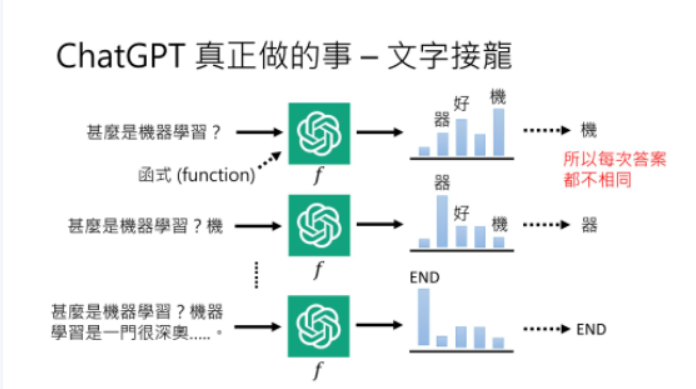
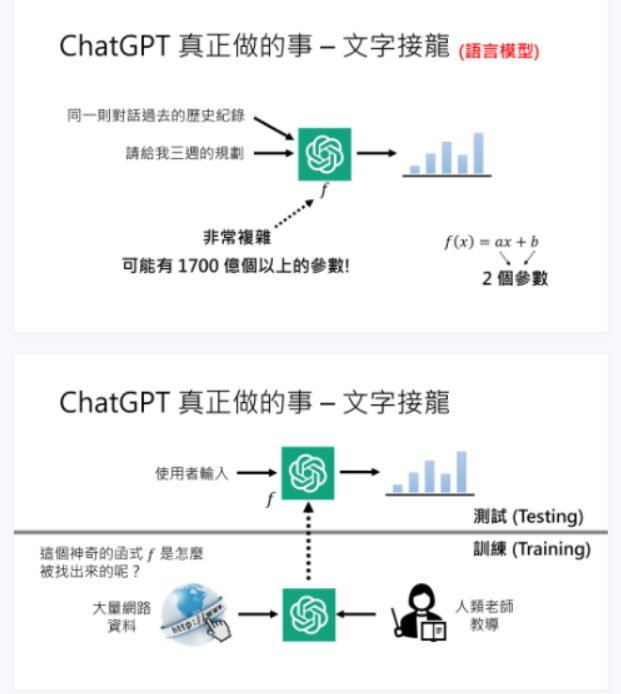
# 1. 对ChatGPT的认识

## 对ChatGPT的常见误解

1. ChatGPT用事先准备好的“罐头”回应
2. ChatGPT是网络搜索的结果，实际上ChatGPT并未联网，目前部分可以
3. ChatGPT真正在做的事情是-->文字接龙（大语言模型）





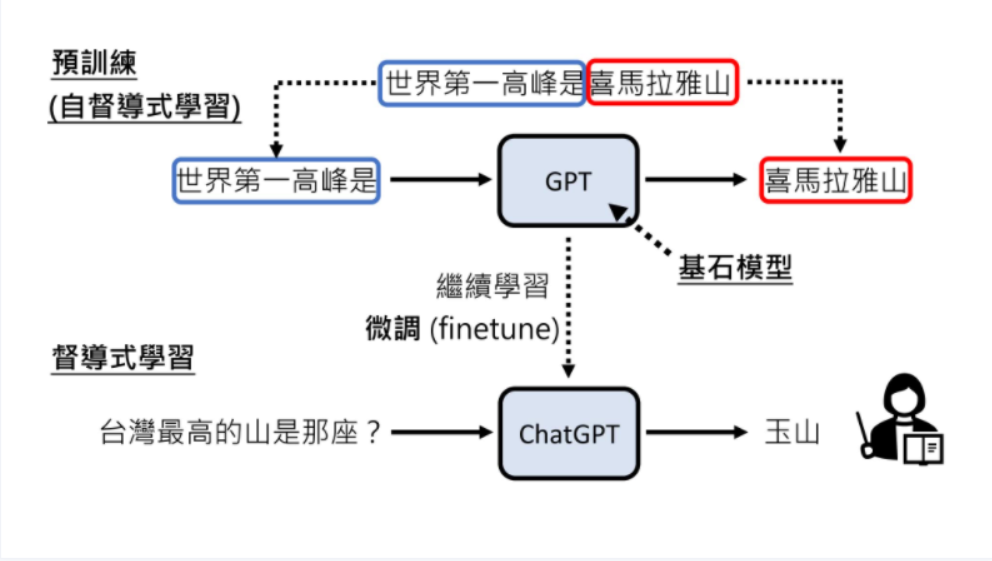
## **ChatGPT的预训练**

1. 无监督学习（预训练）与监督学习

监督学习（Supervised Learning）：

监督学习是一种利用带有标签（标记）的数据进行训练的机器学习方法。在监督学习中，训练数据包含输入样本和对应的标签（预期输出）。学习的目标是通过训练数据构建一个模型，该模型能够对新的未标记数据进行预测或分类。监督学习的典型应用包括分类（如垃圾邮件识别）和回归（如房价预测）等。

