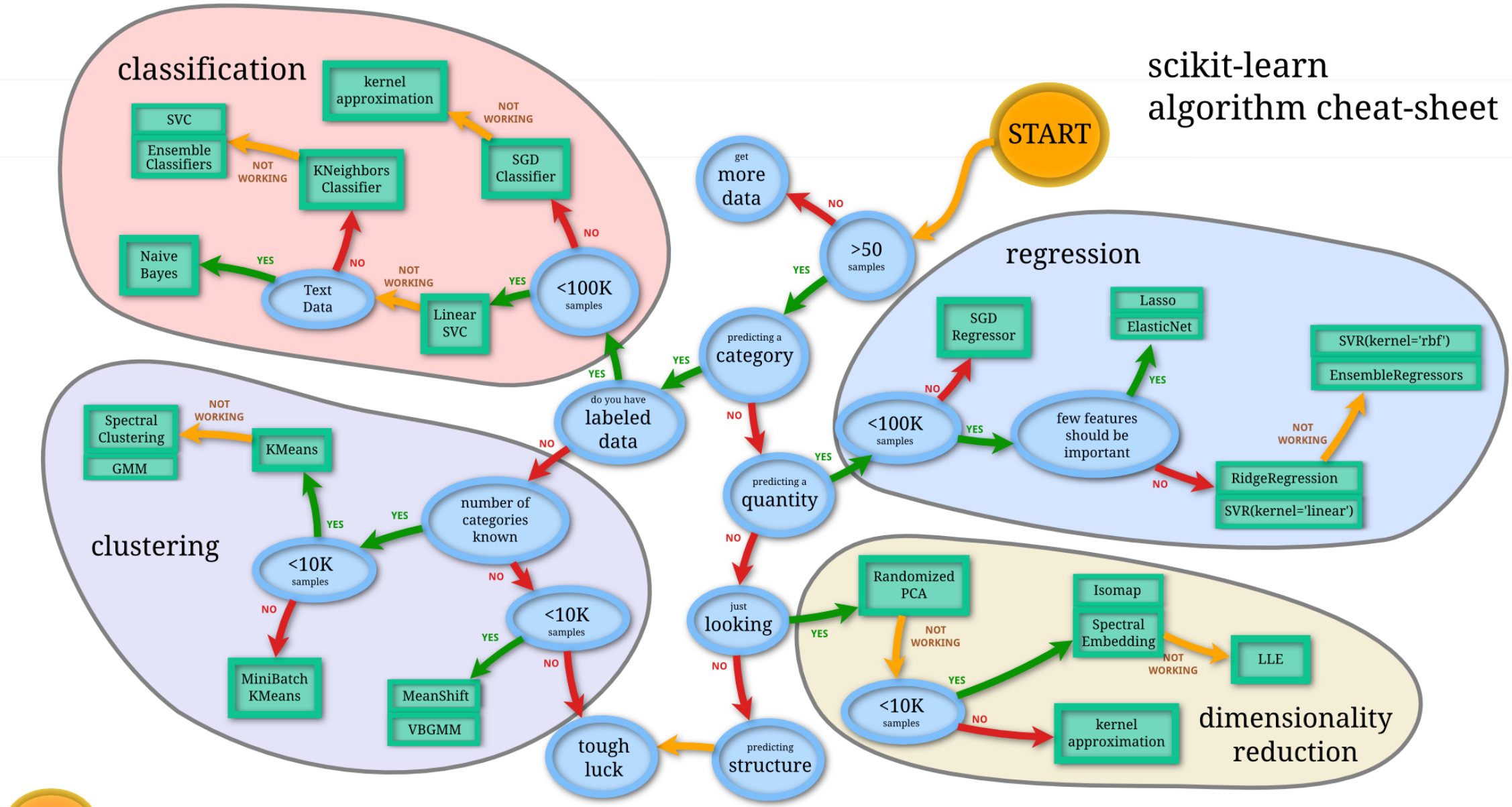
Logistic Regression Basics

Week 5 Day 4

scikit-learn algorithm cheat-sheet



Back

AGENDA

1. Motivation
2. What is Logistic Regression?
3. Why use Logistic Regression
4. Lab

SUPERVISED LEARNING - REGRESSION & CLASSIFICATION

If the y variable is numeric then we have a regression problem - we are trying to predict a continuous number

If the y variable is a category (for example trying to predict a type of flower) the we have a classification problem - we are trying to classify what group that y belongs to.

