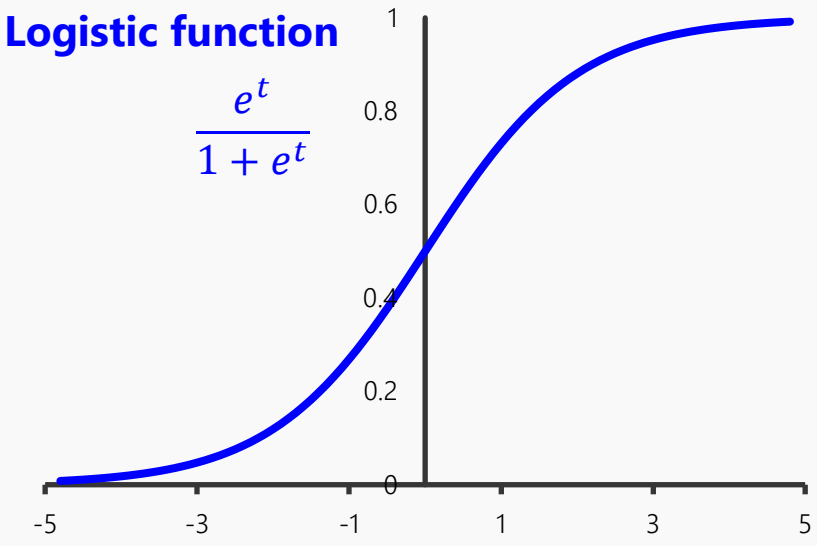
maximum likelihood a/b perspective

An expansion on the intuition behind **Logistic Regression**:



The **Logistic Function** models itself through a visible growth and saturation as seen in the illustration.

The function was developed in the mid-19th century by Adolphe Quetelet and his pupil, Pierre Francois Verhulst. The mathematicians were modeling the growth of populations with the intuition when countries become full, the population growth levels off and the population will saturate.

As a property known of probabilities, there can be no value above 1 or less than 0. The function $\frac{e^t}{1+e^t}$ therefore returns a probability measuring the inputs (t) to the model.

Logistic Regression enters when the probability of modeling the function will result as 1 or 0:

Predicting the value of 1: $P(Y_i = 1|x_i, \beta) = \frac{e^{\sum_{j=1}^{p} \beta_j x_{ij}}}{1+e^{\sum_{j=1}^{p} \beta_j x_{ij}}}$

Written in matrix notation: $P(Y_i = 1|x_i, \beta) = \frac{e^{x_i\beta}}{1+e^{x_i\beta}}$

Predicting the value of 1: $P(Y_i = -1|x_i, \beta) = 1 - \frac{e^{x_i\beta}}{1+e^{x_i\beta}}$

Written in simplified notation: $P(Y_i = -1|x_i, \beta) = \frac{1}{1+e^{x_i\beta}}$

The Likelihood of the observations will be calculated:

$$\text{Likelihood}(x_i, y_i) = P(Y_i = y_i|x_i, \beta)$$

Therefore:

$$\text{If} \begin{cases} P(Y_i = -1|x_i, \beta) = 1 - \frac{e^{x_i\beta}}{1+e^{x_i\beta}} = \frac{1}{1+e^{x_i\beta}} = \frac{1}{1+e^{-y_i x_i\beta}} \\ P(Y_i = 1|x_i, \beta) = \frac{e^{x_i\beta}}{1+e^{x_i\beta}} = \frac{1}{1+e^{-x_i\beta}} = \frac{1}{1+e^{-y_i x_i\beta}} \end{cases}$$

$$\text{Likelihood}(x_i, y_i) = P(Y_i = y_i|x_i, \beta)$$

$$P(Y_i = y_i|x_i, \beta) = \frac{1}{1+e^{-y_i x_i\beta}}$$

Finally adding the product property to **Logistic Regression** as follows:

$$\prod_{i=1}^{n} \text{Likelihood}(x_i, y_i) = \prod_{i=1}^{n} P(Y_i = y_i | x_i, \beta)$$

$$\prod_{i=1}^{n} P(Y_i = y_i | x_i, \beta) = \prod_{i=1}^{n} \frac{1}{1 + e^{-y_i x_i \beta}}$$

Ultimately summarized as:

$$\prod_{i=1}^{n} \text{Likelihood}(x_i, y_i) = \prod_{i=1}^{n} \frac{1}{1 + e^{-y_i x_i \beta}}$$

Proceeding to take the logarithm of each side and simplifying appropriately:

$$-\log \prod_{i=1}^{n} \text{Likelihood}(x_i, y_i) = -\log \prod_{i=1}^{n} \frac{1}{1 + e^{-y_i x_i \beta}}$$