无监督学习（Unsupervised Learning）：

无监督学习是一种在没有标签（标记）的数据中发现模式和结构的机器学习方法。在无监督学习中，训练数据只包含输入样本，没有相应的标签或预期输出。学习的目标是从数据中推断出隐藏的结构、关系或规律。无监督学习的典型应用包括聚类（将数据划分为类别）、降维（减少数据的维度）和关联规则挖掘等。

## **ChatGPT带来的研究问题**

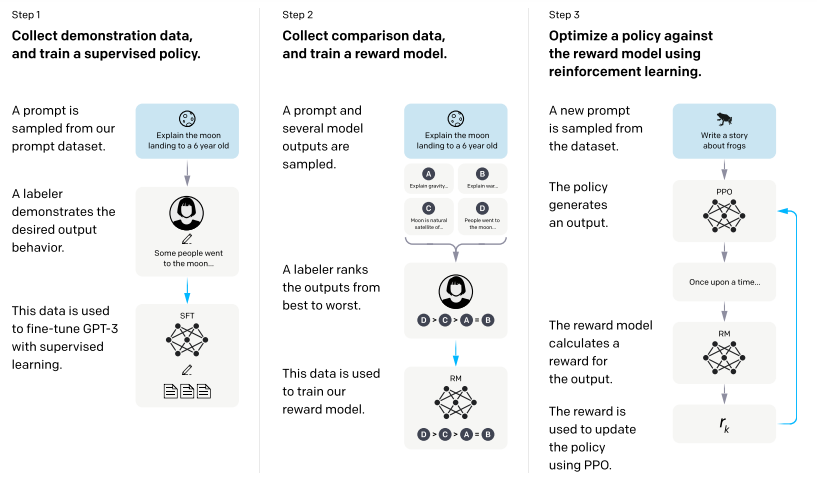
（1）如何精确提出需求。

（2）如何更正错误。

（3）判断物件是否由AI生成。

（4）隐私问题。

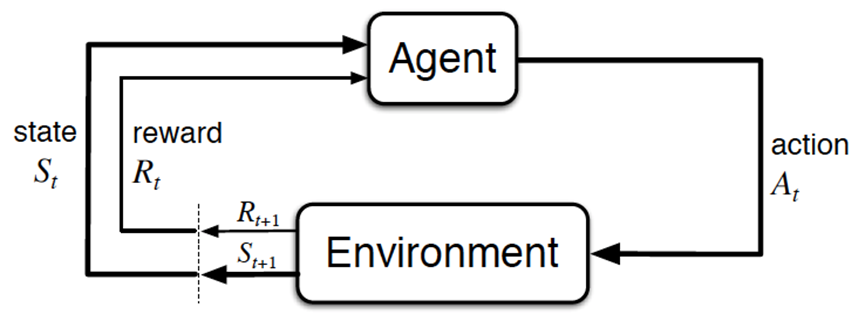
## **ChatGPT的学习与社会化过程**





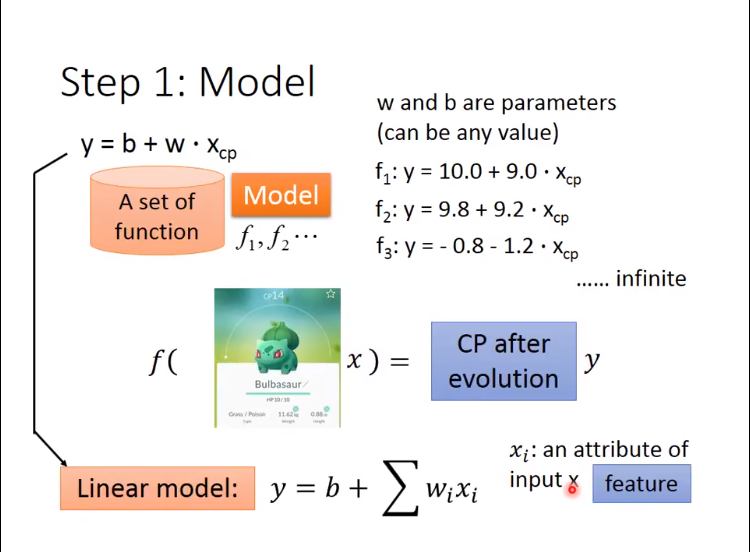
GPT的输出是几率分布问题，因此每次的输出都不太可能一致。

经典的强化学习模型可以总结为下图的形式（任何强化学习都包含这几个基本部分：智能体、行为、环境、状态、奖励）：