WHAT TO USE & WHEN

- The more _____ and the more ____, but the less ____ means a higher probability of _____

- Logistic regression will probably work
- The model will be comprehensible

WHAT TO USE & WHEN

- The more young children you are looking after and the more hungry you are, but the less alternative eating options in your area means a higher probability of you eating dinner tonight at McDonalds

- Logistic regression will probably work
- The model will be comprehensible

WHEN LOGISTIC REGRESSION *WON'T* WORK WELL

- You are more likely to ____ if _____ unless _____, in which case _____
  - A decision tree will probably work better
- There's this complicated mathematical function that relates ____ and ____ to the probability of _____
  - Try a support vector machine

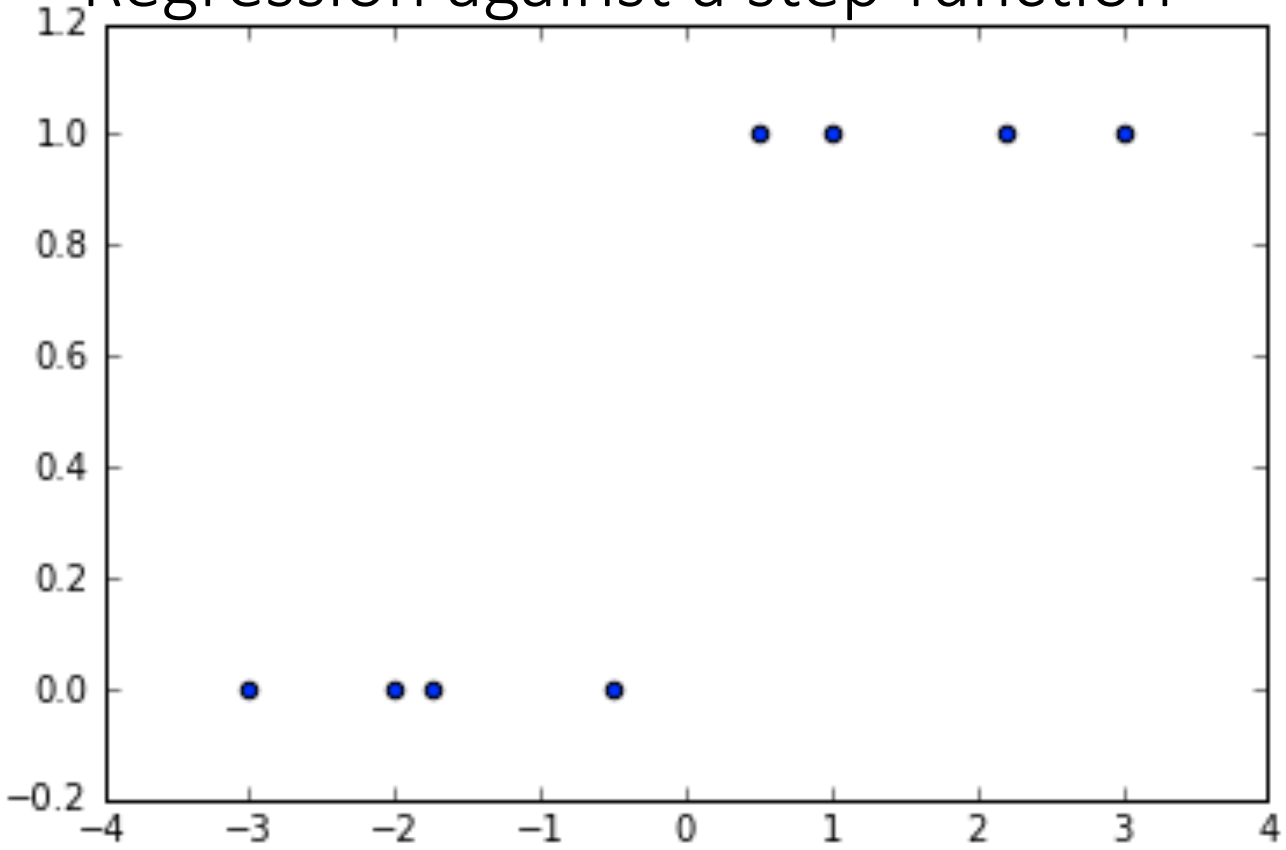
WHEN LOGISTIC REGRESSION *WON'T* WORK WELL

