Classification and Discrimination

Nicholas Brust

Looking to allocate a sample X into a population group.

Methods

Classification Methods

Methods for classification Include the following.

- Minimizing the ECM
- Minimizing the TPM
- Fishers Model
- Logistic Regression
- Deep Learning Models.

Fisher Discriminant Rule

An Allocation Rule Based on Fisher's Discriminant Function[5]

Allocate $\mathbf{x}_0$ to π_1 if

$$\hat{y}_0 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x}_0$$
$$\geq \hat{m} = \tfrac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$$

or (11-25)

$$\hat{y}_0 - \hat{m} \geq 0$$

Allocate $\mathbf{x}_0$ to π_2 if

$$\hat{y}_0 < \hat{m}$$

or

$$\hat{y}_0 - \hat{m} < 0$$

This is a really great first example of a basic discrimination/ thresholding function to discriminate data.

Discrimination Example SAS

```
title 'Two Sample Discriminant Analysis';
data bankrupt;
input x1 x2 x3 x4 group;
datalines;
-0.45  -0.41  1.09  0.45  0
-0.56  -0.31  1.51  0.16  0
0.06  0.02  1.01  0.40  0
-0.07  -0.09  1.45  0.26  0
-0.10  -0.09  1.56  0.67  0
-0.14  -0.07  0.71  0.28  0
0.04  0.01  1.50  0.71  0
-0.07  -0.06  1.37  0.40  0
0.07  -0.01  1.37  0.34  0
-0.14  -0.14  1.42  0.43  0
-0.23  -0.30  0.33  0.18  0
0.07  0.02  1.31  0.25  0
0.01  0.00  2.15  0.70  0
-0.28  -0.23  1.19  0.66  0
0.15  0.05  1.88  0.27  0
0.37  0.11  1.99  0.38  0
-0.08  -0.08  1.51  0.42  0
0.05  0.03  1.68  0.95  0
0.01  0.00  1.26  0.60  0
0.12  0.11  1.14  0.17  0
-0.28  -0.27  1.27  0.51  0
0.51  0.10  2.49  0.54  1
0.08  0.02  2.01  0.53  1
0.38  0.11  3.27  0.35  1
0.19  0.05  2.25  0.33  1
0.32  0.07  4.24  0.63  1
0.31  0.05  4.45  0.69  1
0.12  0.05  2.52  0.69  1
-0.02  0.02  2.05  0.35  1
0.22  0.08  2.35  0.40  1
0.17  0.07  1.80  0.52  1
0.15  0.05  2.17  0.55  1
-0.10  -0.01  2.50  0.58  1
0.14  -0.03  0.46  0.26  1
0.14  0.07  2.61  0.52  1
0.15  0.06  2.23  0.56  1
0.16  0.05  2.31  0.20  1
0.29  0.06  1.84  0.38  1
0.54  0.11  2.33  0.48  1
-0.33  -0.09  3.01  0.47  1
0.48  0.09  1.24  0.18  1
0.56  0.11  4.29  0.44  1
0.20  0.08  1.99  0.30  1
0.47  0.14  2.92  0.45  1
0.17  0.04  2.45  0.14  1
0.58  0.04  5.06  0.13  1
;
proc discrim data=bankrupt method=normal pool=yes manova wcov pcov listerr crosslisterr;
class group;
var x1 x2 x3 x4;
run;
```

Two Sample Discriminant Analysis

The DISCRIM Procedure

Total Sample Size	46	DF Total	45
Variables	4	DF Within Classes	44
Classes	2	DF Between Classes	1

Number of Observations Read	46
Number of Observations Used	46

Class Level Information

group	Variable Name	Frequency	Weight	Proportion	Prior Probability
0	0	21	21.0000	0.456522	0.500000
1	1	25	25.0000	0.543478	0.500000

