Machine learning: prediction, classification and clustering

UBB Faculty of Sociology

Course Agenda

#1 Intro, Simple Linear Regression

#2 Python recap, Git, Handling data, EDA

#3 Regression, Decision Trees

#4 Bias, Variance, Overfitting, Classification

#5 Random Forest Classifier, Clustering

#6 Neural Networks

#7 Help Final Project

#8 Help Final Project

4 Bias, Variance, Overfitting, Classification

#4.1 Catch up
#4.2 Bias, variance tradeoff
#4.3 Overfitting
#4.4 Classification
#4.5 Classification Metrics
#4.6 Final Project Task 3 - Census Data Regression
#4.7 Questions & Further reading

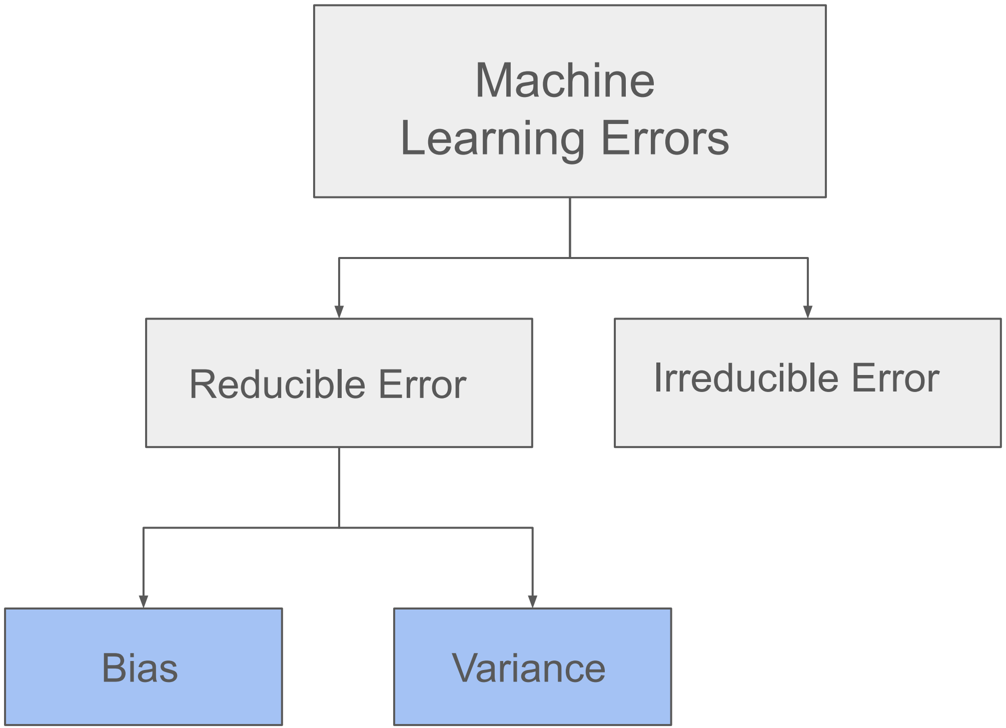
4.1 Catch up

Share one thing that stood out to you:

one thing you found surprising / interesting / useful etc.

4.2 Bias, variance tradeoff

2 types of errors



- **Bias** refers to the error introduced by approximating a real-world problem with a simplified model, leading to underfitting.
- **Variance** refers to the model's sensitivity to small changes in the training data, often resulting in overfitting.

Bias and Variance

Bias can be also inherent to our model: a classifier can be biased to a particular kind of solution (e.g. linear classifier).

HIGH bias and **LOW** variance ---> underfitting

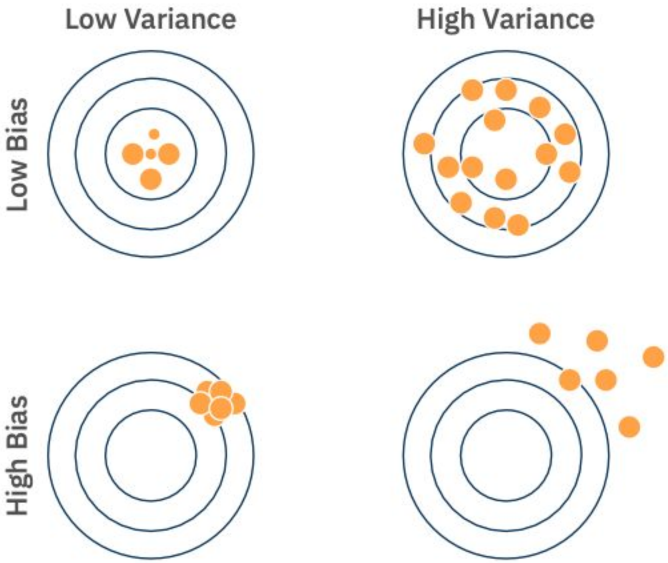
- the model is unable to capture the underlying pattern of the data
- train error > test error
- is not robust enough to produce accurate predictions
- use more complex model
- Boosting