- You are more likely to buy a bright pink phone case if your age is between 8 and 14 unless you have more than twenty cases already, in which case it is between the ages of 40 and 60.
  - A decision tree will probably work better
- If the square of the number of items bought divided by the average cost is greater than 1 then probability of an item return increases.
  - Try a support vector machine

WHAT IS STEP REGRESSION?

In statistics, stepwise regression is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure. In each step, a variable is considered for addition to or subtraction from the set of explanatory variables based on some prespecified criterion.

Regression against a step-function



Turn your category into a number:

- "In category" = 1
- "Not in category" = 0

Find the break point.

Suggested regression:

F(x) = 0 when x < 0

F(x) = 1 when x >= 0

And buy yourself a lottery ticket, because nothing ever works out this well

WHAT IS LOGISTIC REGRESSION?

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).

In logistic regression, the dependent variable is binary or dichotomous, i.e. it only contains data coded as 1 (TRUE, success, pregnant, etc.) or 0 (FALSE, failure, non-pregnant, etc.).

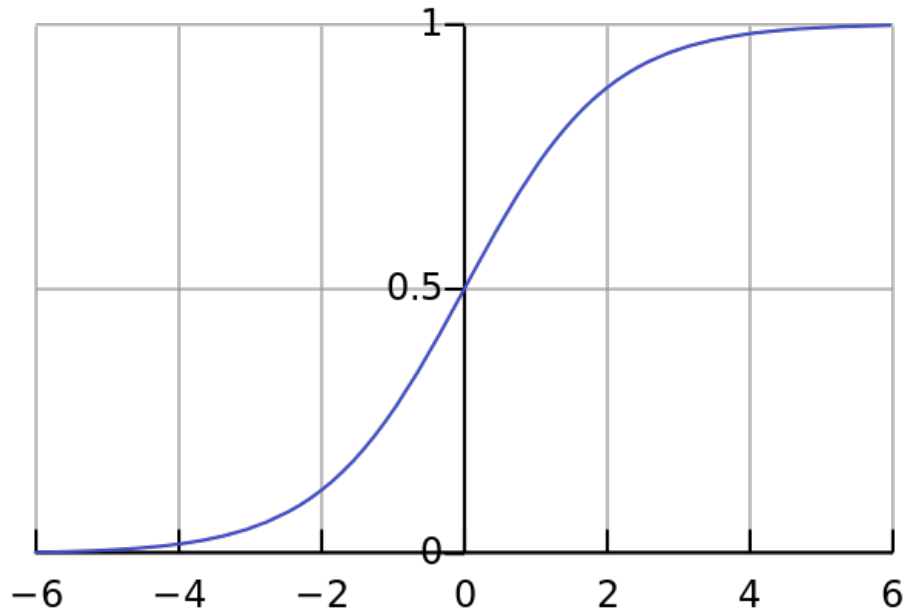
An explanation of logistic regression can begin with an explanation of the standard logistic function. The logistic function is useful because it can take any real input, whereas the output always takes values between zero and one and hence is interpretable as a probability.

$$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$$

$$t = \beta_0 + \beta_1 x$$

Explanetory variable is x

$$F(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$



Multiple Explanetory variables

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m.$$

The goal of logistic regression is to find the best fitting (yet biologically reasonable) model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables.

Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a *logit transformation* of the probability of presence of the characteristic of interest:

$$logit(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k$$

where p is the probability of presence of the characteristic of interest. The logit transformation is defined as the logged odds:

$$odds = \frac{p}{1-p} = \frac{probability\ of\ presence\ of\ characteristic}{probability\ of\ absence\ of\ characteristic}$$

$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$

Type of questions that a logistics regression can examine.

