How does Data Science fit into Consulting?

Pharmaceutical Client
Optimizing how drugs are brought to market

Business Goals
- Improve clinical operations - clinical trial productivity by more than 10%
- Better monitor site level risks

Data Science Goals
- Country footprint optimization
- site selection and risk modeling
- trial management and forecasting

https://www.mckinsey.com/business-functions/mckinsey-analytics/how-we-help-clients

Agenda

- Types of Machine Learning
- Classification Regression Problems
- Train - Test Split
- Preprocessing
- Logistic Regression
- Loss Function
- Model Evaluation
- Assignment
- Further Reading

Types of Machine Learning

- **Supervised Learning**

$$f(x_i) \approx y_i$$

Examples: price prediction, spam detection, medical diagnosis, ad click prediction

- **Unsupervised Learning**

$$x_i \propto p(x)$$

Examples: Clustering, Dimensionality reduction

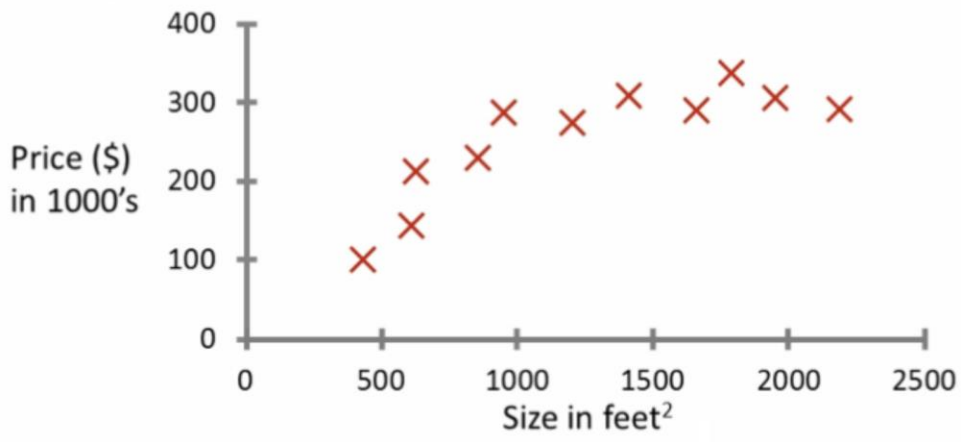
Classification and Regression Problems

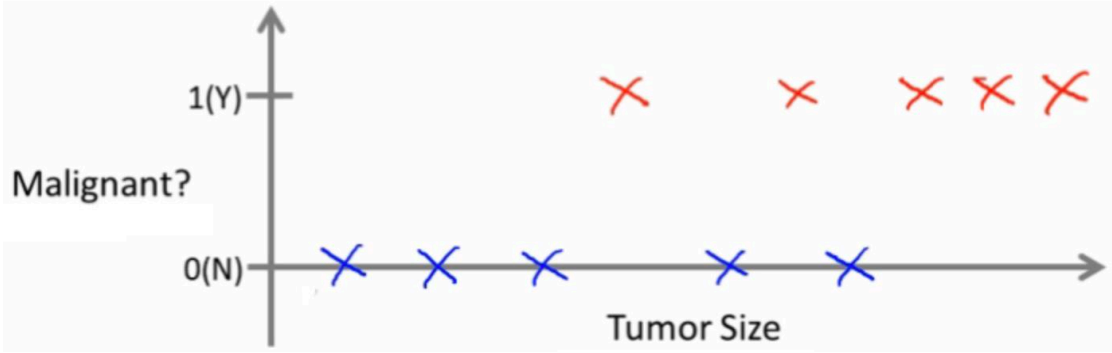
Regression

- Target y is continuous
- Example: House price prediction

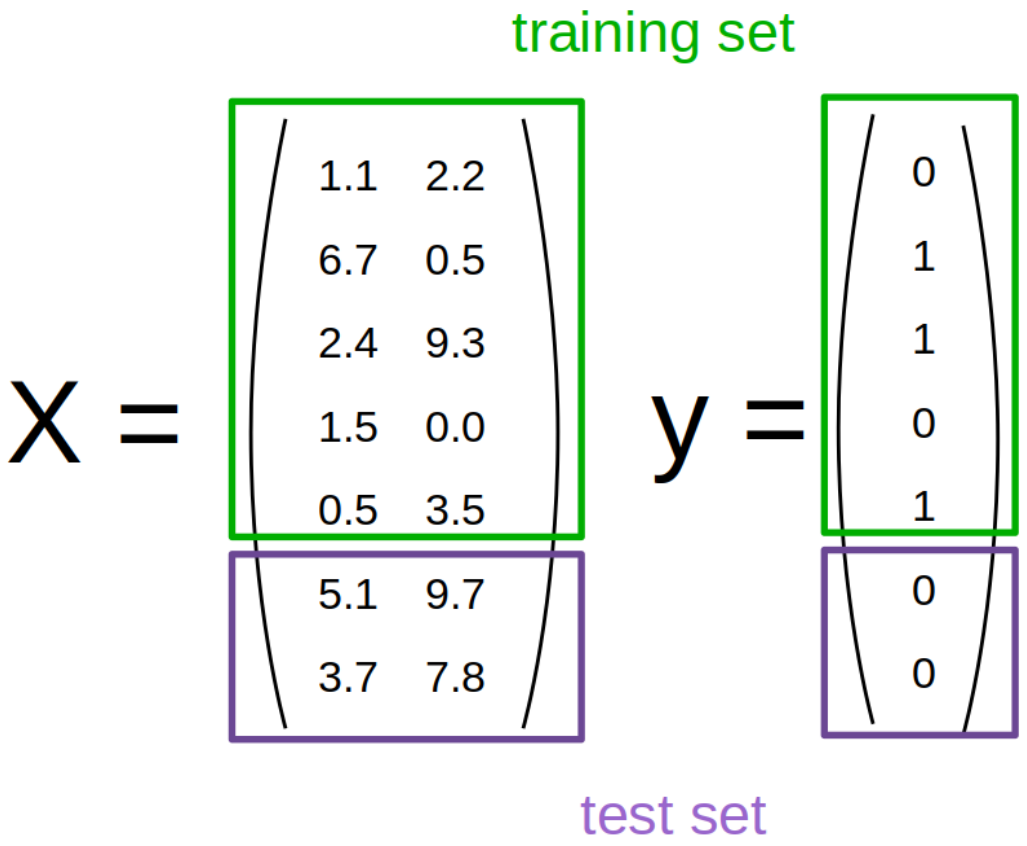
Classification

- Target y is discrete
- Example: Whether a tumor is malignant or benign based on tumour size





Train-Test Split



https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

Logistic Regression

- **Example:** "Gender Recognition by Voice"
- **Output:** 0 – Male and 1 –Female
- **Type of Problem:** Binary Classification Problem
- **What's the most basic thing you can do?:** Weighted sum

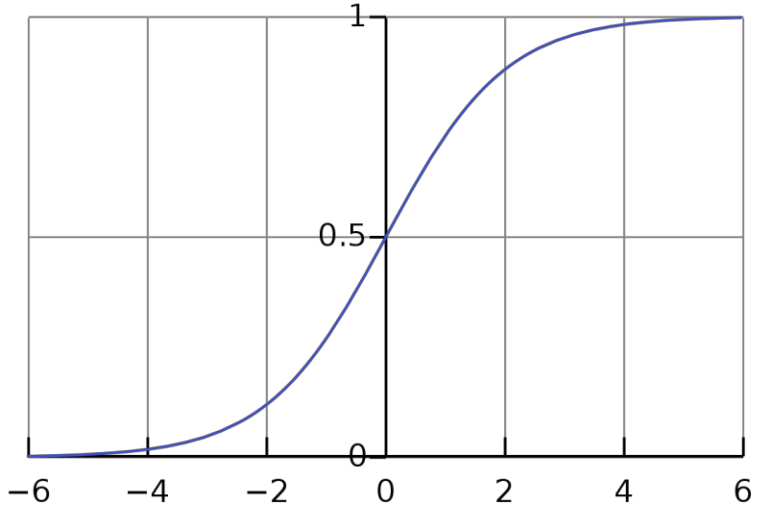
$$h_\theta(x) = \theta_0 x_0 + \theta_1 x_1 + \ldots + \theta_j x_j \quad = \theta^T x \quad = \begin{bmatrix} \theta_0 & \theta_1 & \ldots & \theta_j \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ \ldots \\ x_j \end{bmatrix}$$

