Michael Grossberg

Intro to Data Science CS59969

Naive Bayes and Logistic Regression Classification

Naive Bayes Classifier

Conditional Probability

X = measurements

C= condition

$$P(C \mid X)$$

What we know

What we want to know

Normalized for fixed X

X = measurements

C= condition

$$\sum_i P(C_i | X) = 1$$

X is fixed

Whats our best guess?

X = measurements

C= condition

$$C^{*} = \text{argmax}_C P(C \mid X)$$

Maximum a posteriori (MAP) Estimate

Joint vs Conditional

$$P(C \mid X)\,P(X) = P(C,X)$$

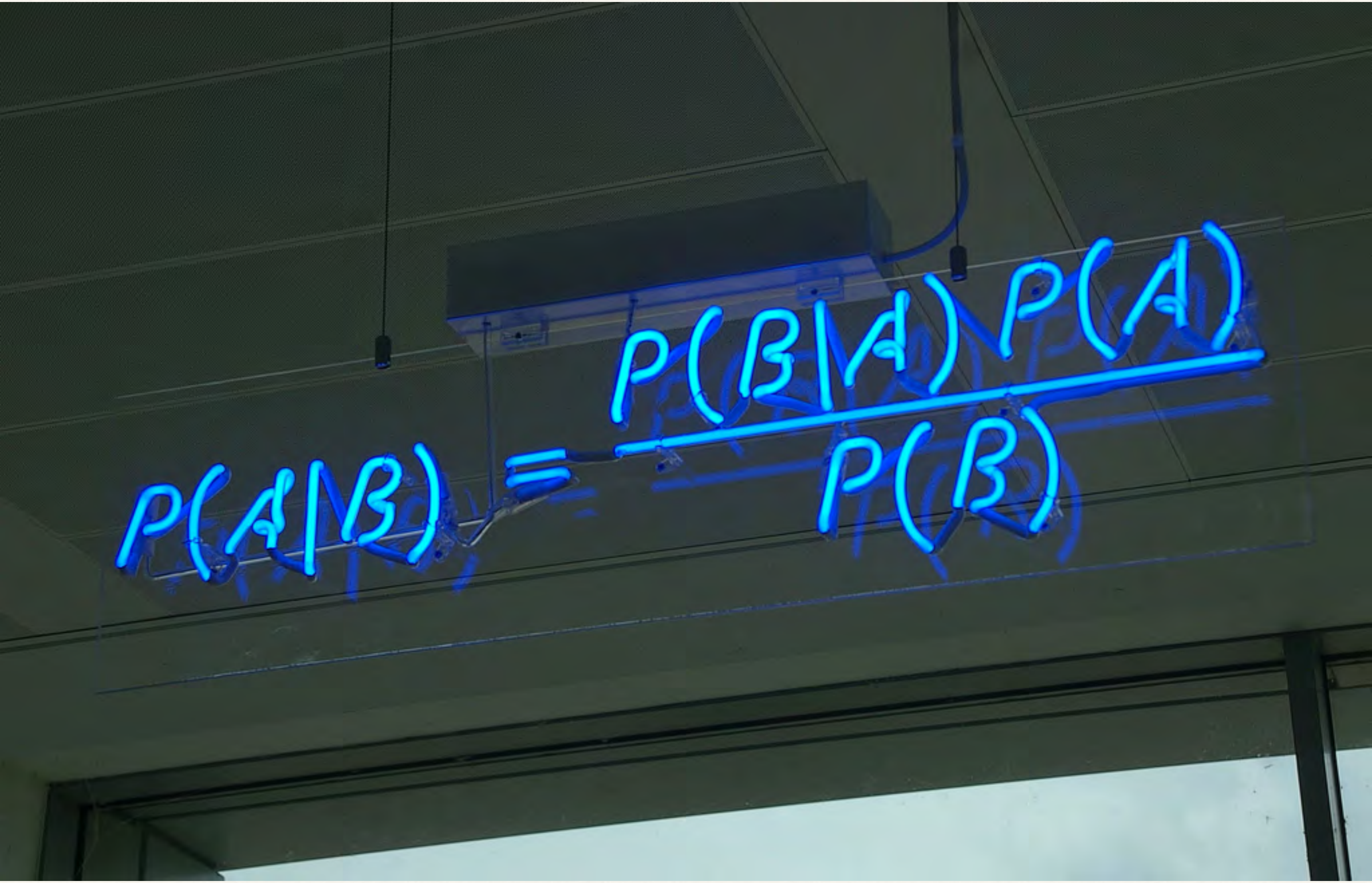
$$P(C \mid X) = \frac{P(C,X)}{P(X)}$$

Joint vs Conditional

$$P(C \mid X)\, P(X) = P(C,X)$$

$$= P(X,C) = P(X \mid C)\, P(C)$$

$$P(C \mid X) = \frac{P(X \mid C)\, P(C)}{P(X)}$$


P(A|B) = P(B|A) P(A) / P(B)

Bayes Theorem

Likelihood Prior

$$P(C \mid X) = \frac{P(X \mid C)\, P(C)}{P(X)}$$

Evidence

A-posteriori probability

Likelihood

$$P(X \mid C)$$

Measurement

Given the condition

Prior

$$P(C)$$

What is P(C) without knowing a measurements?

Maximum Likihood

What if we don't know P(C)?

$$P(C|X) = \frac{P(X|C)\,\cancel{P(C)}}{\cancel{P(X)}}$$

$$P(C|X) \propto P(X|C)$$

$$C^* = \text{argmax}_C P(C|X) = \text{argmax}_C P(C|X)$$

Normal Distribution

$$\mathcal{N}(X \mid \boldsymbol{\theta}) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp(-(X-\mu)^2/2\sigma^2)$$

$$\boldsymbol{\theta} = (\mu, \sigma)$$

Given data what are the parameters?

$$P(Z \mid \boldsymbol{\theta}) = \mathcal{N}(Z \mid \boldsymbol{\theta})$$