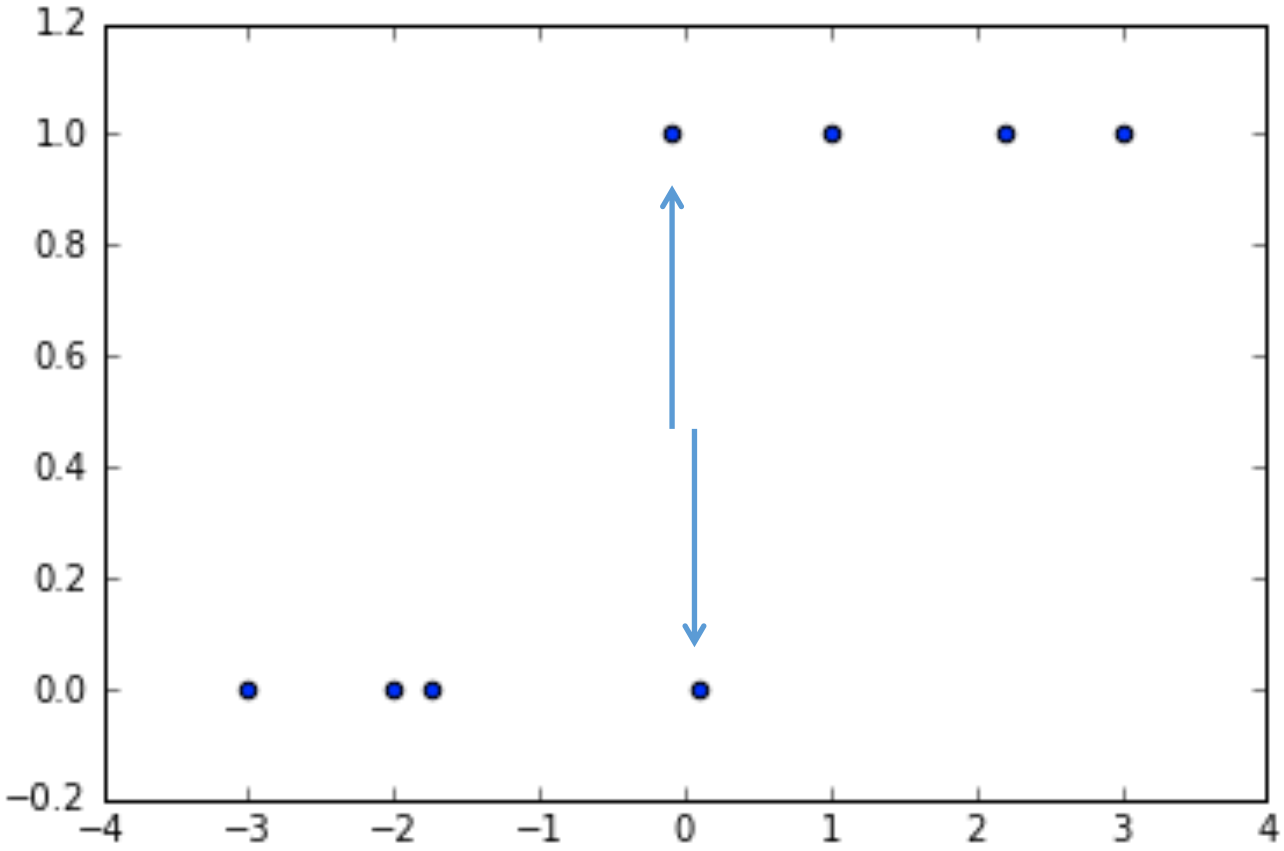
How does the probability of getting lung cancer (yes vs. no) change for every additional pound of overweight and for every pack of cigarettes smoked per day?

Do body weight calorie intake, fat intake, and participant age have an influence on heart attacks (yes vs. no)?