LOW bias and **HIGH** variance ---> overfitting

- train error << test error
- add more data, reduce model complexity
- Bagging

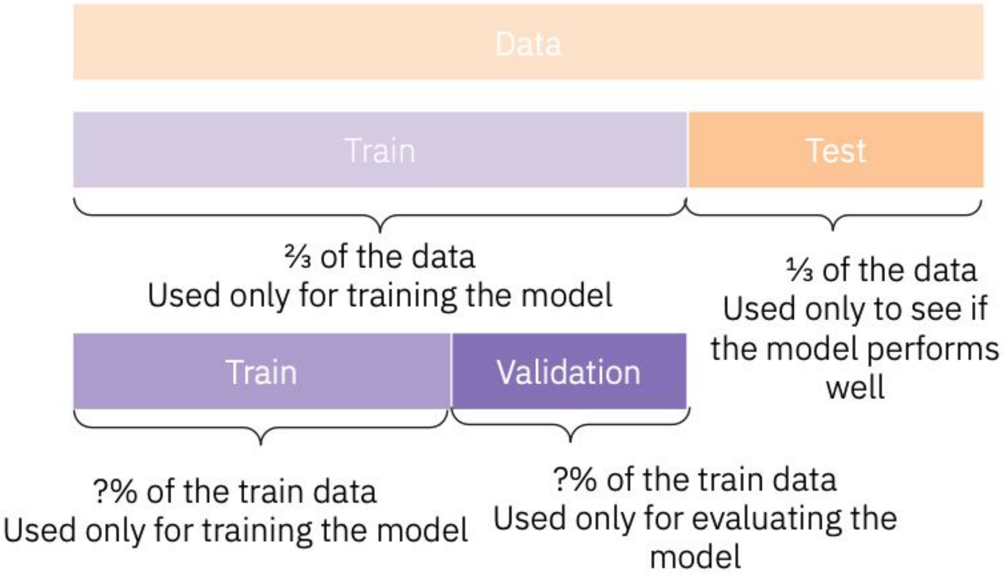
Bias, variance tradeoff

LOW bias and **LOW** variance will give us a balanced mode



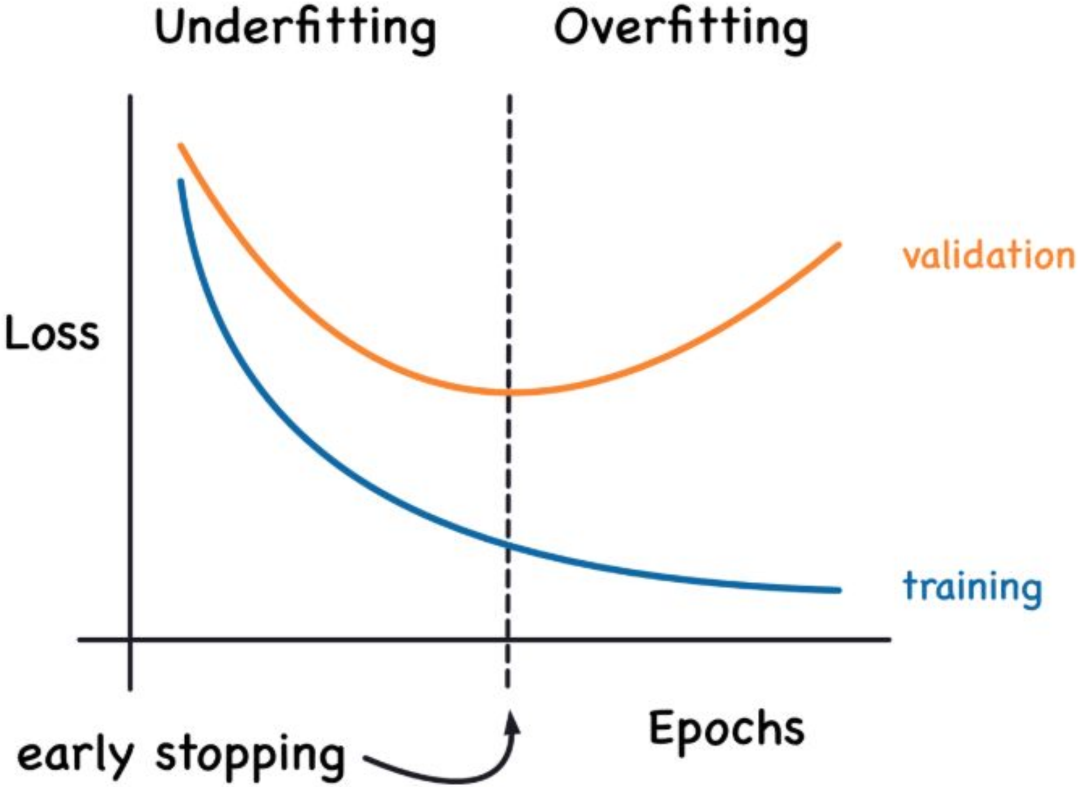
4.3 Overfitting

Validation set

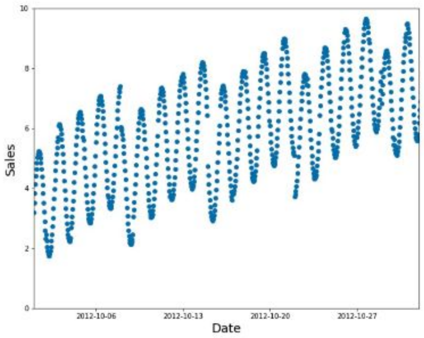


¾ - ¼ or ⅔ -⅓ are good choices

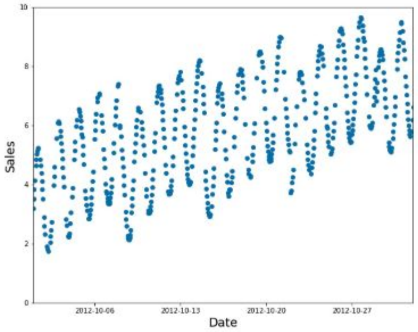
- **Train Set:** Used to train the model by adjusting its parameters to minimize error.
- **Validation Set:** Used to tune hyperparameters and evaluate the model's performance during training to avoid overfitting.
- **Test Set:** Used for the final evaluation of the model's performance on unseen data.