$$\boldsymbol{\theta}^{*} = \arg\max_{\boldsymbol{\theta}} \mathcal{N}(Z_1 \mid \boldsymbol{\theta}) \; \ldots . \; \mathcal{N}(Z_N \mid \boldsymbol{\theta})$$

Z same variable (attribute),
different independent trials

Given data what are the parameters?

$$P(Z \mid \boldsymbol{\theta}) = \mathcal{N}(Z \mid \boldsymbol{\theta})$$

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \mathcal{N}(Z_1 \mid \boldsymbol{\theta}) \ldots. \mathcal{N}(Z_N \mid \boldsymbol{\theta})$$

answer:

$$\Rightarrow \quad \boldsymbol{\theta}^* = (\mu, \sigma)$$

Mean

Standard Deviation

Bayes

$$P(C \mid X_1, \ldots, X_N) \propto P(X_1, \ldots, X_N \mid C)\, P(C)$$

Independence

$$P(X_1, \ldots, X_N) = P(X_1) \cdots P(X_N)$$

$$P(C \mid X_1, \ldots, X_N) = P(C \mid X_1) \cdots P(C \mid X_N)$$

Different attributes

Usually Not True ….

Naive Bayes (Assume Independence)

$$P(C \mid X_1, \dots, X_N) \propto P(C \mid X_1) \cdots P(C \mid X_N)\, P(C)$$

Argmax

Banana

Example

Fruit	Long	Sweet	Yellow	Total
Banana	400	350	450	500
Orange	0	150	300	300
Other	100	150	50	200
Total	**500**	**650**	**800**	**1000**

50% of the fruits are bananas
30% are oranges
20% are other fruits

P(Banana) = .5
P(Orange) = .3
P(Other) = .2

Priors

Measurement

Features

X_1 = long
X_2 = sweet
X_3= yellow

What is it? Banana? Orange? or Other?

Naive Bayes Formula

Banana:

$$
\begin{aligned}
&P(Banana|Long, Sweet, Yellow) \\
&= \frac{P(Long|Banana) \cdot P(Sweet|Banana) \cdot P(Yellow|Banana) \cdot P(Banana)}{P(Long) \cdot P(Sweet) \cdot P(Yellow)} \\
&= \frac{0.8 \times 0.7 \times 0.9 \times 0.5}{P(evidence)} \\
&= \frac{0.252}{P(evidence)}
\end{aligned}
$$