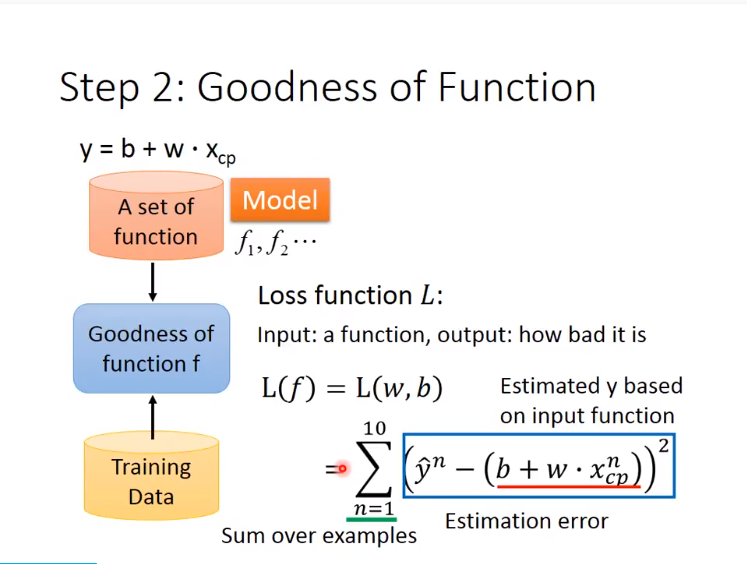
# Regression-Case Study

## Step 1: Model



建立一个函数，从输入到输出，模型的表达式和参数可以自定，也可由机器自动生成与调整。

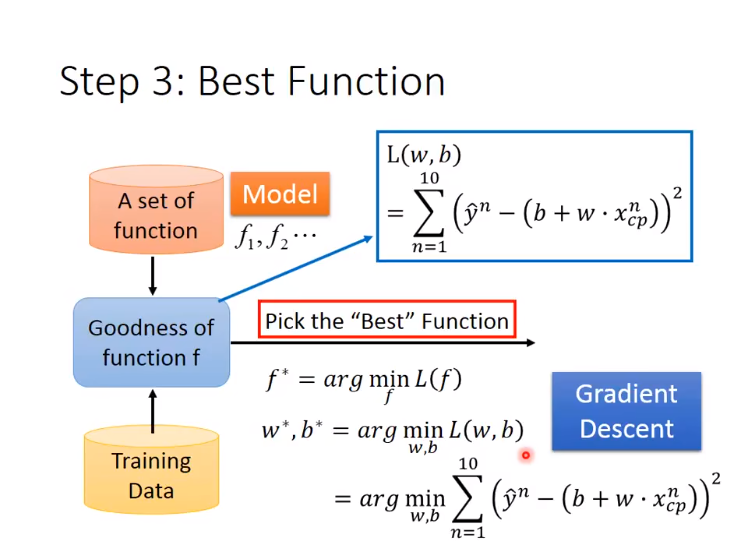
## Step 2: Goodness of Function

、

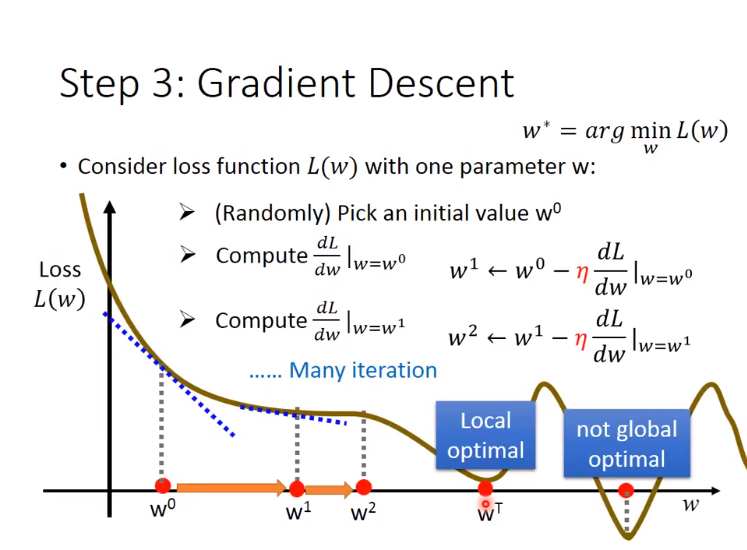
损失函数是衡量模型好坏的重要指标，此处是预测值与实际值平方差的和。

## Step 3: Best Function

方法：梯度下降 Gradient Decent



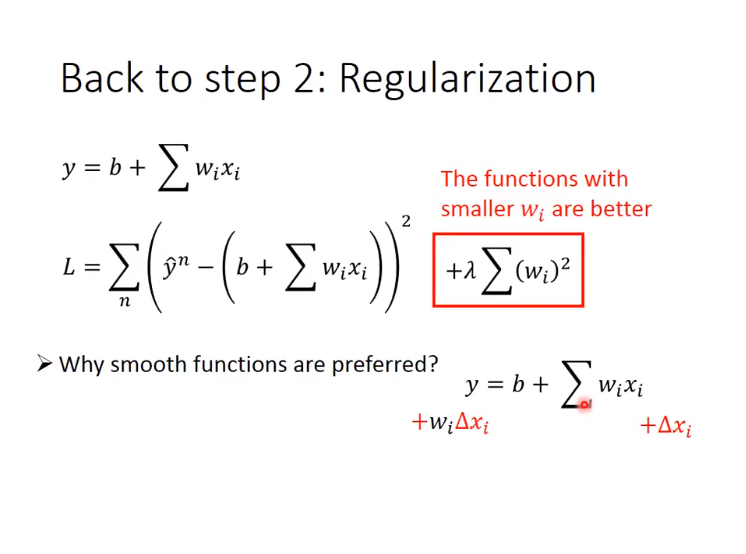
argmin即取loss最小时的参数。



梯度下降方法在回归模型中常用，主要是对参数进行微分，判断参数下一步如何调整才能降低loss，但会存在局部最优解的问题。