Underfitting
Overfitting
Loss
validation
training
Epochs
early stopping

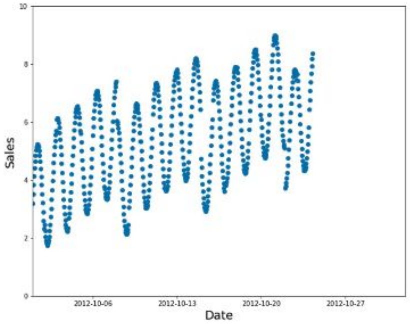
Time series data



Time series data



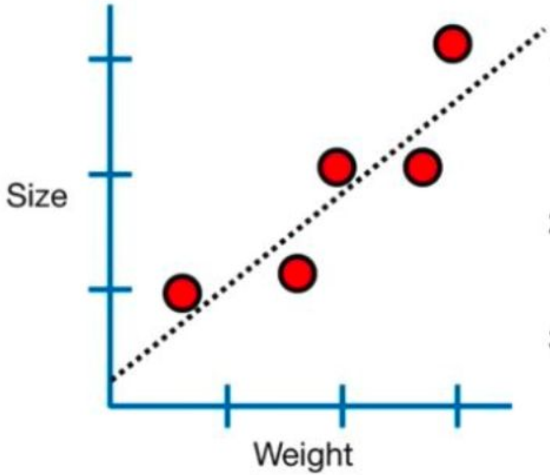
a poor choice for your training set



a better choice for your training set

4.4 Classification

Linear Regression - statistics



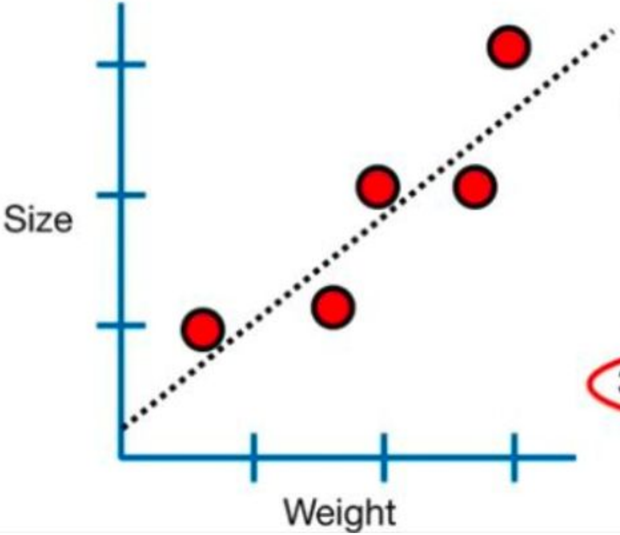
...and with that line, we could do a lot of things:

1) Calculate R^2 and determine if ***weight*** and ***size*** are correlated. Large values imply a large effect.

2) Calculate a p-value to determine if the R^2 value is statistically significant.

3) Use the line to predict ***size*** given ***weight***.