Assuming
Features
Independent

Orange:

$$P(Orange|Long, Sweet, Yellow) = 0$$

Other Fruit:

$$
\begin{aligned}
&P(Other|Long, Sweet, Yellow) \\
&= \frac{P(Long|Other) \cdot P(Sweet|Other) \cdot P(Yellow|Other) \cdot P(Other)}{P(Long) \cdot P(Sweet) \cdot P(Yellow)} \\
&= \frac{0.5 \times 0.75 \times 0.25 \times 0.2}{P(evidence)} \\
&= \frac{0.01875}{P(evidence)}
\end{aligned}
$$

Naive Bayes Formula

Banana:

$$P(Banana|Long, Sweet, Yellow)$$
$$= \frac{P(Long|Banana) \cdot P(Sweet|Banana) \cdot P(Yellow|Banana) \cdot P(Banana)}{P(Long) \cdot P(Sweet) \cdot P(Yellow)}$$
$$= \frac{0.8 \times 0.7 \times 0.9 \times 0.5}{P(evidence)}$$
$$= \frac{0.252}{P(evidence)}$$

Orange:

$$P(Orange|Long, Sweet, Yellow) = 0$$

Other Fruit:

$$P(Other|Long, Sweet, Yellow)$$
$$= \frac{P(Long|Other) \cdot P(Sweet|Other) \cdot P(Yellow|Other) \cdot P(Other)}{P(Long) \cdot P(Sweet) \cdot P(Yellow)}$$
$$= \frac{0.5 \times 0.75 \times 0.25 \times 0.2}{P(evidence)}$$
$$= \frac{0.01875}{P(evidence)}$$

P(evidence) same => irrelevant for argmax

Naive Bayes: argmax

Banana:

$$P(Banana|Long, Sweet, Yellow)$$
$$= \frac{P(Long|Banana) \cdot P(Sweet|Banana) \cdot P(Yellow|Banana) \cdot P(Banana)}{P(Long) \cdot P(Sweet) \cdot P(Yellow)}$$
$$= \frac{0.8 \times 0.7 \times 0.9 \times 0.5}{P(evidence)}$$
$$= \frac{0.252}{P(evidence)}$$

Winner

Orange:

$$P(Orange|Long, Sweet, Yellow) = 0$$

Other Fruit:

$$P(Other|Long, Sweet, Yellow)$$
$$= \frac{P(Long|Other) \cdot P(Sweet|Other) \cdot P(Yellow|Other) \cdot P(Other)}{P(Long) \cdot P(Sweet) \cdot P(Yellow)}$$
$$= \frac{0.5 \times 0.75 \times 0.25 \times 0.2}{P(evidence)}$$
$$= \frac{0.01875}{P(evidence)}$$

Many Variations of Naive Bayes

$$p(C_k \mid x_1, \dots, x_n) = \frac{1}{Z} p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)$$

Gaussian naive Bayes
Multinomial naive Bayes
Bernoulli naive Bayes
… many more

Naive Bayes

Pros:

- Doesn’t need lots of data
- Very fast
- Easy to interpret

Con:

- Variables rarely really independent

Frequently works well enough (even if variables not independent)!

PCA can even help!

Logistic Regression (Classifier)