Two Sample Discriminant Analysis

The DISCRIM Procedure
Within-Class Covariance Matrices

group = 0, DF = 20

Variable	x1	x2	x3	x4
x1	0.0441290476	0.0284764286	0.0344933333	0.0041473810
x2	0.0284764286	0.0210028571	0.0260200000	0.0034414286
x3	0.0344933333	0.0260200000	0.1643033333	0.0327816667
x4	0.0041473810	0.0034414286	0.0327816667	0.0445790476

group = 1, DF = 24

Variable	x1	x2	x3	x4
x1	0.047051000	0.008507167	0.074930500	-0.006703500
x2	0.008507167	0.002375667	0.008583167	0.000185333
x3	0.074930500	0.008583167	1.046774000	0.032632833
x4	-0.006703500	0.000185333	0.032632833	0.026381000

Two Sample Discriminant Analysis

The DISCRIM Procedure

Pooled Within-Class Covariance Matrix, DF = 44

Variable	x1	x2	x3	x4
x1	0.0457228398	0.0175841039	0.0565499697	-.0017712814
x2	0.0175841039	0.0108425714	0.0165090000	0.0016653766
x3	0.0565499697	0.0165090000	0.6456509697	0.0327004848
x4	-.0017712814	0.0016653766	0.0327004848	0.0346528398

Pooled Covariance Matrix Information

Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
4	-12.63389

Two Sample Discriminant Analysis

The DISCRIM Procedure

Generalized Squared Distance to group

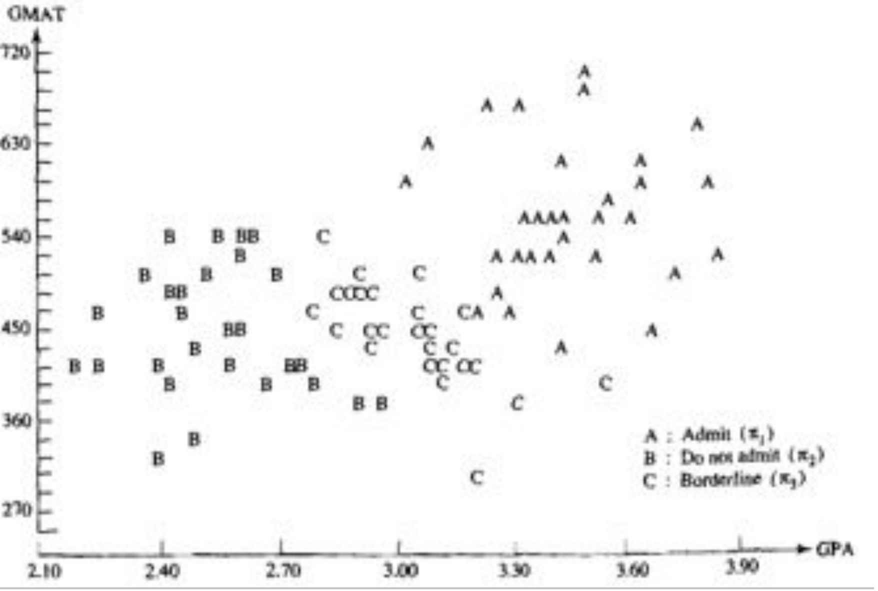
From group	0	1
0	0	3.62722
1	3.62722	0

TPM Example

$$n_1 = 31 \qquad n_2 = 28 \qquad n_3 = 26$$

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} 3.40 \\ 561.23 \end{bmatrix} \qquad \bar{\mathbf{x}}_2 = \begin{bmatrix} 2.48 \\ 447.07 \end{bmatrix} \qquad \bar{\mathbf{x}}_3 = \begin{bmatrix} 2.99 \\ 446.23 \end{bmatrix}$$

$$\bar{\mathbf{x}} = \begin{bmatrix} 2.97 \\ 488.45 \end{bmatrix} \qquad \mathbf{S}_{\text{pooled}} = \begin{bmatrix} .0361 & -2.0188 \\ -2.0188 & 3655.9011 \end{bmatrix}$$



Our goal with the TPM rule is to use the distance from the mean of each group in order to place our sample.

This data is uses GPA and GMAT to determine the admission status of individuals applying to graduate school.

TPM Example Continued

X0 is a sample student with minimizing at D3 we find the student right on the edge of admission.

