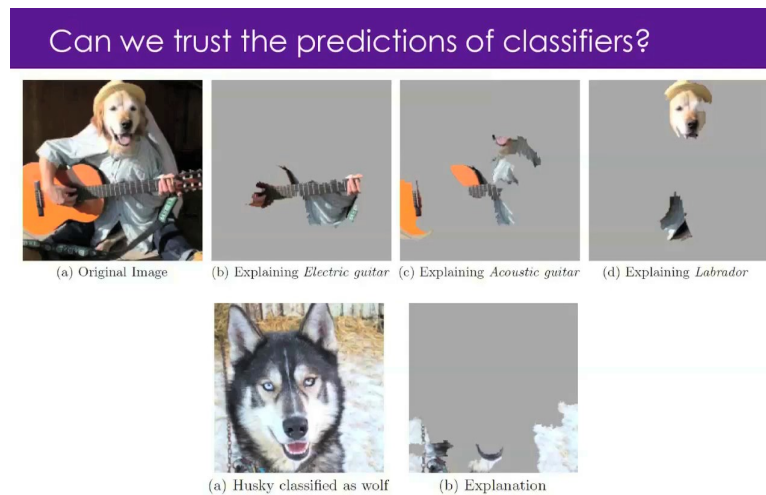
111

And this is something that happens to us a lot in biomarker discovery and so, this you will have these slides, so it is definitely worth reading. So, David Ransohoff has written several papers on this problem and here this is just some lists of what can go wrong. But it is definitely by should spend quite a lot of time thinking about what for example, if one would develop a blood test for early discovery; one should not collect the normal samples in a different clinic than for the samples from people with that have cancer.

It should be, but there, so there is a lot of its worth reading these and a lot of things to think about. So, then the again I am said this several times, so the test set data has to be independent, otherwise it is not if you train if you test your model on something you have trained with its really not going to tell you how good the model is. So, we talked about a little bit than they have very complex models, it is difficult to know when if the model tells us something about reality.



(Refer Slide Time: 12:52)




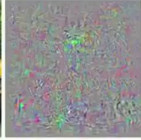

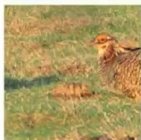
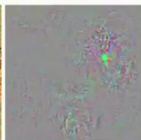

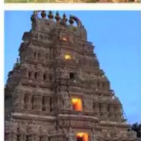
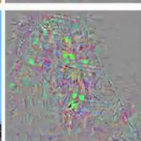
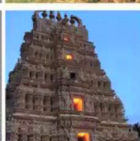
Ribeiro, Singh and Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier", In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2016 114

So, one thing especially with images it is easy to. So, here we have a an image and I think the predict the neural network in this case was able to say that it this was it found that it there is an electrical guitar it thought from this, acoustic guitar from that, and Labrador from that. So, that sounds pretty reasonable. But often they we can have this case where this was classified as a wolf, but what was used for classification was the background snow.

And this can easily happen in proteogenomics that we since we have, so maybe we if we build a very complex model we can get something that is irrelevant; what happened probably was in this in the training set all the wolves had snow in the background and that is something that can happen.

(Refer Slide Time: 14:05)

### Adversarial Fooling Examples
















Original correctly classified image	Perturbation	Classified as ostrich
		
		
		

Szegegy et al., "Intriguing properties of neural networks", <https://arxiv.org/abs/1312.6199>

So, another thing that is still with image analysis with all its success. Here we have an original image three images that were classified like you would you see it, but then if you add a little bit of perturbation that is barely, you can barely see any difference. Then all these three images are classified as ostriches. And so, so there is a lot than we especially with complex deep learning methods, there are lot of things that we do not understand.

(Refer Slide Time: 14:42)

### Adversarial Fooling Examples

					
network decision:	9	8	2	9	1
	"Normal" images are classified perfectly well but...				
					
network decision:	3	3	9	4	4
	... small modifications can completely derail the network decisions.				
					
	These images are classified with high confidence as "zero" by a neural network.				

Wieland Brendel, <https://medium.com/bethgelab/ai-still-fails-on-robust-handwritten-digit-recognition-and-how-to-fix-it-a432d84ede18>

And it is actually even worse these are quite complex images, but even for simple handwritten digits, so these are classified correctly. So, this is the handwritten image and

below is the, but then neural network says, but again if you added a little bit of noise you see that it barely definitely does not disturb us. We can still see clearly what it is. And then for example, this 9 here becomes according to the network becomes 3 and this is now people that develop this that were classification knew about this problem and try to fix it, but did not succeed. And even worse these are all classified as 0, even though there is nothing there.

(Refer Slide Time: 15:32)



So on, few books these are very easily accessible books I would say, An Introduction to Statistical Learning, more Applied Predictive Modeling. Both of them teaches you it gives you a good starting point for starting to do predictive modeling and feature extraction.

The other thing I recommend and then we are going to start that during hands on session, you really no need to learn how to program it, there is no way around it. And so, this is a good starting point R is probably; since you now have all have R studio installed and you should go home and continue using it and this there is a PDF available of this book online. So, I think all of these books are available as PDFs online also.

(Refer Slide Time: 16:37)

## Learning Objectives

### **How to train a model:**

- Gradient descent
- Regularization
- Feature selection
- Selection of machine learning method

### **How to test a predictive model:**

- Overfitting and underfitting



118

So, I hope that you have learned a little bit about how to train predictive models and then how to test them to avoid overfitting.

Student: Hi, David I have a more general question. So, you gave us example of 1940s presidential election 2 min ago. So, we have to nominate in 2018, but we still seems to be getting this thing wrong time to time getting.

Yes, 2016 was another example.

Student: That was an good example and right now again so I am sure we will get it wrong, many of the outlets will get it wrong. So, what is your take on that? Why we are still getting wrong?