Linear Regression - machine learning



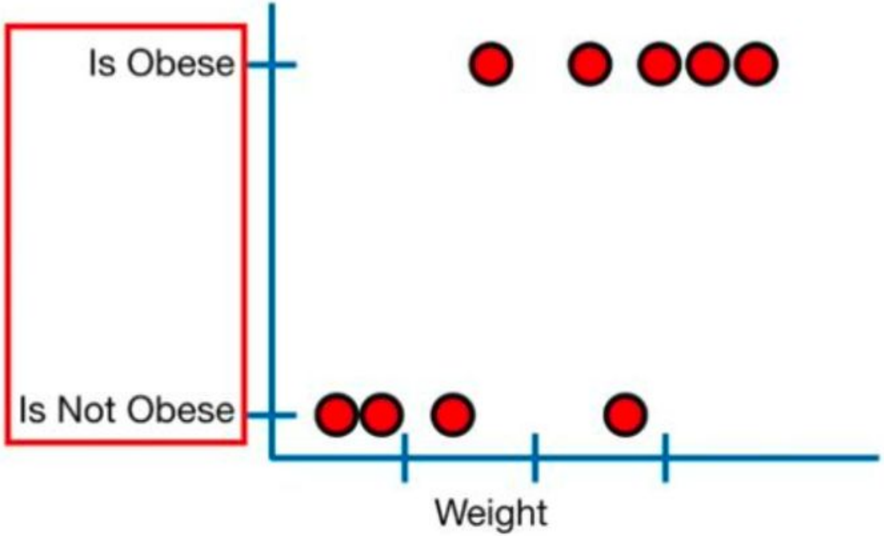
Although we didn't mention it at the time, using data to predict something falls under the category of "machine learning".

So plain old linear regression is a form of machine learning.

3) Use the line to predict ***size*** given ***weight***.

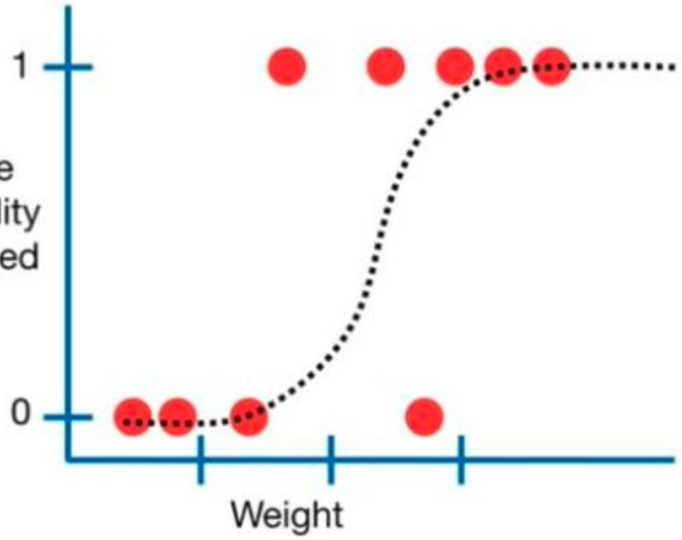
Logistic Regression

Logistic regression predicts whether something is ***True*** or ***False***, instead of predicting something continuous like ***size***.



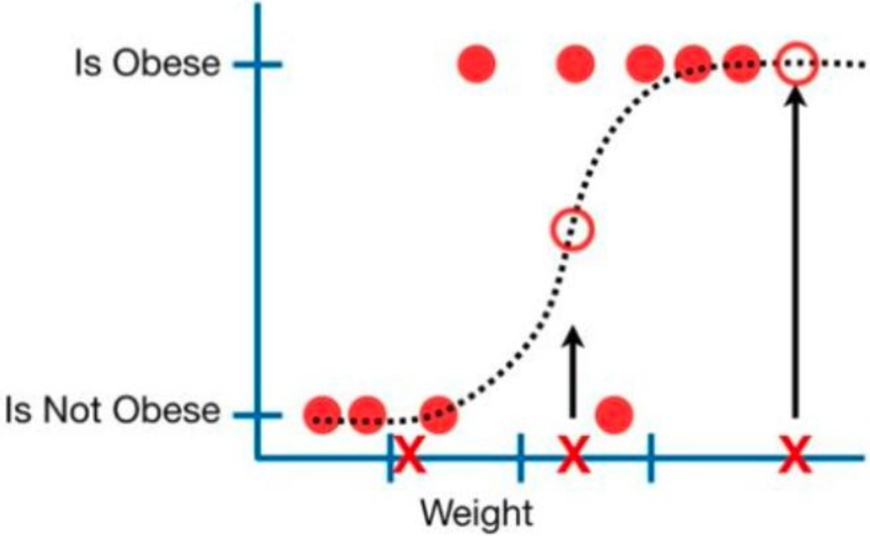
Logistic Regression

…and that means that the curve tells you the probability that a mouse is ***obese*** based on its ***weight***.

