GENERAL ASSEMBLY

# INTRO TO DATA SCIENCE

## LECTURE 7: PROBABILITY & LOGISTIC REGRESSION

# LAST TIME:

- LINEAR REGRESSION
- REGULARIZATION

QUESTIONS?

I. REVIEW OF REGULARIZATION

II. LOGISTIC REGRESSION

*These regularization problems can also be expressed as:*

**OLS:** $\min_\beta(\|y - X\beta\|_2^2)$

**L1 regularization:** $\min_\beta(\|y - X\beta\|_2^2 + \alpha\|\beta\|_1)$

**L2 regularization:** $\min_\beta(\|y - X\beta\|_2^2 + \alpha\|\beta\|_2^2)$

*We are no longer just minimizing error but also an additional term to penalize model complexity.*

# II. LOGISTIC REGRESSION

| | continuous | categorical |
|---|---|---|
| supervised | regression | classification |
| unsupervised | dimension reduction | clustering |

*Q: What is* **logistic regression***?*

*Q: What is* **logistic regression***?*

*A: A generalization of the linear regression model to classification problems.*

*In linear regression, we used a set of covariates to predict the value of a (continuous) outcome variable.*

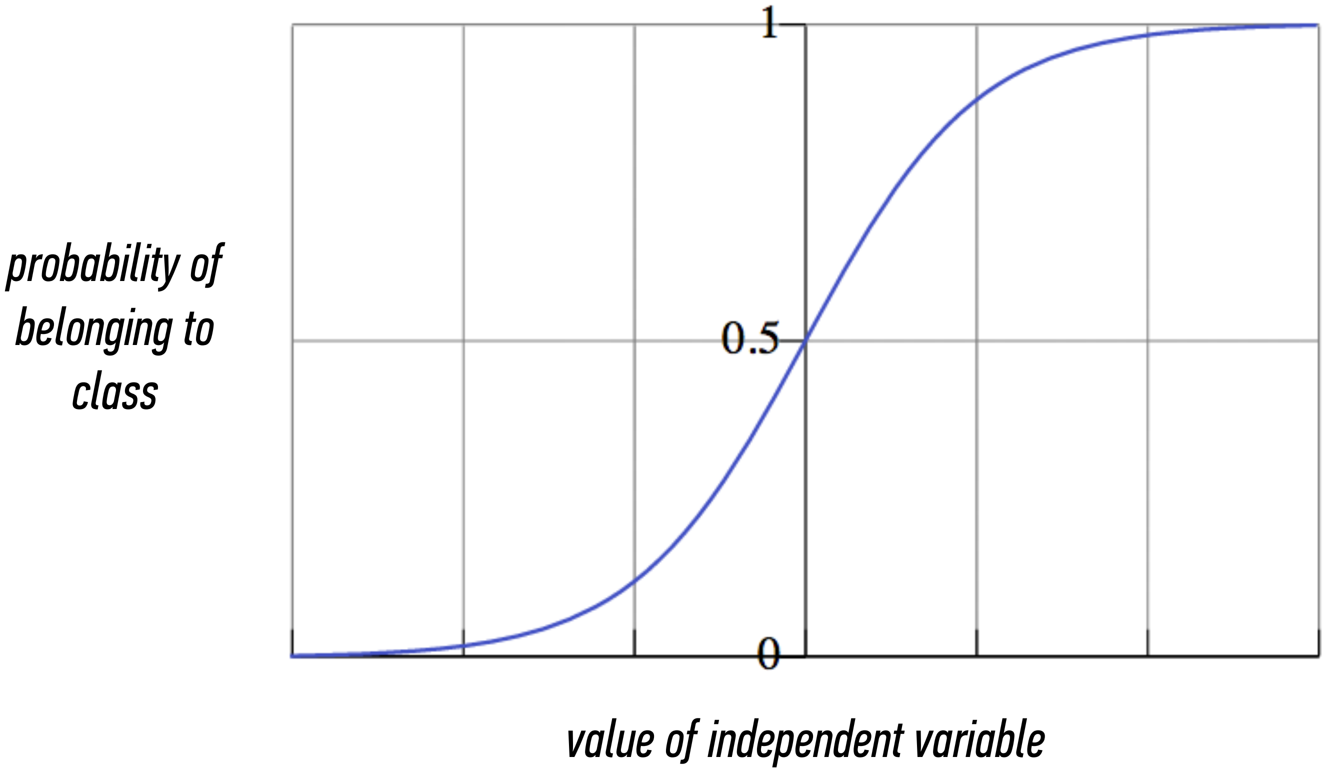
*In **linear regression**, we used a set of covariates to predict the **value of a (continuous) outcome variable**.*