$$Where\ x_0 = 1$$

THE UNIVERSITY OF CHICAGO

Logistic Regression

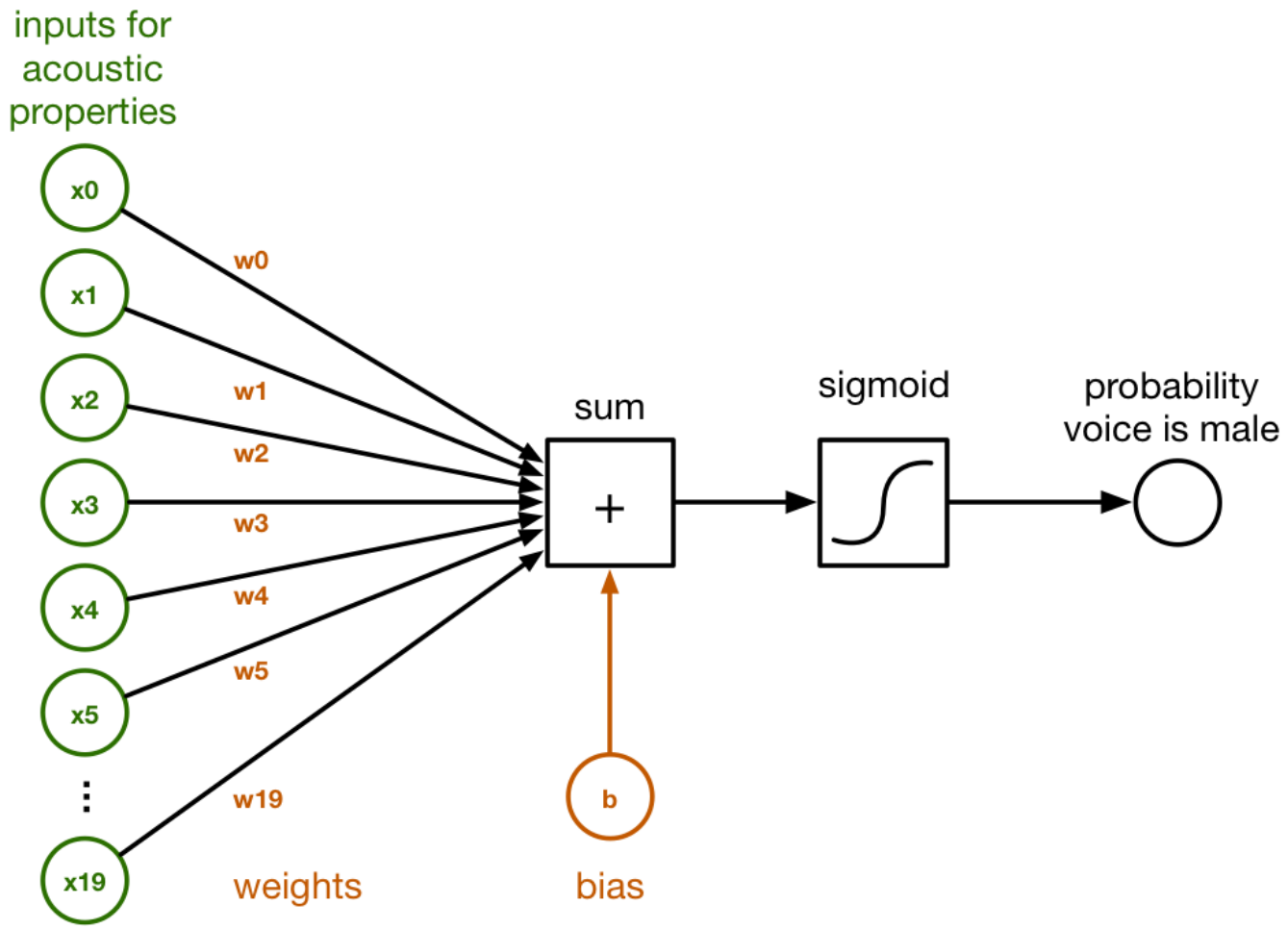
- **Next step:** *"What function would help our classifier to output values between 0 and 1"*
- **Sigmoid Function:**



- **What does y axis represent?:** outputs a number, we treat that value as the estimated probability that y=1. If p>0.5 we say it's 1 else we say it's a 0.