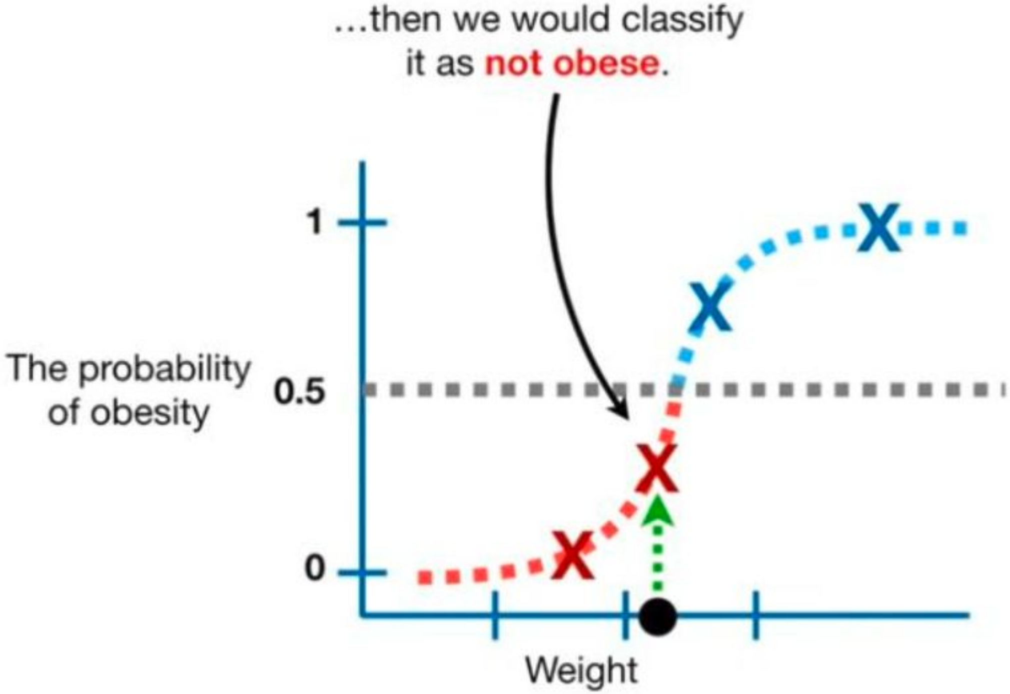


Logistic Regression

For example, if the probability a mouse is obese is > 50%, then we'll classify it as obese, otherwise we'll classify it as "not obese".



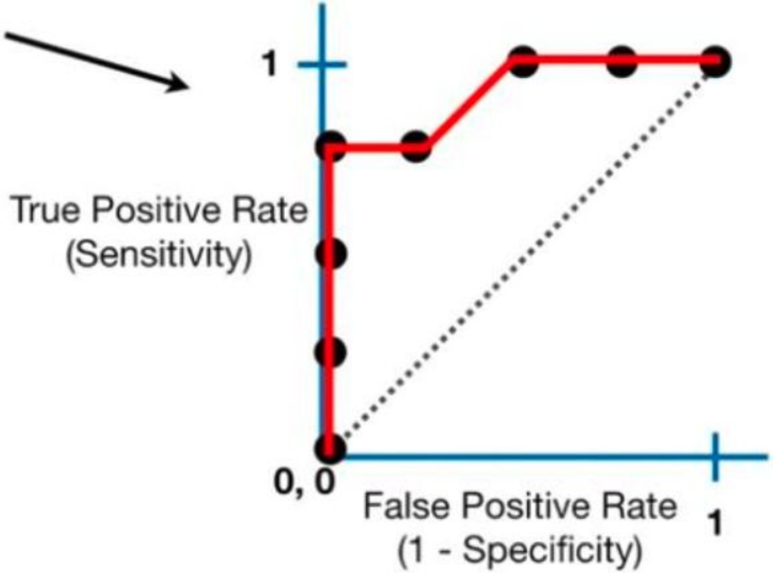
ROC curve



...then we would classify
it as not obese.
1
The probability
of obesity
0.5
0
Weight

ROC curve

So instead of being overwhelmed
with confusion matrices,
Receiver Operator Characteristic
(**ROC**) graphs
provide a simple way to summarize
all of the information.



4.5 Classification Metrics

Misclassification
Confusion Matrix

Perfect classification is very rare.
We use the Confusion Matrix to evaluate our classification results.
In general there are 4 values in the 2 class case:

true positive (TP) = Predicted True & Actual True

true negative (TN) = Predicted False & Actual False

false positive (FP)= Predicted True & Actual False

false negative (FN)= Predicted False & Actual True

n=165	**Predicted: NO**	**Predicted: YES**
Actual: NO	TN	FP
Actual: YES	FN	TP

However, there are other important statistics.

Important statistics

Accuracy: Overall, how often is the classifier correct?

(TP+TN) / total = (100+50)/165 = 0.91

Precision: When it predicts yes, how often is it correct?