复杂的模型在训练集上有较好的表现，但应用于测试集时表现较差，说明出现了过拟合（Overfitting），因此需要尝试多种Function来对比，从而选取最合适的模型。

出现过拟合时。可以在损失函数中添加正则化（regularization）参数来限制参数的变化，防止出现过拟合情况。正则化参数也不能过大，否则会出现欠拟合的情况。



# Classification

## Classification分类：概率生成模型

**线性回归算法的output是一个数值，而分类算法的input依旧是features，但output是一个类别。**



**分类问题的三步骤**

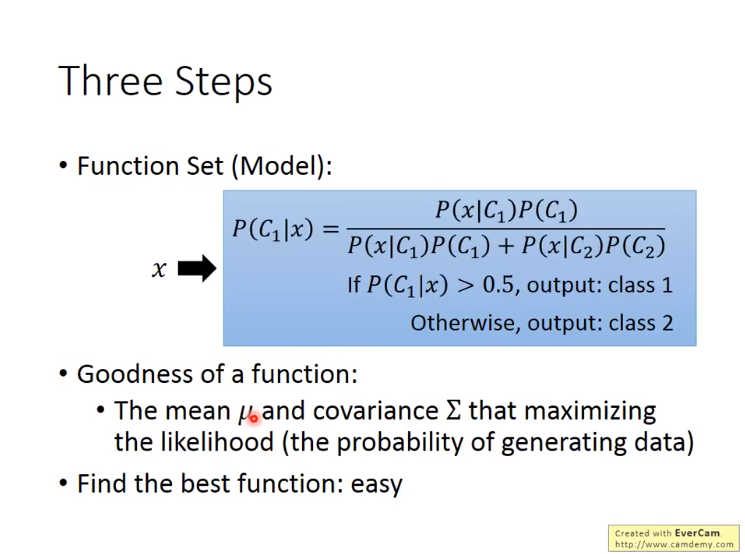
(1)定义模型集合，样本1属于类别1的概率

(2)定义损失函数loss来评价模型好坏

(3)找到最佳模型

**概率分布选择**

选择自己喜欢的概率分布，二元特征可用伯努利，所有维度独立可用朴素贝叶斯



## 朴素贝叶斯分类方法

朴素贝叶斯算法的核心思想是通过考虑特征概率来预测分类，即对于给出的待分类样本，求解在此样本出现的条件下各个类别出现的概率，哪个最大，就认为此待分类样本属于哪个类别。