With $\mathbf{x}_0' = [3.21, 497]$, the sample squared distances are

$$D_1^2(\mathbf{x}_0) = (\mathbf{x}_0 - \bar{\mathbf{x}}_1)'\mathbf{S}_{\text{pooled}}^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_1)$$

$$= [3.21 - 3.40, \quad 497 - 561.23]\begin{bmatrix} 28.6096 & .0158 \\ .0158 & .0003 \end{bmatrix}\begin{bmatrix} 3.21 - 3.40 \\ 497 - 561.23 \end{bmatrix}$$

$$= 2.58$$

$$D_2^2(\mathbf{x}_0) = (\mathbf{x}_0 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_2) = 17.10$$

$$D_3^2(\mathbf{x}_0) = (\mathbf{x}_0 - \bar{\mathbf{x}}_3)'\mathbf{S}_{\text{pooled}}^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_3) = 2.47$$

Logistic regression

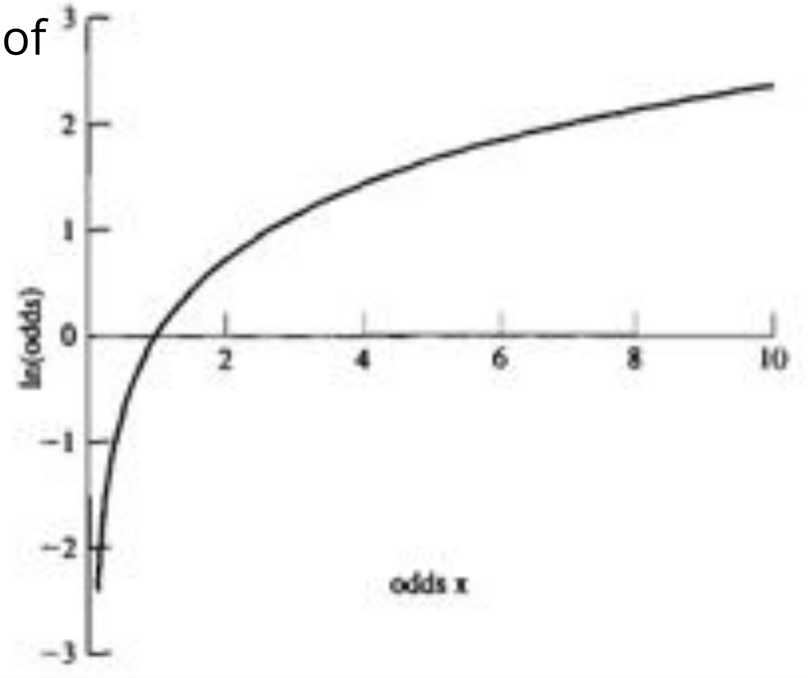
The Method

This is an approach where at least some of the variables are qualitative and in its simplest form works well with data labeled in a binary fashion.

An example of this would be Male or Female. Or in code 0 or 1.

The Model

Where normally we would just use a linear model in this case we will consider the ratio of the classes: odds = p or 1-p