The best you could hope for really…

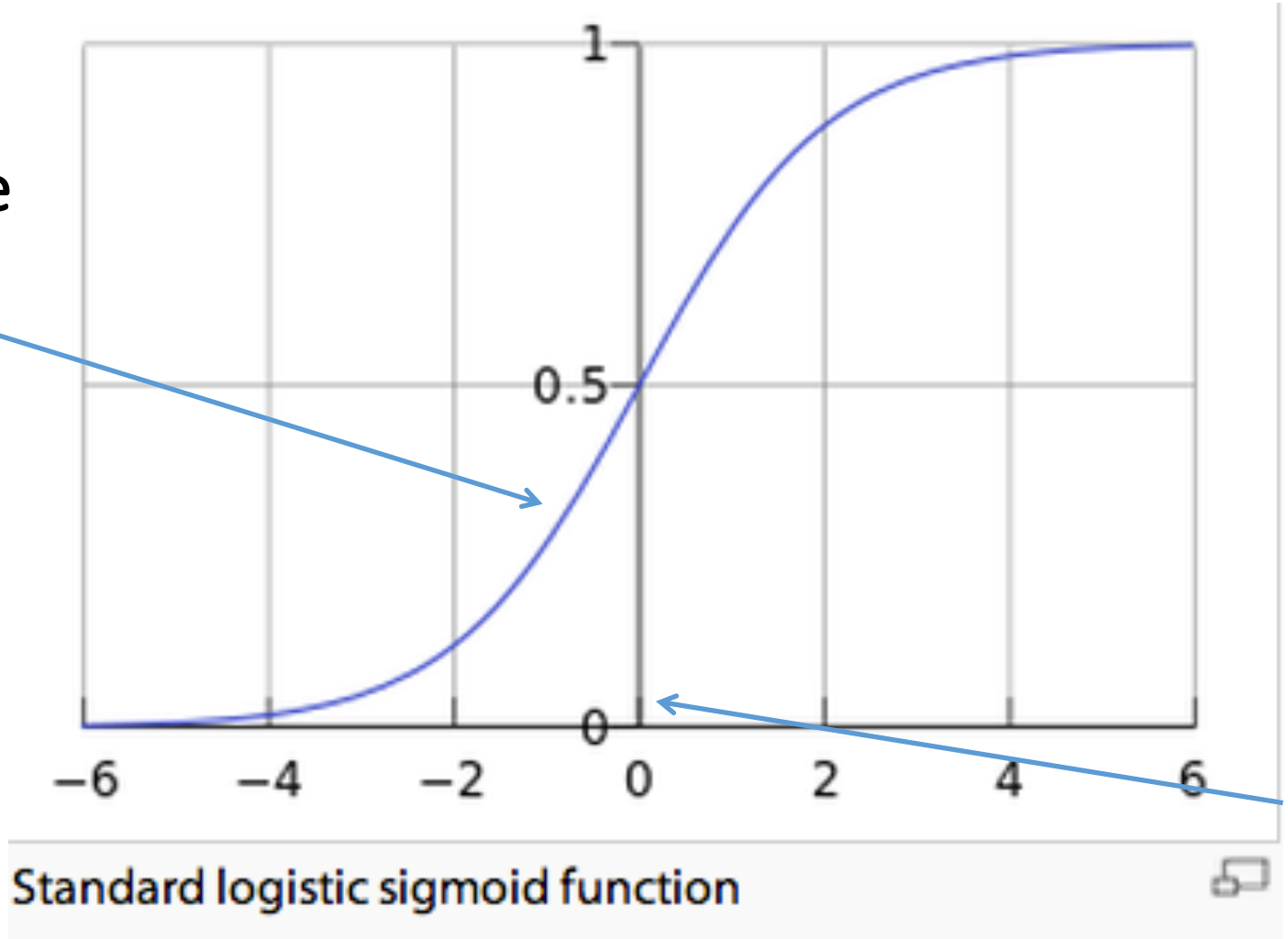
Near x=0, we don't know what category to choose: 50% either way

```
F(x) = close to zero if x is
really negative

F(x) = close to one if x is
really positive

F(x) = 0.5 if x is near zero
```



Let's smooth that function a bit

We can adjust the slope.