*In **logistic regression**, we use a set of covariates to predict **probabilities of class membership**.*

*In linear regression, we used a set of covariates to predict the value of a (continuous) outcome variable.*

*In logistic regression, we use a set of covariates to predict probabilities of class membership.*

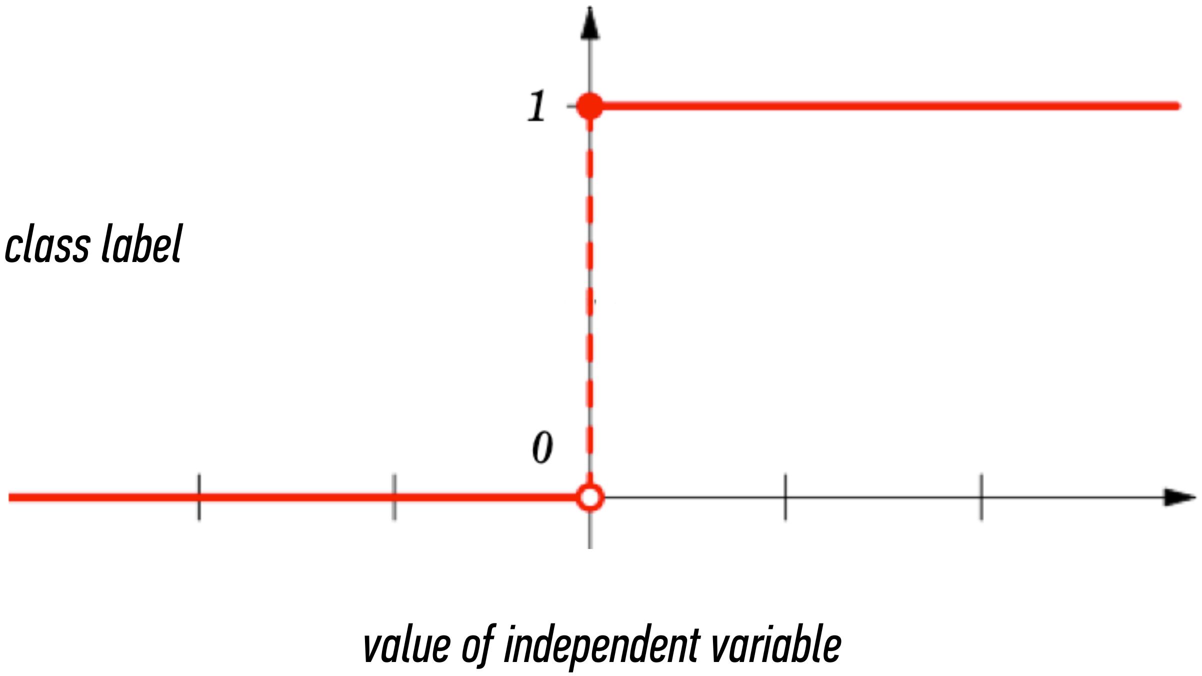
*These* ***probabilities are then mapped to class labels****, thus solving the classification problem.*




**NOTE**

Probability predictions look like this.

# CLASS LABELS




**NOTE**

Probabilities are "snapped" to class labels (eg by threshholding at 50%).