I mean that I am not do not know that much about predicting elections, but in general in let us say biomarker discovery or I mean we get it wrong for one thing is that we do not think about it maybe well enough, and we take shortcuts. That is and, but it is also very difficult to do it right. That is the main thing.

Student: Number of variant are there.

Yeah, and also I mean it is not easy to collect enough sample, so we have to do them over many hospitals and it gets very complicated and it is actually, it is not easy to think it through.

Student: So, you mentioned that the normal tissues in cancerous normal that would not be collected from a hospital which was removed from the cancerous. So, one is there could be cancerous here, like if you get the samples from the same hospital there could be cancerous and the other is such a scenario is generally not seen where normal tissues and cancer tissues are same.

So, there could be another institute where other regions different combinations are being held and for which are not really cancerous, but other diseased tissues which we could consider as normal, but not cancerous let say normally; cancerous not cancerous normally not cancerous could be used. So, in that kind of scenario how would we.

I mean.

Mani talked about some solution like trying to do batch correction and things like that, but again it is as I said better to avoid that if possible so, yeah. So, one should definitely try to have that is a normal or in cancer patients or I mean for example, for blood testing one could imagine that there is a clinic where people come, and when they come there they do not even know if their cancer, the doctors do not know.

So, that is a good situation that during the testing no one knows and they are treated the same way, and only after the samples are taken and have been tested, if that is if one can do that, that is the ideal situation I think. But usually, I mean often we do have to make compromises, but we should try to make as few compromises as possible.

Two comments to make, the political thing that we brought on is lot more complicated because human psychology and how people behave is involved. I think the example David gave was for showing bias, but one you brought about the 2016 election is a lot more complicated. So, who goes to that, who goes to vote is also part of the prediction and so, the polling does not take that into account in a proper way.

So, I hate Hillary Clinton so I get all my friends to come and work for Trump. So, that dynamic is not taking into account and calling is done. So, I think that is much more complex example that involves human behavior and psychology. So, I think that is how we can think for this course.

The other question that I want to make was David brought up feature election and he also talked about keeping your best data set separate and using cross validation. So, one of the common mistakes that people make and one of the mistakes I made too earlier on is to use your entire dataset to do features selection and then split your training datasets

Yeah. Yes.

So, that is a very bad. So, again if you are working in the business lab and you have millions of data points I think it does not make that big of a difference, but in biology where you have only 100 or 50 samples; if you do that then you already have your answers in your features. And so, you will do very well on first set, but the next set of new samples that come from the hospital you fill back.

So, I used to work previously in telephone industry where we get hundreds of thousands of samples or records, but when I move on to working on biological samples you really have to pay attention to not contaminating your training and test set and keep them separate.

Student: My question is that, I we have develop some brain bank management some feature related selected from a large number of features as well as small number of features. Now, when I see the patient database is somehow misses mass feature that feature is NA that would not be collected from patient. In such scenario multiple way again go back and develop new models and features selection strategy is recommendable?

Yeah. So, some machine learning methods are better at handling missing data like that. So, especially some of the tree based methods there you can set them up. So, that they can even if you have in your training set, if each of them in the test set not it can still work well.

Student: So are you not trying for it

So, the tree based method as David mentioned so especially the random forest when tried in these kind of data can come storing surrogate features, so when your main feature is missing it will surrogate feature instead of that. So, if you have that 5 surrogates for each main feature than if one or two of your main feature are missing use the surrogates. So obviously, not regression because the surrogate is as good or better then main feature that would have been main feature. So, if you do some performance but you can still be your prediction.

Student: Is the same thing that interpolation and fixing some more data point.

No, that is missing value imputation that is different. I show it that is that is another thing we can do that I can using surrogates is more robust way of dealing with it because when you impute so, when you will impute its one think to look at all your data and impute. But, when you have tested or which is just a few samples how do you impute; you use only the test data to impute or you can use entire training data and test at all impute. So, I do not recommend on combining everything to impute because then you are kind of making your test data look like your training data by design that is not what we want.

And so, the ideal way to be have large test training set you can try imputing, usually we get a few samples of prediction and in that case I think using a method that will deal with surrogates or missing features would be the best of course. I know we ran fast but I can do that. I think for some of the others you can calculate variable importance, but I do not know some methods other methods that to use surrogates automatically but not others

(Refer Slide Time: 25:07)

### Points to Ponder

- Most of the protein candidate may not directly relate to the phenotype, so MAS (Marker Associated Selection) help us to feature those candidate and build a good predictive model.
- Features should be chosen which can build a simple model and give correct prediction.
- Sampling biasness and constructing pre-planned interference should be avoided.



So, today Dr. Fenyo provided a brief idea about different biomarker selection methods and how it could help in optimal selection of features. We also learnt that cross validation, forward selection and backward selection plays an important role in feature selection. Lasso is a very good software to explore these kind of extraction features. We should choose features keeping in mind two important things, models should be as simple as possible so that we could interpret easily, at the same time the model should also provide a good prediction.

Today we also learnt curse of dimensionality. In simple words, the complex algorithm and data frame having a big number of dimensions or features frequently make the target function very complex and it may lead to the model overfitting. Finally, Dr. Fenyo talked about sampling biasness and biomarker discovery. He also mentioned about data snooping, it refers to the statistical inference that the researcher decides to perform after looking at the data.

We should avoid sampling biasness and construct pre-planned interferences. For example, a group of researchers planned to compare 3 dosages of a drug in clinical trial. They pre-planned the data comparison on the basis of record of patients and grouped the patients on the basis of that which is an example of data snooping.

The next lecture will be hands on session on WebGestalt by Dr. Bing Zhang.

Thank you.