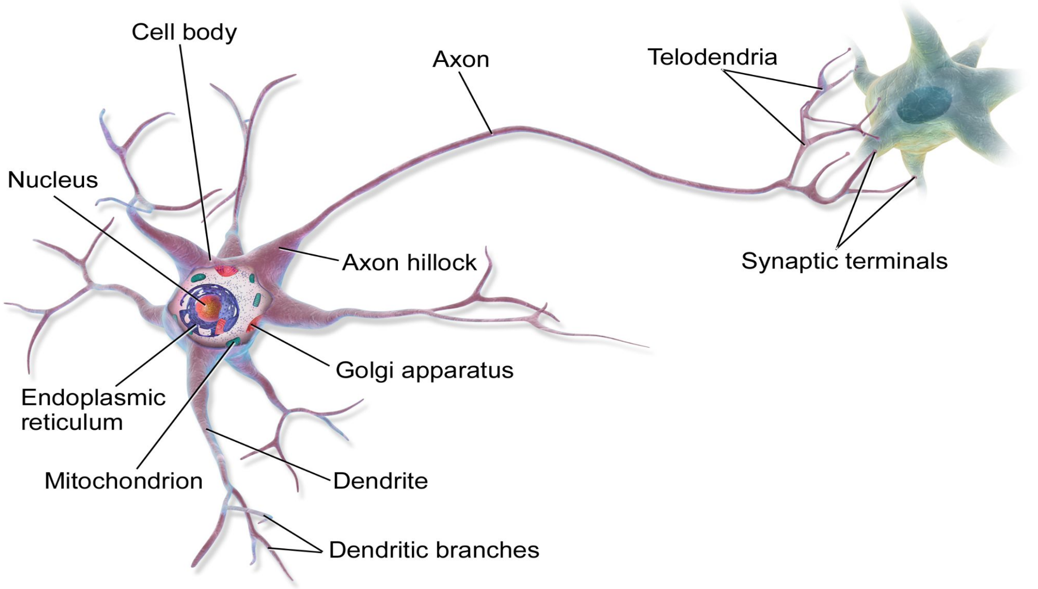


Deep Learning Networks

Neurons

A neuron is Simply a function applied to the data followed by an activation function.

A group of these functions applied to the data is called a layer.



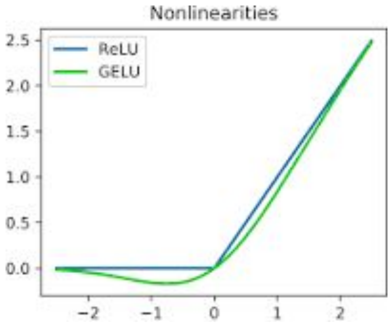
Linear Neuron

- Applies a linear transformation to the data.

$$y = x\mathbf{A}^{\mathsf{T}} + \boldsymbol{\beta}$$

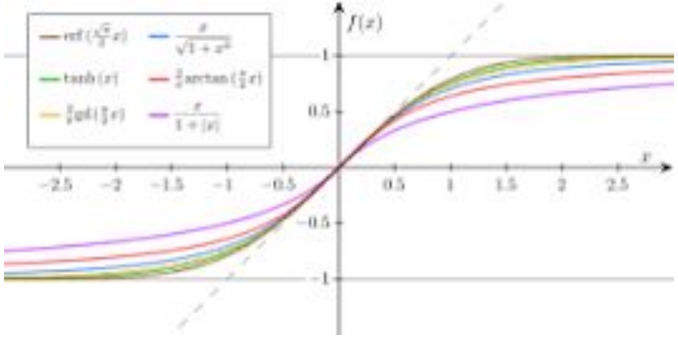
Activation Functions

- Relu applies a rectified linear unit function which takes the max between input value and 0.



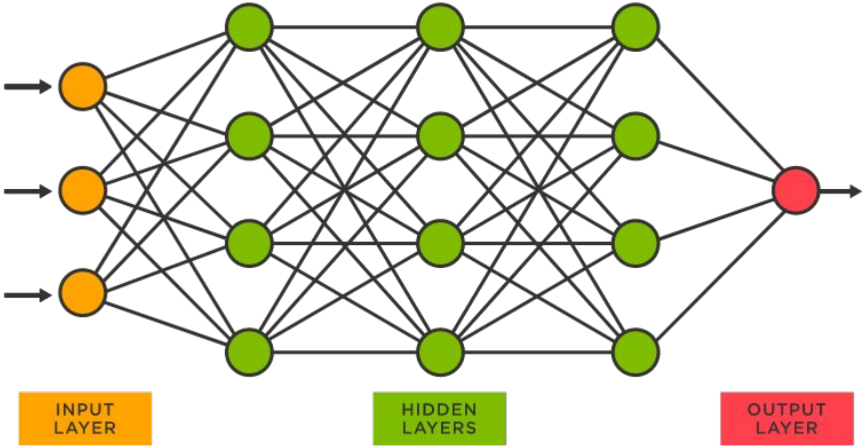
- Sigmoid Function applies:

$$S(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1} = 1 - S(-x).$$



Layers

Groups of neurons strong together to either apply a linear Function(Linear, BiLinear, etc.), Activation function(Sigmoid, ReLU), or Convolutional(We will talk about this later).