*The logistic regression model is an extension of the linear regression model, with a couple of important differences.*

*The logistic regression model is an extension of the linear regression model, with a couple of important differences.*

*The main difference is in the outcome variable.*

*The key variable in any regression problem is the* **response type** *of the outcome variable* y *given the value of the covariate* x*:*

$$E(y|x)$$

*The key variable in any regression problem is the* **conditional mean** *of the outcome variable* y *given the value of the covariate* x*:*

$$E(y|x)$$

*In linear regression, we assume that this conditional mean is a linear function taking values in* $(-\infty, +\infty)$*:*

$$E(y|x) = \alpha + \beta x$$

*In logistic regression, we've seen that the conditional mean of the outcome variable takes values only in the unit interval* `[0, 1]`.

*In logistic regression, we've seen that the conditional mean of the outcome variable takes values only in the unit interval* `[0, 1]`.

*The first step in extending the linear regression model to logistic regression is to map the outcome variable* $E(y|x)$ *into the unit interval.*

*In logistic regression, we've seen that the conditional mean of the outcome variable takes values only in the unit interval* $[0, 1]$.

*The first step in extending the linear regression model to logistic regression is to map the outcome variable* $E(y|x)$ *into the unit interval.*

*Q: How do we do this?*

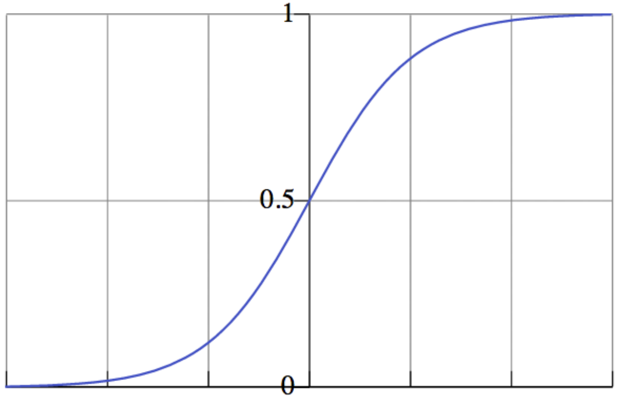
*A: By using a transformation called the* **logistic function***:*

$$E(y|x) = \pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

*A: By using a transformation called the* **logistic function***:*

$$E(y|x) = \pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

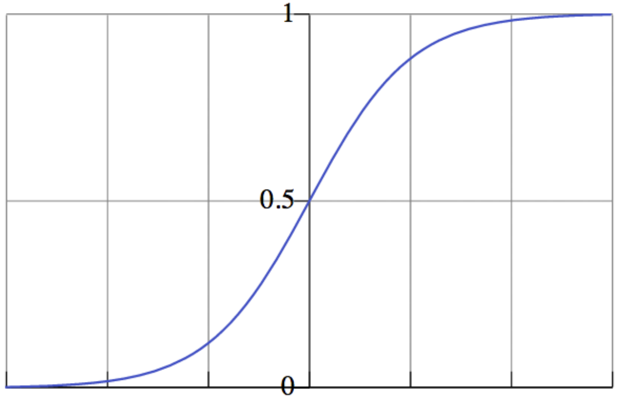
*We've already seen what this looks like:*



*A: By using a transformation called the* **logistic function***:*

$$E(y|x) = \pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

*We've already seen what this looks like:*




**NOTE**

For any value of x, y is in the interval [0, 1]

This is a nonlinear transformation!

*The* **logit function** *is an important transformation of the logistic function. Notice that it returns the linear model!*

$$g(x) = ln(\frac{\pi(x)}{1-\pi(x)}) = \alpha + \beta x$$

*The* **logit function** *is an important transformation of the logistic function. Notice that it returns the linear model!*

$$g(x) = ln(\frac{\pi(x)}{1-\pi(x)}) = \alpha + \beta x$$

*The logit function is also called the* **log-odds function***.*