之所以称为“朴素”或“朴素贝叶斯”，是因为如果对每种标签的生成模型进行非常简单的假设，就能找到每种类型生成模型的近似解，然后就可以使用贝叶斯分类。不同类型的朴素贝叶斯分类器是由对数据的不同假设决定的。

**朴素贝叶斯分类器**

先验概率P(X)：先验概率是指根据以往经验和分析得到的概率。

后验概率P(Y|X)：事情已发生，要求这件事情发生的原因是由某个因素引起的可能性的大小，后验分布P(Y|X)表示事件X已经发生的前提下，事件Y发生的概率，称事件X发生下事件Y的条件概率。

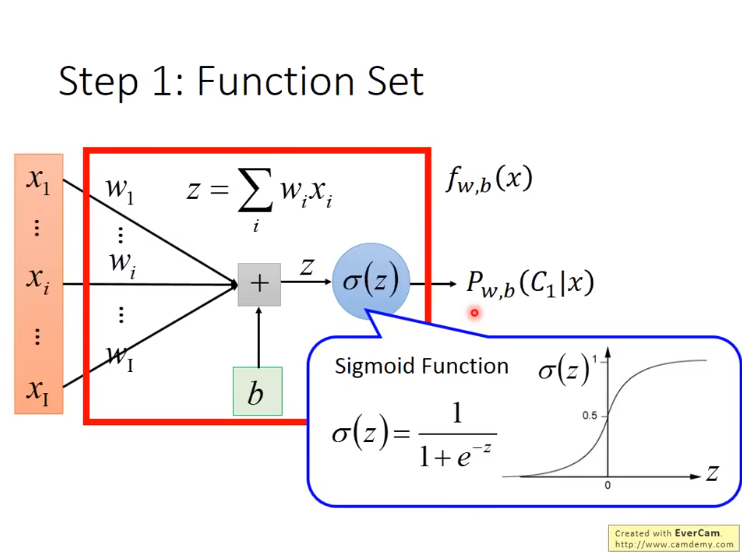
后验概率P(X|Y)：在已知Y发生后X的条件概率，也由于知道Y的取值而被称为X的后验概率。

朴素：朴素贝叶斯算法是假设各个特征之间相互独立。

# Logistic Regression

逻辑回归是机器学习领域中一种用于[二分类问题](https://so.csdn.net/so/search?q=%E4%BA%8C%E5%88%86%E7%B1%BB%E9%97%AE%E9%A2%98&spm=1001.2101.3001.7020" \t "https://blog.csdn.net/qq_51447496/article/details/_blank)的常用算法。尽管其名字中包含"回归"一词，但实际上，逻辑回归是一种分类算法，用于估计输入特征与某个事件发生的概率之间的关系。

## Step 1: Model



**W是权重，b是偏差，X是输出。**

**激活函数 sigmiod function**

Sigmoid函数：

Sigmoid函数的图像呈现一种S形曲线，当输入值较大时，趋近于1；当输入值较小时，趋近于0。虽然Sigmoid函数在某些情况下被广泛使用，但由于其存在的梯度消失问题，随着深度学习的发展，其在隐藏层中的应用逐渐被ReLU、Leaky ReLU、ELU等激活函数所取代。