Loss functions and Optimizers

Loss Functions:

- MSE Loss:
- Cross Entropy Loss
- L1 Loss:

Optimizers:

- Adam
- Adagrad
- Stochastic Gradient Descent

The Basic Algorithm

Forward Pass: Sends the data forward through the layers in order to get a prediction.

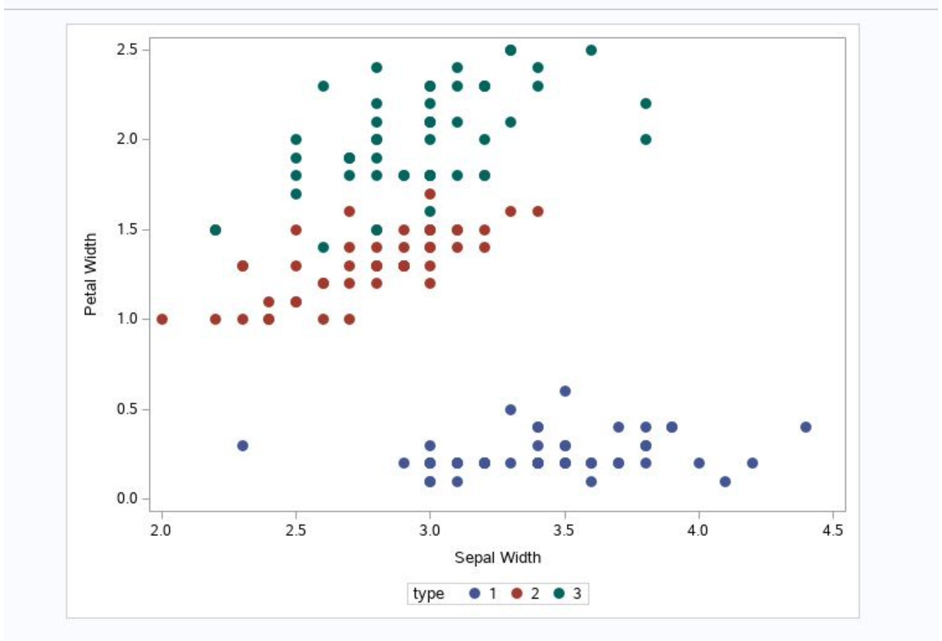
Backward Propagation: through the forward pass we calculate loss and then propagate it backwards through the network.

Repeat.

Logistics Regression and Neural Net on the Iris Dataset

The Data

THe Iris data set is a set of 150 observations of 3 different species of Iris plant, with Four Features(Sepal length and width, Petal length and Width). Using an 80/20 data split to construct a training and testing data set.



Logistic Regression Using SAS Demo

Results

Using the Iris dataset with the code shown to the left we created a model using the training data that was able to place each testing data in its proper place.

```
proc surveyselect data = Iris out=shuffled rate=1 outorder=random;

Data Train;
    set shuffled (obs = 120);

Data Test;
    set shuffled (firstobs = 121 obs=150);
run;

proc sgplot Data = Iris;
scatter x = sw y = pw / group= type markerattrs=(symbol = circlefilled size = 10);
xaxis label="Sepal Width";
yaxis label="Petal Width";
run;

proc logistic data=train outmodel=irislr;
model type = pw sw sl pl;
run;

proc logistic inmodel=irislr;
score data=Test(drop=type) out=score;
run;

data final;
    merge score Test;
    if I_type-type = 0 Then answer = 1;
    else answer = 0;
run;
```

The LOGISTIC Procedure

Model Information	
Data Set	WORK.TRAIN
Response Variable	type
Number of Response Levels	3
Model	cumulative logit
Optimization Technique	Fisher's scoring

Number of Observations Read	120
Number of Observations Used	120

Response Profile		
Ordered Value	type	Total Frequency
1	1	42
2	2	39
3	3	39

Probabilities modeled are cumulated over the lower Ordered Values.