$$= \sum_{i=1}^{n} -\log \frac{1}{1 + e^{-y_i x_i \beta}}$$

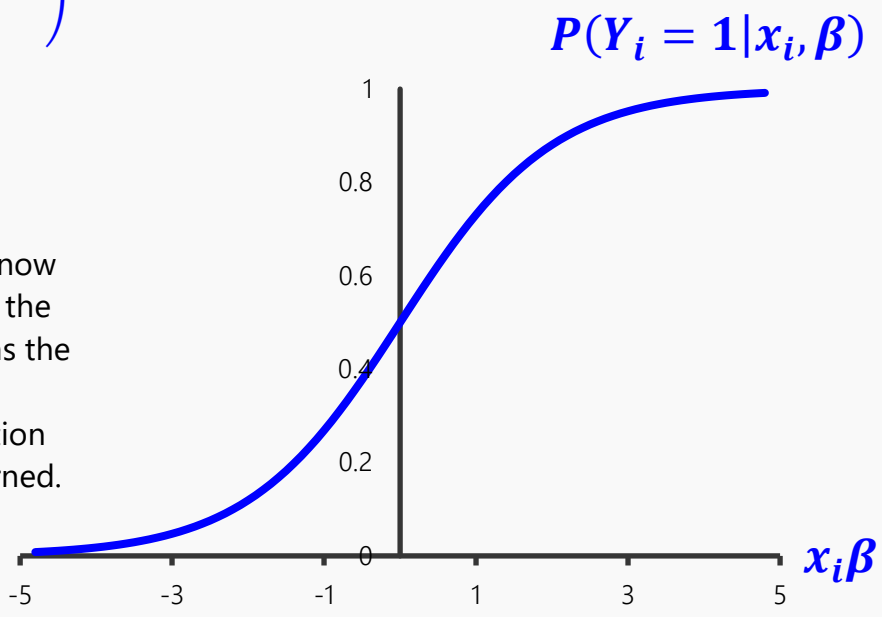
$$= \sum_{i=1}^{n} \log\left(1 + e^{-y_i x_i \beta}\right)$$

The above derivation ultimately returns us to the original **Logistic Function**:

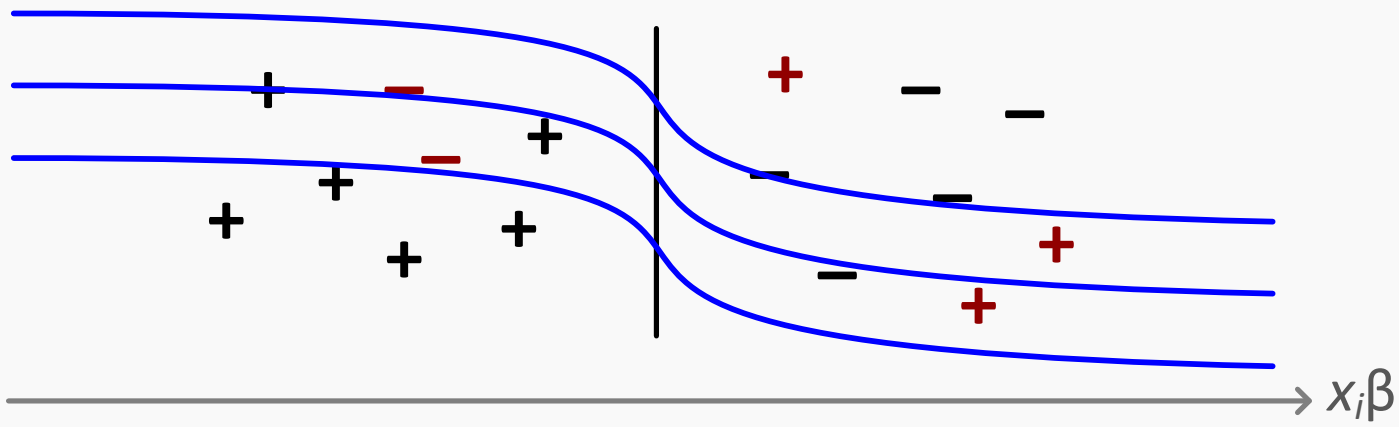
$$\min_{\beta_1, \beta_2, \ldots, \beta_p} \frac{1}{n} \sum_{i=1}^{n} \left(1 + e^{-y \sum_{j=1}^{p} \beta_j x_{ij}}\right)$$

$$P(Y_i = 1 | x_i, \beta) = \frac{e^{x_i \beta}}{1 + e^{x_i \beta}}$$

The derivation of the **Logistic Function** now provides a probabilistic interpretation of the model. Whatever score the model assigns the observation, the model also assigns the probability (likelihood). Both a Classification and Probability of Classification are returned.



The above in **Logistic Function** in geometric notation:



The geometric illustration above demonstrates a high probability received in a prediction of $y = 1$ on the left side of the decision boundary. Conversely the probability logarithmically decreases as the values move towards and over the decision boundary into the right side of the illustration.

Logistic Regression in Summary:

1. Randomly partition the data into training and test sets

2. Estimate the coefficients and train the model:

$$\hat{\beta} = \underset{\beta_1,\beta_2,\dots,\beta_p}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\left(1 + e^{-y\sum_{j=1}^{p}\beta_j x_{ij}}\right)$$

3. Score the model: Compute scores for each x_i in the test set

$$f(x_i) = \sum_{j=1}^{p}\beta_j x_{ij}$$

4. Evaluate the model's performance

5. Improving the performance through **Regularization** (discussed later)

$$f^* = \underset{\text{models } f}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\ell\big(y_i f(x_i)\big) + C + \text{Regulatization}(f)$$

$$f(x_i) = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_p x_{ip}$$

$$\text{Regularization}(f) = \beta_1^2 + \beta_2^2 + \beta_3^2 + \cdots + \beta_p^2 = \|\beta\|_2^2 \text{ (referred to as } \ell_2)$$

$$\text{Regularization}(f) = |\beta_1^2| + |\beta_2^2| + |\beta_3^2| + \cdots + |\beta_p^2| = \|\beta\|_1 \text{ (referred to as } \ell_1)$$