Standard logistic sigmoid function

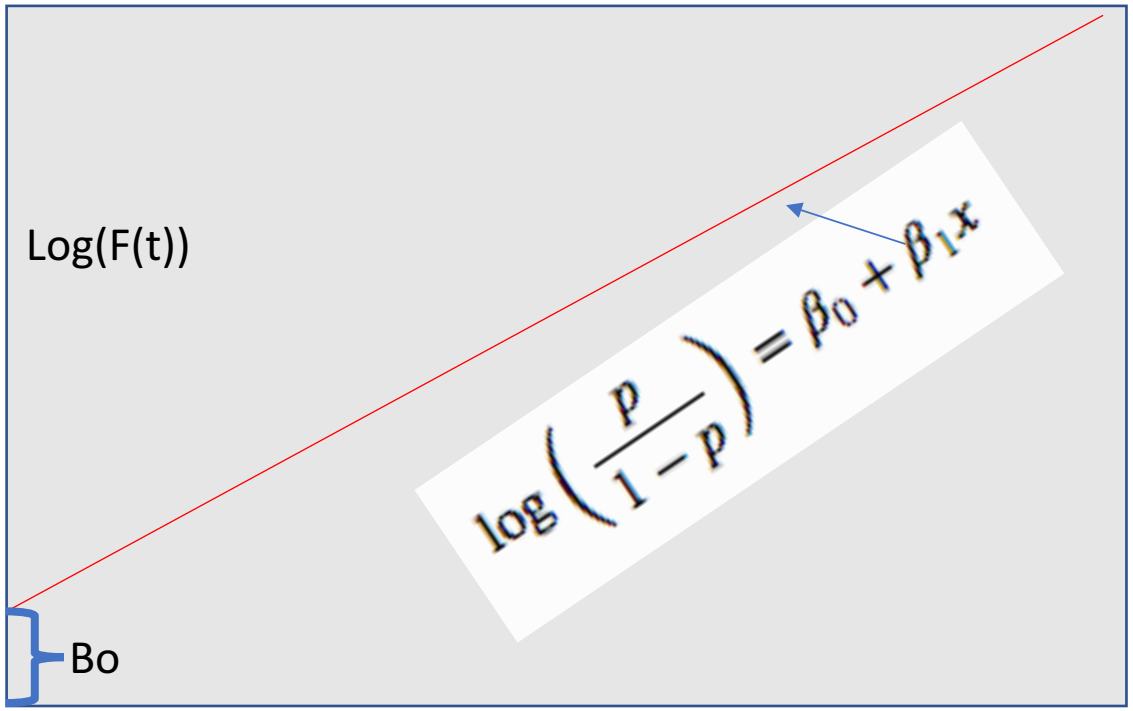
We can adjust the half-way point.

