Introduction

OK, now we're going to talk about the supervised learning problem and set down a little bit of notation. So we'll have an outcome measurement Y, which goes by various names, dependent variable, response, or target. And then we'll have a vector of p predictor measurements,

which we usually call X. They go by the name inputs, regressors, covariates, features,

or independent variables.

And we distinguish two cases. One is the regression problem, Y is quantitative, such as

price or blood pressure. In the classification problem, Y takes values in a finite,

unordered set, such as survived or died, the digit classes zero to nine, the cancer class

of the tissue sample. Now we have training data pairs. X1, y1, x2, y2, up to xN, yN.

So again, x1 is a vector of p measurements, y1 is usually a single response variable.

And so these are examples or instances of these measurements. So the objectives of supervised

learning are as follows.

On the basis of the training data, we would like to accurately predict unseen test cases, understand which inputs affect the outcome and how, and also to assess the quality

of our predictions and the inferences. So by way of philosophy, as you take this course, we want

not just to give you a laundry list of methods, but we want you to know that it is important to understand the ideas behind the various techniques, so you know where and when to use them.

Because in your own work, you're going to have problems that we've never seen before, you never seen before, and you want to be able to judge which method are likely to work well

and which ones are not likely to work well. As well, not just prediction accuracy is important, but it's important to do to try simpler methods first in order to grasp the more sophisticated ones.

We're going to spend quite a bit of time on linear models, linear regression, and linear logistic regression. These are simple methods, but they're very effective. And it's also important to understand how well method is doing. It's easy to apply an algorithm. Nowadays you can just run software, but it's difficult but also very important to figure out how well is the method actually working so you can tell your boss or your collaborator that, when you apply this method we developed, how well you're likely to do it tomorrow. And in some cases, you won't do well enough to actually use the method, and you'll have to improve your algorithm or maybe collect better data. The other thing we want to convey just through the course and hopefully through examples is that this is a really exciting area in research.

I mean, statistics in general is very hot area. Statistical learning and machine learning is of more

and more importance, and it's really exciting that the area's not gelled in anyway in the sense that there's a lot of good methods out there, but a lot of challenging problems aren't solved.

Especially in recent years with the onset of big data and coined the word data science.

Statistical learning is a fundamental ingredient in this new area of data science. So you might be wondering where's the term supervised learning come from? It's actually a very clever term, and it was developed by someone in the machine learning area. The idea of supervised learning, [? I'm thinking ?] in kindergarten of a teacher trying to teach a child to discriminate between what, say, a house is and a bike.

So he might show the child, maybe Johnny, say Johnny, here are some examples of what a house looks like.

And maybe in LEGO blocks. And here's some examples of what a bike looks like. He tells Johnny this and shows him examples of each of the classes. And the child then learns, oh I see.

Houses got sort of square edges, and a bike has got some more rounded edges, et cetera.

That's supervised learning, because he's been given examples of label training observations.

He's been supervised. And as Trevor just sketched out on the previous slide, the Y there is given and the child tries to learn to classify the two objects based on the features, the X's. Now, unsupervised learning is another thing that we're talking in this course. Which I grew up with.

See, that's the problem. So in unsupervised learning, now in the kindergarten, Trevor's in kindergarten, and Trevor was not given examples of what a house and a bike was. He just sees on the ground lots of things. He sees maybe some houses, some bikes, some other things.

And so this data is unlabeled. There's no Y. I was pretty sharp. So the problem there now is for the child, it's unsupervised, to try to organize in his own mind the common patterns of what he sees. He may look at the objects and say, oh these three things are probably houses. Or he doesn't know they're called houses, but they're similar to each other because they have common features.

These other objects, maybe they're bikes or other things, they're similar to each other, because I see some commonality.

And that brings the idea of trying to group observations by similarity of features, which is going to be a major topic of this course, unsupervised learning. So more formally, there's no outcome variable measure, just a set of predictors. And the objective is more fuzzy. It's not just to predict Y, because there is no Y. It's rather to learn about how the data is organized, and to find which features are important for the organization of the data.

So we'll talk about the clustering of principle components which are important techniques for unsupervised learning. One of the other challenges is it's hard to know how well you're doing.