TP / (TP+FP) = 100/(100+10) = 0.91

Recall/Sensitivity/TPR: How good is it at positive?

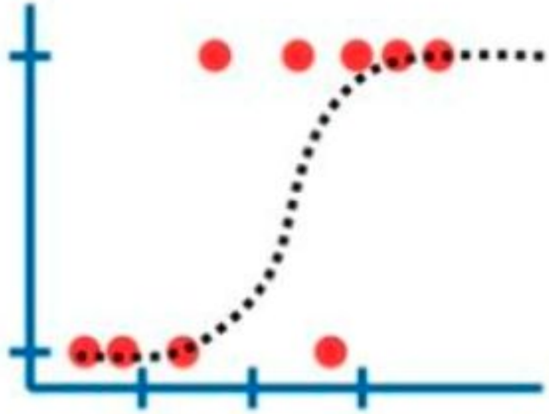
TP / (TP+FN) = 100/(100+5) = 0.95

Specificity: How good is it at negative?

TN / (TN+FP) = 50/(50+10) = 0.83

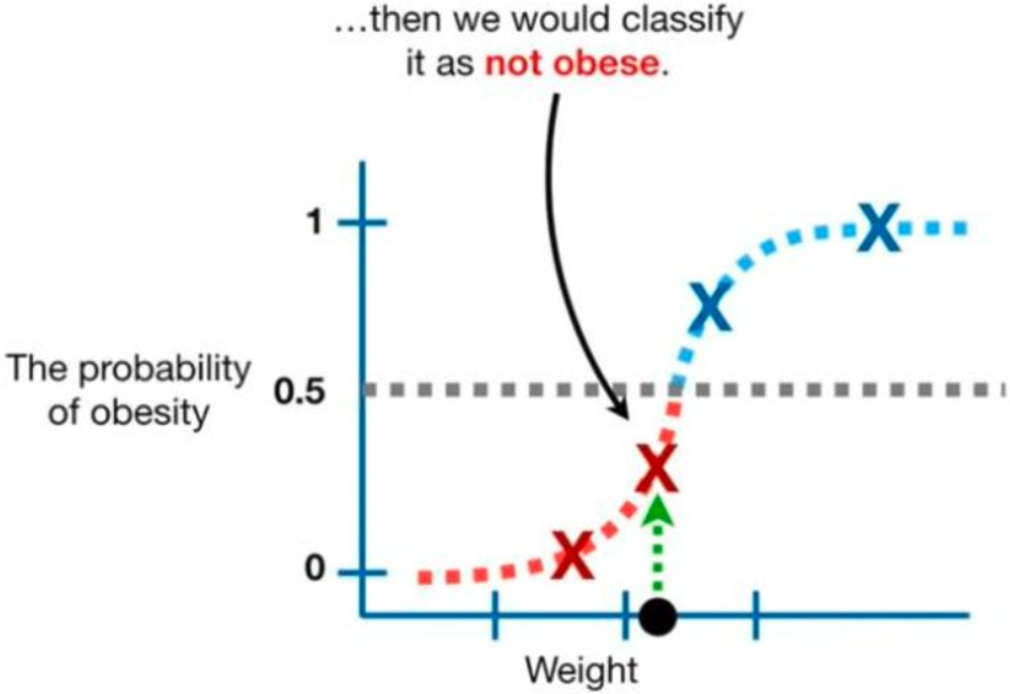