Model Convergence Status
Complete separation of data points detected.

Score Test for the Proportional Odds Assumption		
Chi-Square	DF	Pr > ChiSq
0.7426	4	0.9460

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	267.518	13.621
SC	273.093	30.346
-2 Log L	263.518	1.621

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	261.8972	4	<.0001
Score	113.0090	4	<.0001
Wald	4.6113	4	0.3296

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1	-6.6312	24.9440	0.0707	0.7904
Intercept	2	1	14.9917	23.8217	0.3961	0.5291
pw		1	-14.1464	9.4814	2.2261	0.1357
sw		1	10.0042	9.0610	1.2190	0.2696
sl		1	6.0318	4.9454	1.4877	0.2226
pl		1	-12.0283	7.8707	2.3355	0.1265

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
pw	<0.001	<0.001	84.501
sw	>999.999	<0.001	>999.999
sl	416.481	0.026	>999.999
pl	<0.001	<0.001	29.904

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	100.0	Somers' D	1.000
Percent Discordant	0.0	Gamma	1.000
Percent Tied	0.0	Tau-a	0.672
Pairs	4797	c	1.000

Shallow Neural Network Using Python Demo

Results

Using a neural net I was able to create the same results using 4 linear layers and the relu activation function. Using those we successfully placed most of the data into the right class.

Computer Vision

Image Data: What an Image is made of.

Image data is high dimensional data with many features and is an interesting data type when it comes to classifications.

Feature Extraction(A New Type of Layer)

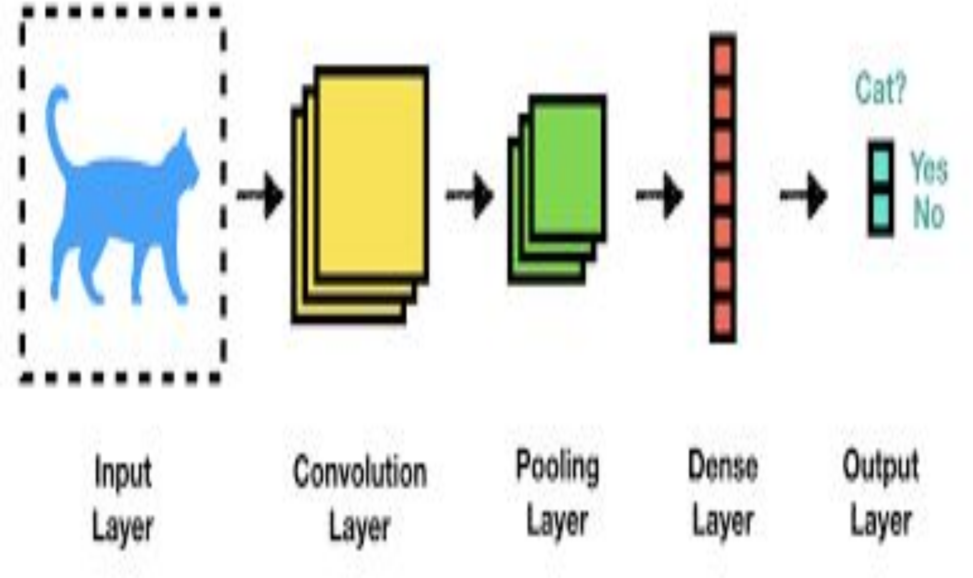
How do we teach a computer what is in an image. It all starts with methods of feature extraction. We have seen this before with PCA which can be done on images. However we have other Methods of extracting features.

Convolutional Neural Nets

Convolutional Layer: Breaks up our Image into smaller images and pulls out features.

Pooling Layer: Reduces Dimensions by combining outputs of CL.

Dense Layer: Is similar to our original NN which produces an answer to the question.



Difficulties of Computer Vision

Computer VIsion takes a lot of data in order to have good results.

It can also take a lot of time in order completely create a brand new model. That is why it is a very common practice to use another model to start and edit the last Layer. This is also known as transfer learning. Or applying a model pre created to complete a new task.

Any Questions?