There's no gold standard. There's no Y. So when you've done a clustering analysis, you don't really know how well you've done. And that's one of the challenges. But nonetheless, it's an extremely important area. One reason is that the idea of unsupervised learning is an

important preprocessor for supervised learning. It's often useful to try to organize your features, choose features based on the X's themselves, and then use those processed or chosen features as input into supervised learning. And the last point is that it's a lot easier, it's a lot more common to collect data which is unlabeled.

Because on the web, for example, if you look at movie reviews, a computer algorithm can just scan the web and grab reviews. Figuring out whether review, on the other hand, is positive

or negative often takes human intervention. So it's much harder and costly to label data. Much easier just to collect unsupervised, unlabeled data. The last example we're going to show you

is a wonderful example. It's the Netflix prize. Netflix is a movie rental company in the US, and now you can get the movies online. There used to be DVDs that were mailed out. And Netflix set up a competition to try and improve on their recommender system. So they created a data set of 400,000 Netflix customers and 18,000 movies, and each of these customers that rated on

average around 200 movies each. So each customer had only seen about 1% of the movies.

And so you can think of this as having a very big matrix which was very sparsely populated with ratings between one and five. And the goal is to try and predict, as in all recommender systems, to predict what the customers would think of the other movies based on what they rated so far.

So Netflix set up a competition where they offered a $1 million prize for the first team that could improve on their rating system by 10% by some measure. And the design of the competition is very clever.

I don't know if it was by luck or not, but the are root-mean-square error of the original

algorithm was about 9.953. So that's on a scale of, again, one to five. And it took the community, when they announced the competition and put the data on the web, it took the community about a month or so to have an algorithm which improved upon that. But then it took the community about another three years to actually for someone to win the competition.

It's a great example, here's the leader's board at the time the competition ended.

It was eventually won by a team called BellKor's Pragmatic Chaos. But a very close second was Ensemble. In fact, they had the same score up to four decimal points. And the final winner was determined by who submitted the final predictions first. So it was a wonderful competition, but what was especially wonderful was the amount of research that it generated. Tens of thousands of teams all over the world entered this competition over the period of three years, and a whole lot of new techniques were invented in the process. A lot of the winning techniques ended up using a form of principal components in the presence of missing data.

How come our name's not on that list, Trevor? Where's our team? That's a good point, Rob.

The page isn't long enough. I think if we went down a few hundred, you might. So actually, seriously, we actually tried with a graduate student when the competition started. We spent about three or four months trying to win the competition. And one of the problems was computation.

The data was so big and our computers were not fast enough. Just to try things out took too long.

And we realized that the graduate student was probably not going to succeed and was probably going to waste three years of his graduate program, which was not a good idea for his career.

So we basically abandoned ship early on.

So I mentioned in the beginning the field of machine learning, which actually led to the statistical learning area which we're talking about in this course. And machine learning itself arose as a sub field of artificial intelligence, especially with the advent of neural networks in the '80s. So it's natural to wonder what's the relationship between statistical learning and machine learning. First of all, the question's hard to answer. We ask that question often. There's a lot of overlap. Machine learning tends to work at larger scales. They tend to work on bigger problems.

Although again, the gap tends to be closing because fast computers now are becoming much cheaper.

Machine learning worries more about peer prediction and how well things predict. Statistical learning also worries about prediction but also tries to come up with models, methods that can be

interpreted by scientists and others. And also by how well the method is doing. We worry more about precision and uncertainty. But again, the distinctions become more and more blurred,

and there's a lot of cross-fertilization between the methods. Machine learning clearly has the upper hand in marketing. They tend to get much bigger grants and their conferences are much nicer places, but we're trying to change that, starting with this course.

So here's the course text, Introduction to Statistical Learning. We're very excited.

And this course will cover this book in its entirety. The book has at the end of each chapter, there's examples run through in our computing language. And we do sessions on R. And so when you do this course, you will actually learn to use R as well. R is a wonderful environment.

It's free, and it's a really nice way of doing data analysis. You see there's a second book there, which is our more advanced textbook, Elements of Statistical Learning.

That's been around for a while. That will serve as a reference book for this course for

people who want to understand some of the techniques in more detail. Now the nice thing is, not only is this course free, but these books are free as well. The Elements of Statistical Learning has been free, and the PDF is available on our websites.