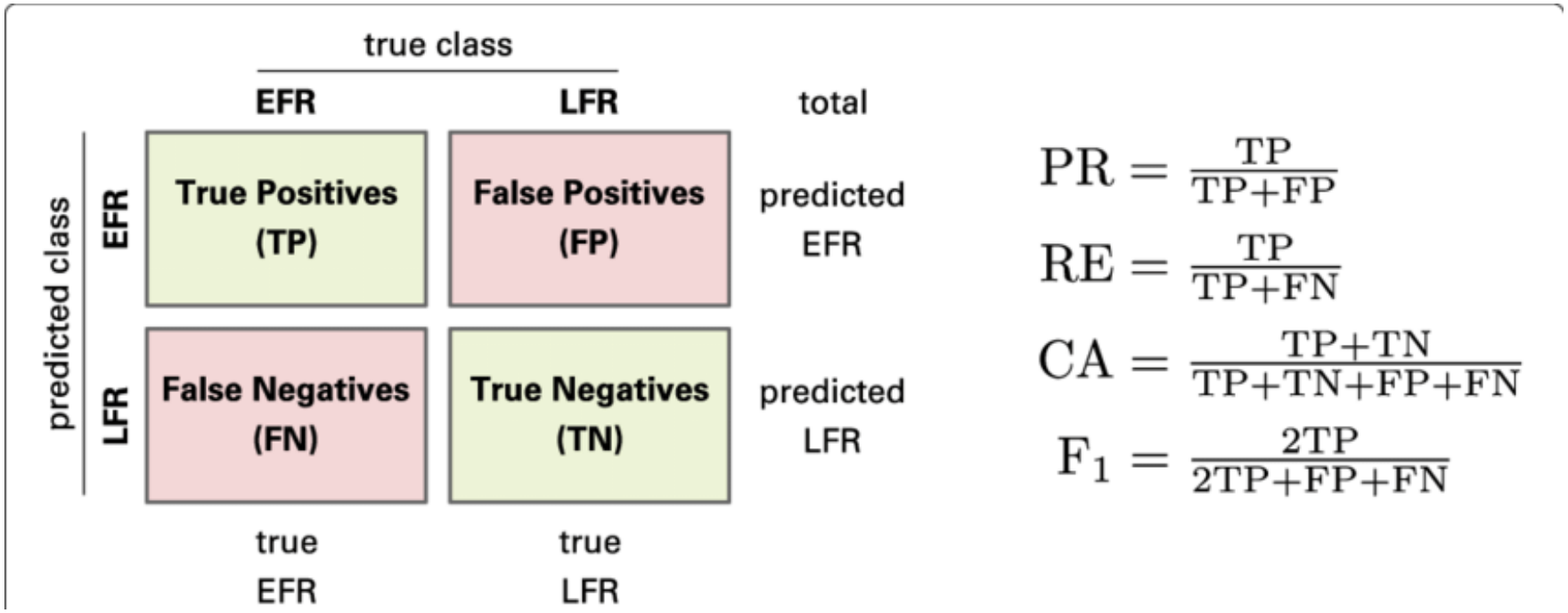
https://en.wikipedia.org/wiki/Sigmoid_function

Logistic Regression



http://machinethink.net/blog/tensorflow-on-ios/

Model Evaluation – Confusion Matrix



$$\mathrm{PR} = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$$

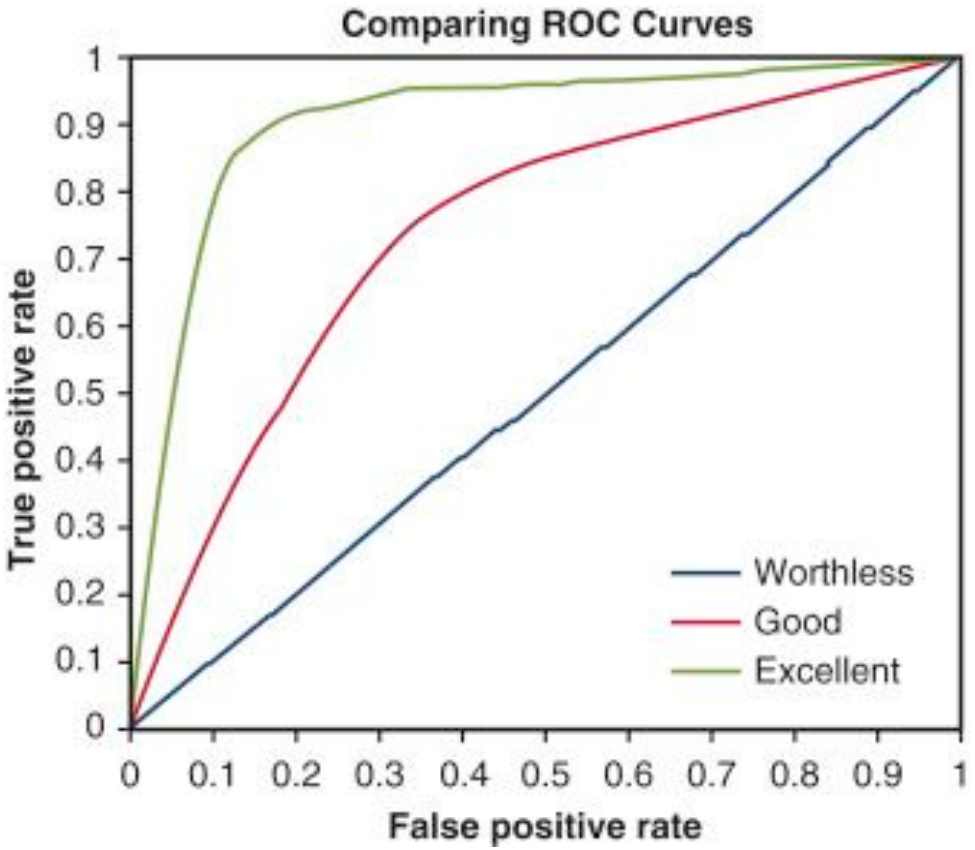
$$\mathrm{RE} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$$