False Positive Rate

FPR = FP / (TN + FP) = 10/(50+10) = 0.6



n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

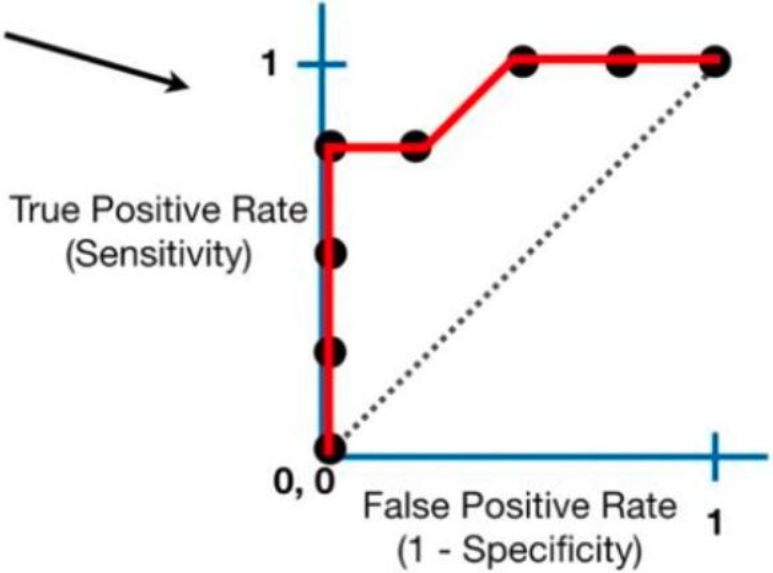
ROC curve

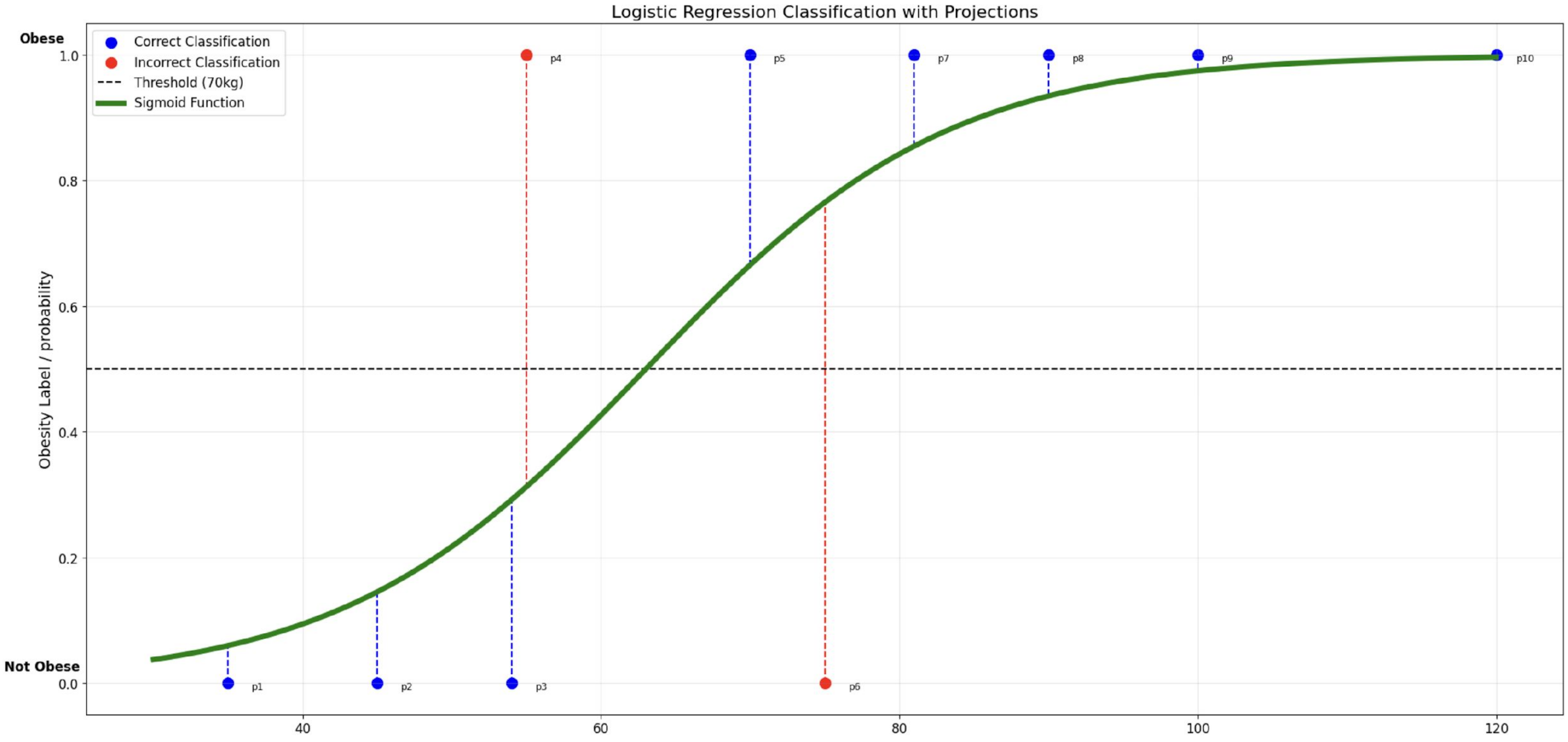


...then we would classify
it as not obese.
1
0.5
0
The probability
of obesity
Weight