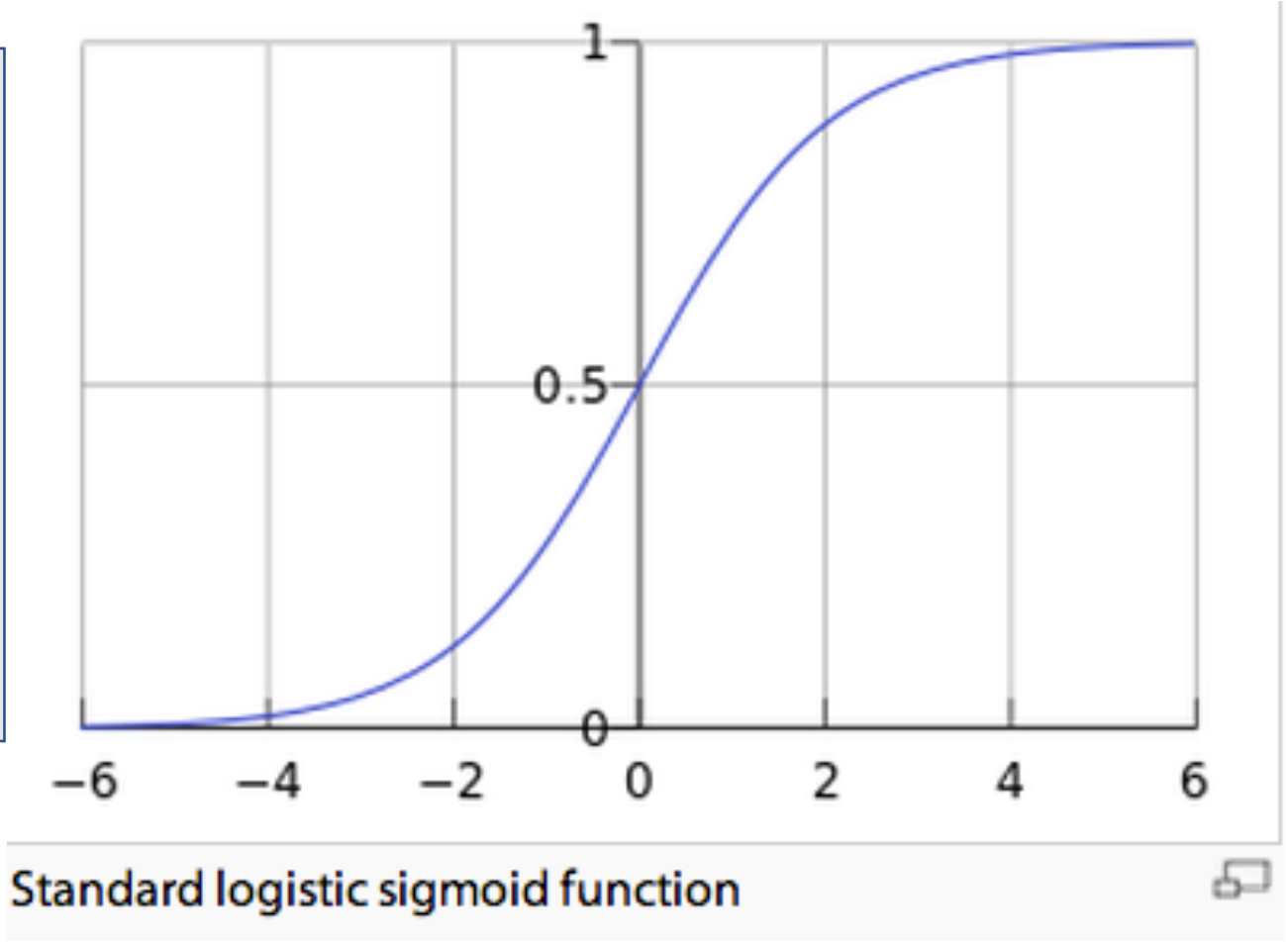
All the maths…

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

- Use the logit function
- *p* is the probability of being in a category (between 0.0 and 1.0)
- We can see that it this function is linear in X
- $\frac{p}{1-p}$ is called the ‘odds’ and can be any value from 0 to ∞
- $\log\left(\frac{p}{1-p}\right)$ is called the ‘log-odds’ or ‘logit’

Linear in Logit





- Fitted Regression line = $B_o + B_1X$

- Logit(F(t))=0.998 – 0.014*weight
- e^{B1} = 0.986 = odds ratio
- Scenario: Data collected on women with babies, some were normal weight and some had low birth weight. Variables thought to be associated are age, weight of mother, smoking and number of visits to a physician.
- Interpretation: Pick weight of mother for x. The odds of having a low birth weight baby are 0.986 times the odds of having a normal birth weight baby for each one pound increase in weight.
- Is a 1 pound weight change really going to have an effect?

- Scale the odds ratio to reflect a larger increment of the variable in question
- Calculate the odds ratio for a 5 pound change in weight
- e^{5*B1} = 0.93 = odds ratio

- The odds of having a low birth weight baby are 0.93 the odds of having a normal birth weight baby for each 5 pound increase in weight
- Testing the slope for weight is zero
- Z-statistic =-2.28, p-value = 0.02
- P-value<0.05 reject H_0 and conclude that result is statistically significant
- -→evidence indicates slope of weight is nonzero

SCIKIT LEARN LOGISTIC REGRESSION


```
import sklearn.linear_model
Regressor = sklearn.linear_model.LogisticRegression()
X = my_dataframe[['var1','var2','var3']]
Y = my_dataframe.target_column
Regressor.fit(X,y)
Regressor.predict(X)
Regressor.predict_proba(X)
```

Other things to be aware of

- LogisticRegressor(C=100000000)
  - “I don’t care if I overfit the data”
- regressor.predict_proba([[…]])
  - Returns list: probability of y=0, y=1, y=2…