$$\mathrm{CA} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}}$$

$$\mathrm{F_1} = \frac{2\mathrm{TP}}{2\mathrm{TP+FP+FN}}$$

https://www.researchgate.net/figure/Confusion-matrix-Exemplified-CM-with-the-formulas-of-precision-PR-recall-RE_fig1_330174519

Model Evaluation – ROC and AOC



THE UNIVERSITY OF CHICAGO

Loss Function

- **Next Step:** How to find thetas?
- **Intuition:** *"This is the cost you want the learning algorithm to pay if the outcome is hϑ(x) and the actual outcome is y"*
- **Binary Cross Entropy**

$$BCE = -\frac{1}{N}\sum_{i=0}^{N} y_i \cdot log(\hat{y}_i) + (1 - y_i) \cdot log(1 - \hat{y}_i)$$

- Yi is the label and ycapi is the probability obtained from the model
- This is a non-convex function i.e., there are many local optimum.

Assignment - Fitness Club

Independent variable
Default/Churn (Attrition Rate)

Dependent variable
Enrollment Date
Price
Downpayment
Months Due
Payment Type
Use
Age
Gender

Assignment - Fitness Club

Q1: Calculate basic stats descriptive statistics (mean, median, min, max, standard deviation) for each field

- Write a function, Use Numpy, Use Pandas - describe

Q2: Visualize distributions of data elements using histograms for key variables and predict which variables you expect to be most correlated with default/churn.

- Matplotlib and Seaborn - distplot, pairplot

Q3: Calculate Linear and/or Logistic Regression Models to Predict Churn/Retention, you may choose to identify groups within the data to narrow your focus on.

- Perform Train test split of 70 – 30 % using train test split function in scikit learn
- Scale the data and create one hot encoded variables if required
- Use logistic regression from scikit learn
- Fit the model and predict the values using .fit and .predict

Assignment - Fitness Club

- Display the output visually using charts of your choosing and explain your choice. (ROC Curve, Confusion Matrix, Gains Table)
  - Use confusion_matrix and classification_report from scikit learn
  - Plot ROC Curve - https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html

- In addition to the spreadsheet/code or programming output you submit, include a separate written document of 250-500 words that summarizes your process.

- Discuss why certain variables you expected to be significant are/are not and any other unexpected insights.

Assignment - Fitness Club

- Eliminate records with Age < 16 and Age =99 (reason age noisy due to customer not giving up age, and records incomplete where age is 99)
- Either create an indicator variable or eliminate records where Price = Down payment (take paid in full records out of analysis for model build)
- Only use records with date <=1998 (eliminate or set aside 1999 records since they dropped a payment type in that last year)
- Feature possible creation (% down payment – I think this would be a cool added variable but not sure if it is helpful with some modeling techniques)

Further Reading

- Code for Logistic Regression
  - https://chrisalbon.com/machine_learning/model_evaluation/plot_the_receiving_operating_characteristic_curve/
  - https://medium.com/@kgpvijaybg/logistic-regression-on-iris-dataset-48b2ecdfb6d3
  - https://www.kaggle.com/mnassrib/titanic-logistic-regression-with-python
  - https://medium.com/jovianml/predicting-survival-of-titanic-passengers-using-logistic-regression-model-14b9559dc4b5