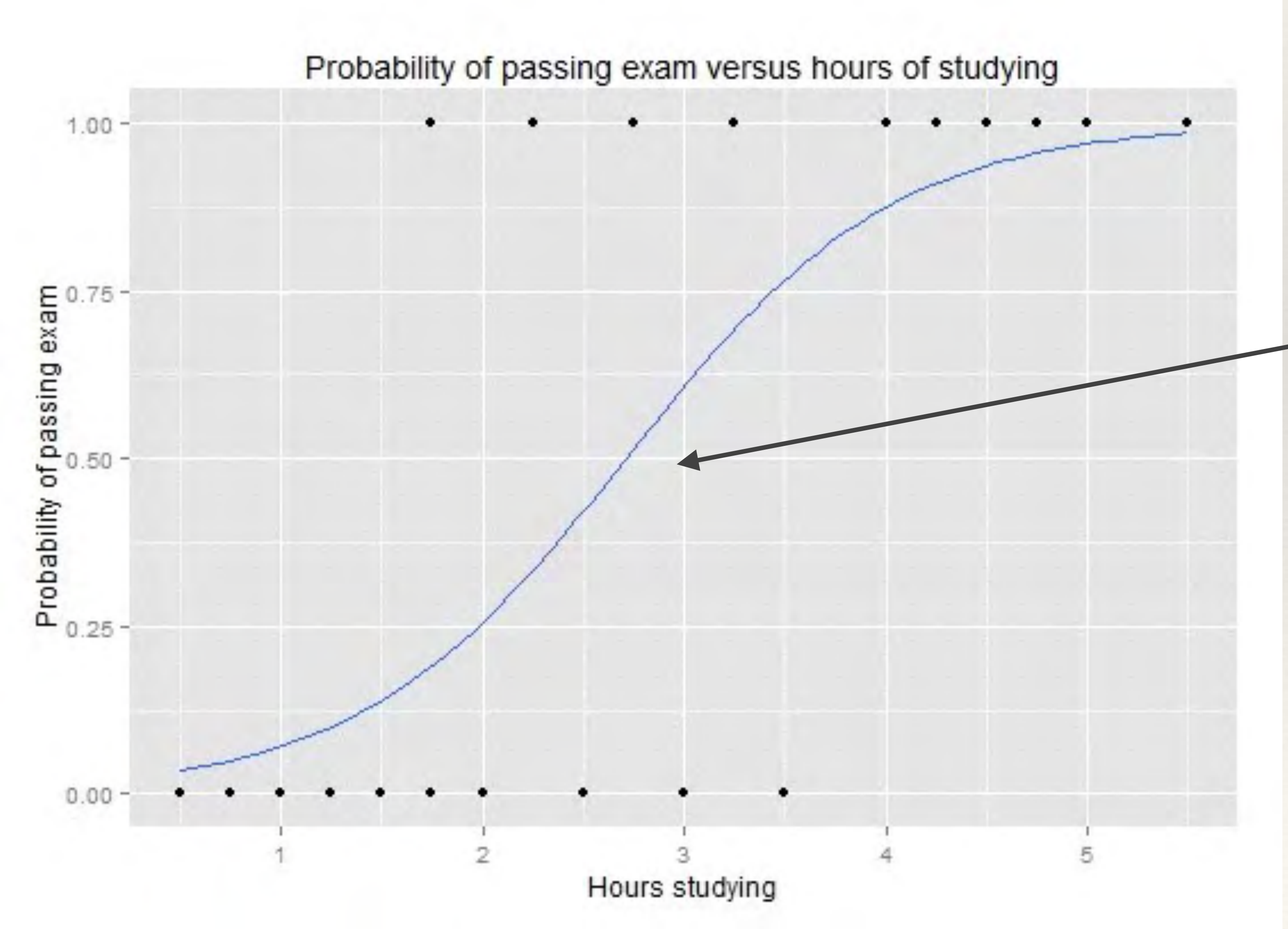
Study vs. Exam Success

A group of 20 students spend between 0 and 6 hours studying for an exam. How does the number of hours spent studying affect the probability that the student will pass the exam?