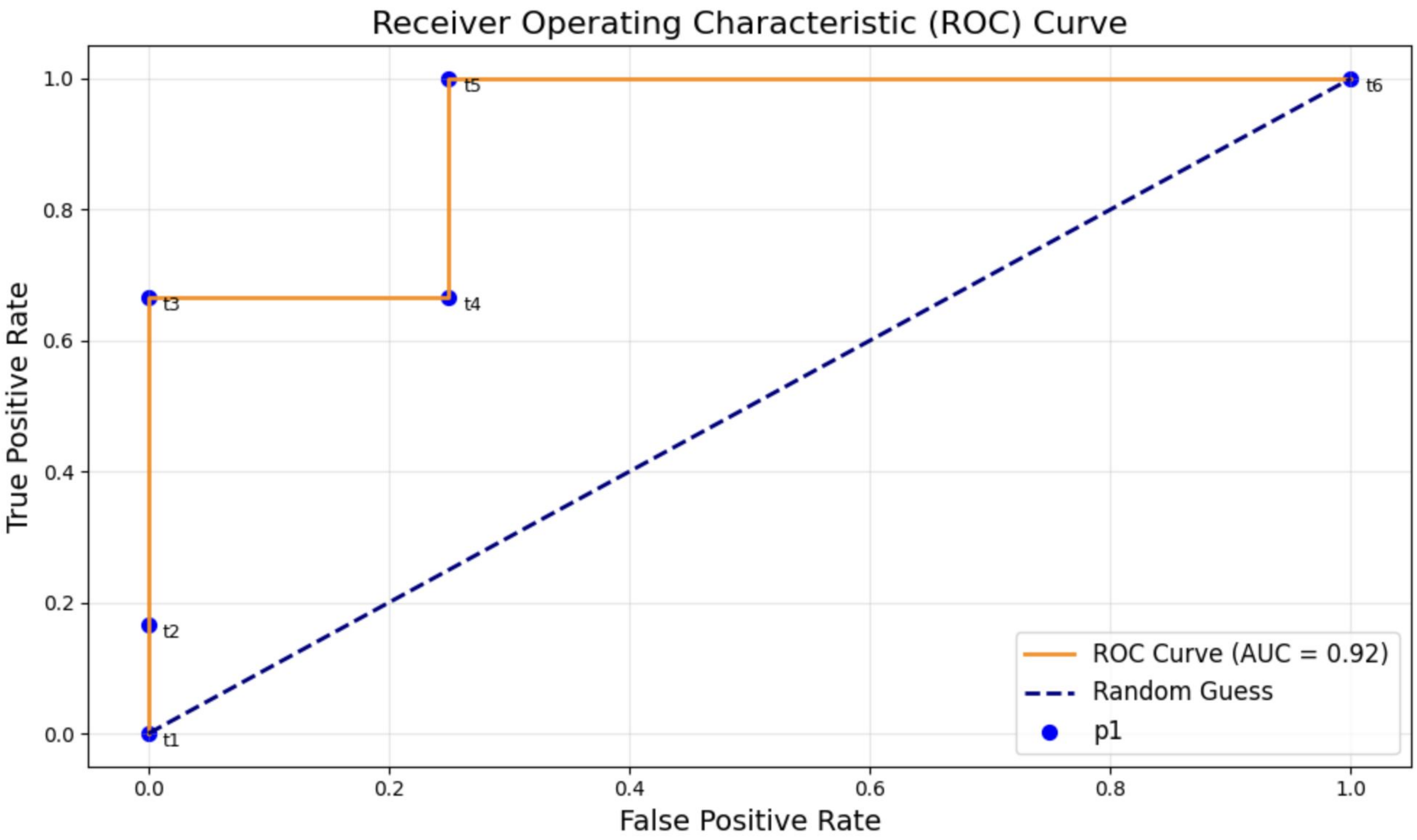
ROC curve

So instead of being overwhelmed
with confusion matrices,
Receiver Operator Characteristic
(**ROC**) graphs
provide a simple way to summarize
all of the information.





[Code on colab]



[Code on colab]

4.7 Further Reading & Questions

#1 Bias versus variance: https://inria.github.io/scikit-learn-mooc/overfit/bias_vs_variance_slides.html

#2 Bias variance decomposition: https://www.youtube.com/watch?v=zUJbRO0Wavo

#3 How (and why) to create a good validation set: https://www.fast.ai/posts/2017-11-13-validation-sets.html

#4 Problems with metrics: https://www.fast.ai/posts/2019-09-24-metrics.html

#5 Evaluating models: https://alan-turing-institute.github.io/rds-course/modules/m4/4.4-ModelEvaluation.html

#6 Sensitivity and Specificity for more than two classes: https://www.youtube.com/watch?v=vP06aMoz4v

#7 Should Computers Run the World? - with Hannah Fry: https://www.youtube.com/watch?v=Rzhpf1Ai7Z

#8 Build a classification tree using Gini impurity: https://www.youtube.com/watch?v=_L39rN6gz7Y&t=196s

#9 Adaptive Boosting (AdaBoost) on Penguin data: https://github.com/INRIA/scikit-learn-mooc/blob/main/notebooks/ensemble_adaboost.ipynb

#10 ROC and AUC: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc#:~:text=The%20points%20on%20a%20ROC,to%20the%20specific%20use%20case.

#11 What is the Classification Threshold in Machine Learning?: https://www.iguazio.com/glossary/classification-threshold/

Thank you !!

Machine Learning Engineer / Data Scientist
zahariesergiu@gmail.com
https://www.linkedin.com/in/zahariesergiu/
https://github.com/zahariesergiu/ubb-sociology-ml