1. 作用

将输入值映射到一个区间[0,1]之间，实现非线性变换。

常用于二分类问题的输出层，将输出值转化为概率值。

在浅层网络中，用于处理二元分类任务。

2. 优点

具有很好的数学性质，具备平滑性和连续性。

输出范围在之间，可以被解释为概率。

相对简单，易于理解和实现。

3. 缺点

容易出现梯度消失问题：在函数两端，梯度接近于零，导致在反向传播过程中，参数更新缓慢，特别是在深度网络中。

输出不是以零为中心的：这可能导致网络的收敛速度变慢。

饱和区域梯度较小：在输入较大或较小时，梯度会变得很小，导致训练过程中的梯度消失问题。

4. 数学公式

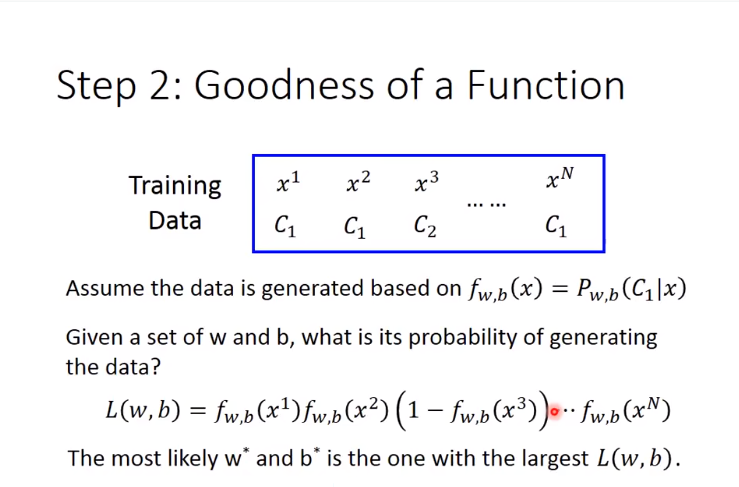
Sigmoid激活函数的数学公式如下：

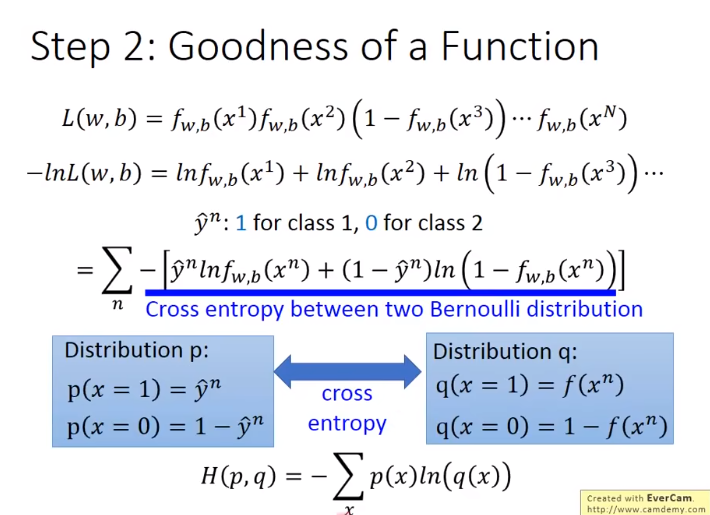


其中，e是自然对数的底数，x是输入值。

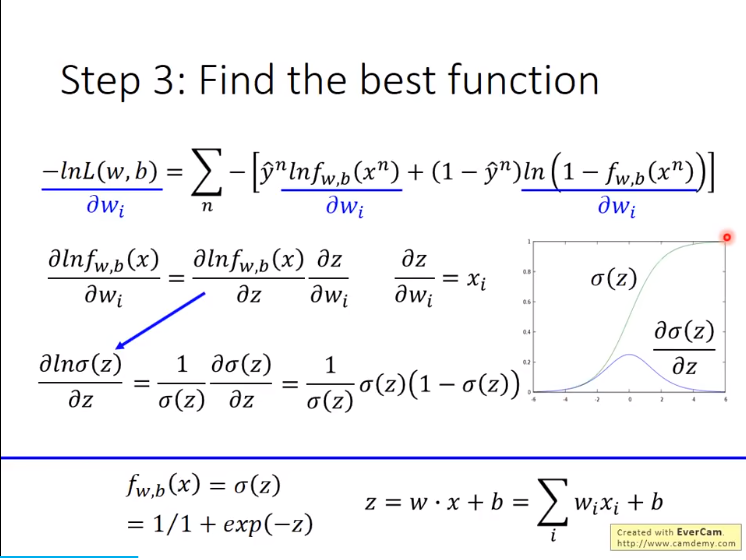
## Step 2: Goodness of Function

已知一组训练数据，假定训练数据来自后验概率，已知一组w和b，求产生这组数据的概率。最大可能性产生这组训练数据的w和b ，函数整理、变换。





## Step 3: Best Function

经过计算找到最好的函数。（计算过程？）

## 逻辑回归与线性回归的对比