Exam data

Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1

0= fail, 1=Pass

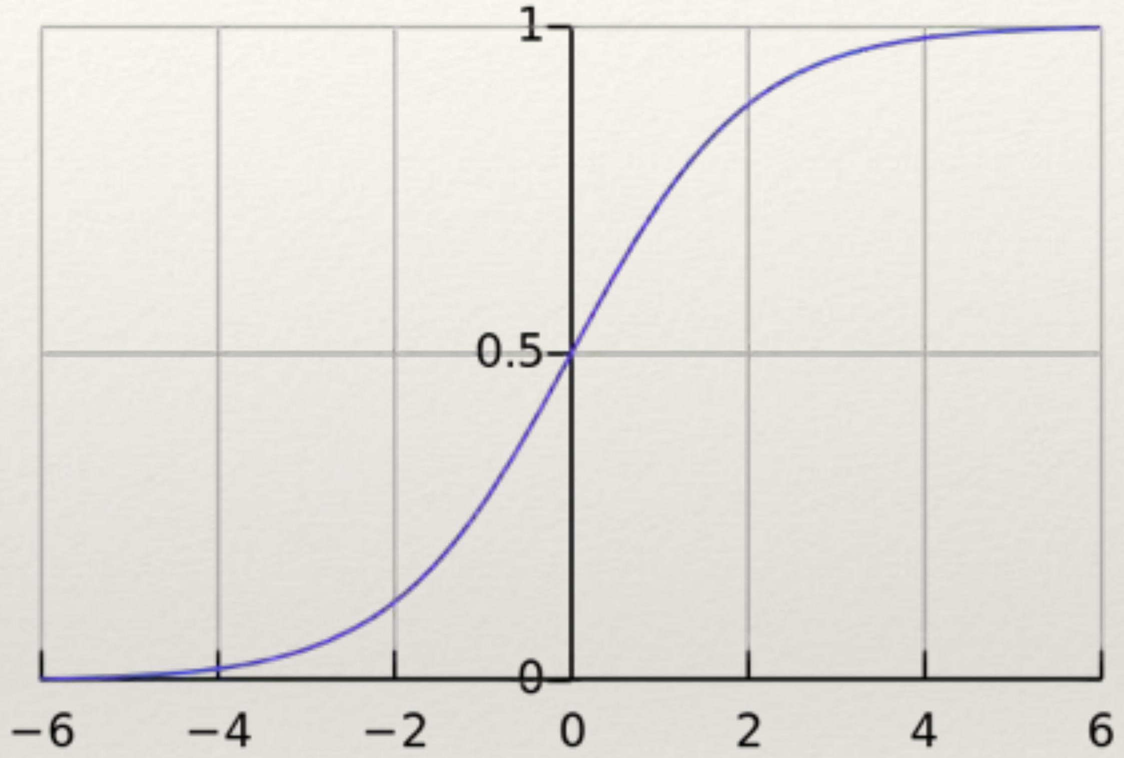


Want a function like this
to represent probability

Logistic Regression
Really Classifier (not regression)

Logistic Function

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



Want prob = 0.0 for x<<0
prob = 0.5 for x=0
prob = 1.0 for x>>0

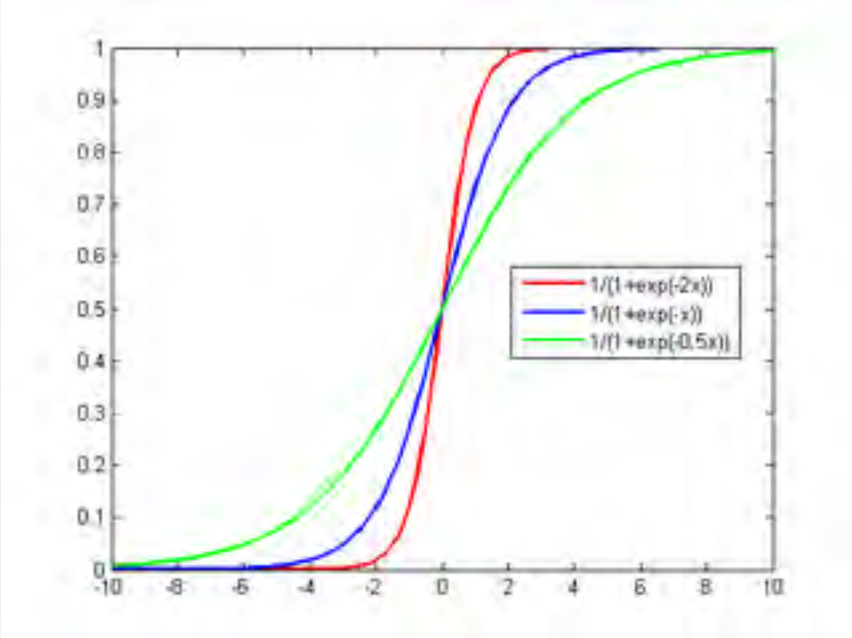
1-variable Logistic Function

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

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

Decision boundary position

Decision boundary slope



Derived Probability

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

```
Probability of passing exam =1/(1+exp(-(-4.0777+1.5046* Hours)))
```

From a curve fit

Hours of study	Probability of passing exam
1	0.07
2	0.26
3	0.61
4	0.87
5	0.97

2.5 hours ≃ coin flip chance of passing

5 hours <
greater than 97% chance of passing

logit

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

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

$$\ln\left(\frac{F(x)}{1 - F(x)}\right) = \beta_0 + \beta_1 x,$$

Linear

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

One Variable

Multi-Variable (linear classifier in a log space)