Review of mathematical foundations for Machine Learning

September 22, 2017

Random Variables

Slide credit: Amit Datta, Carnegie Mellon University

Coin tossing experiment

- **Experiment**
  - Toss a coin twice
- **Sample space:** Possible outcomes of an experiment
  - S = {HH, HT, TH, TT}
- **Event:** subset of possible outcomes
  - A = {HH}, B = {HT, TH}, C = {TT}

Slide credit: Amit Datta, Carnegie Mellon University

Random Variable (RV)

- A ***random variable X*** is a function from the sample space to a real number
- X : (represents number of heads)
  - {HH} → 2
  - {HT, TH} → 1
  - {TT} → 0

- Pr(Experiment yields no heads) = Pr({TT}) = Pr(X=0)

- Discrete RV: takes on finite number of values
- Continuous RV: takes an uncountable number of values

Slide credit: Amit Datta, Carnegie Mellon University

Discrete RV

- Probability Mass Function (PMF) p_X
  - Gives the probability that X will take on a *particular value*

- $p_X(a) = Pr(X=a)$

- $\sum_i p_X(a_i) = 1$

Slide credit: Amit Datta, Carnegie Mellon University

Continuous RV

- Probability Density Function (PDF) f_X
  - Non-negative function such that

$$Pr(a<=X<=b) = \int_a^b f_X(x)dx$$

- The integral from $-\infty$ to $+\infty$ is 1

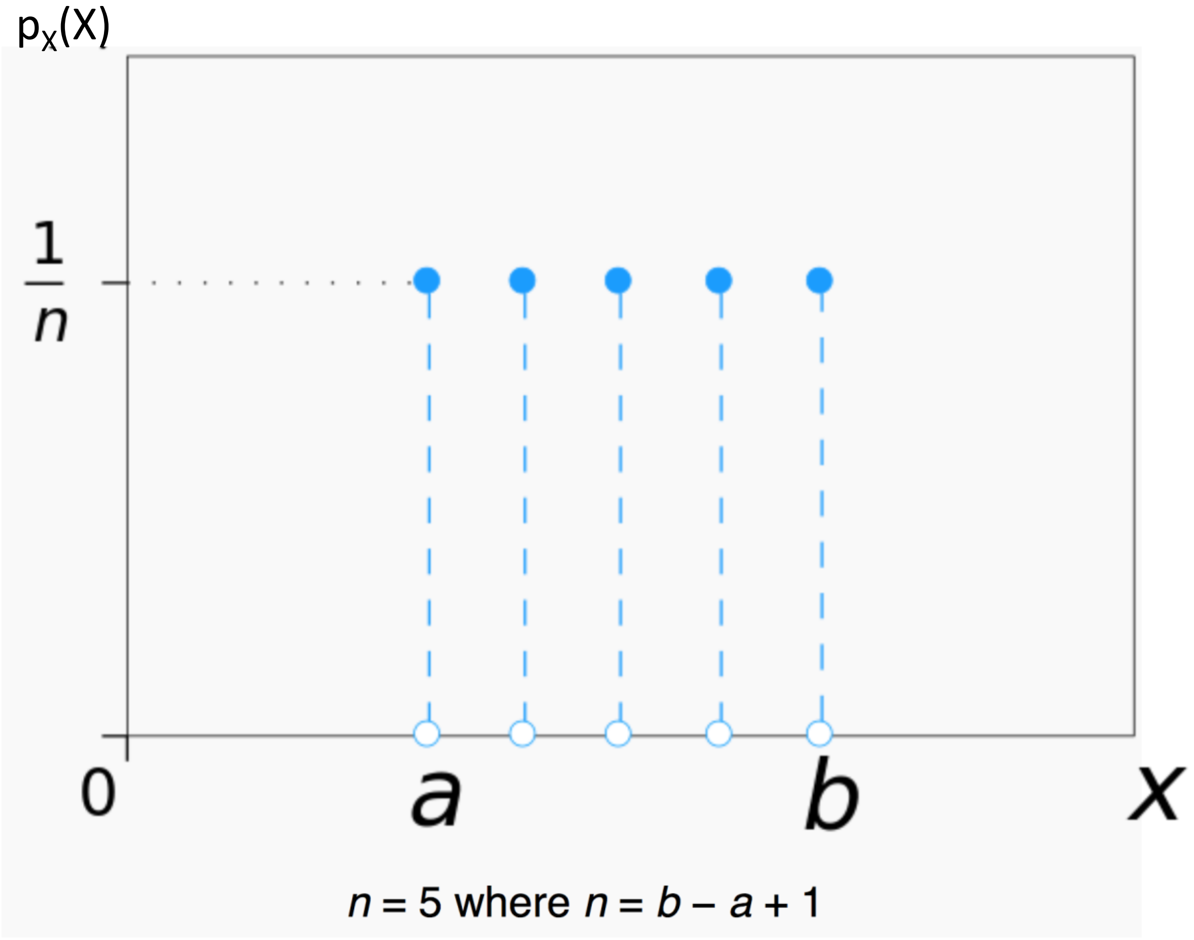
- $Pr(X=a) = 0$

Slide credit: Amit Datta, Carnegie Mellon University

Probability Distribution

- Assigns a probability to each event in the sample space

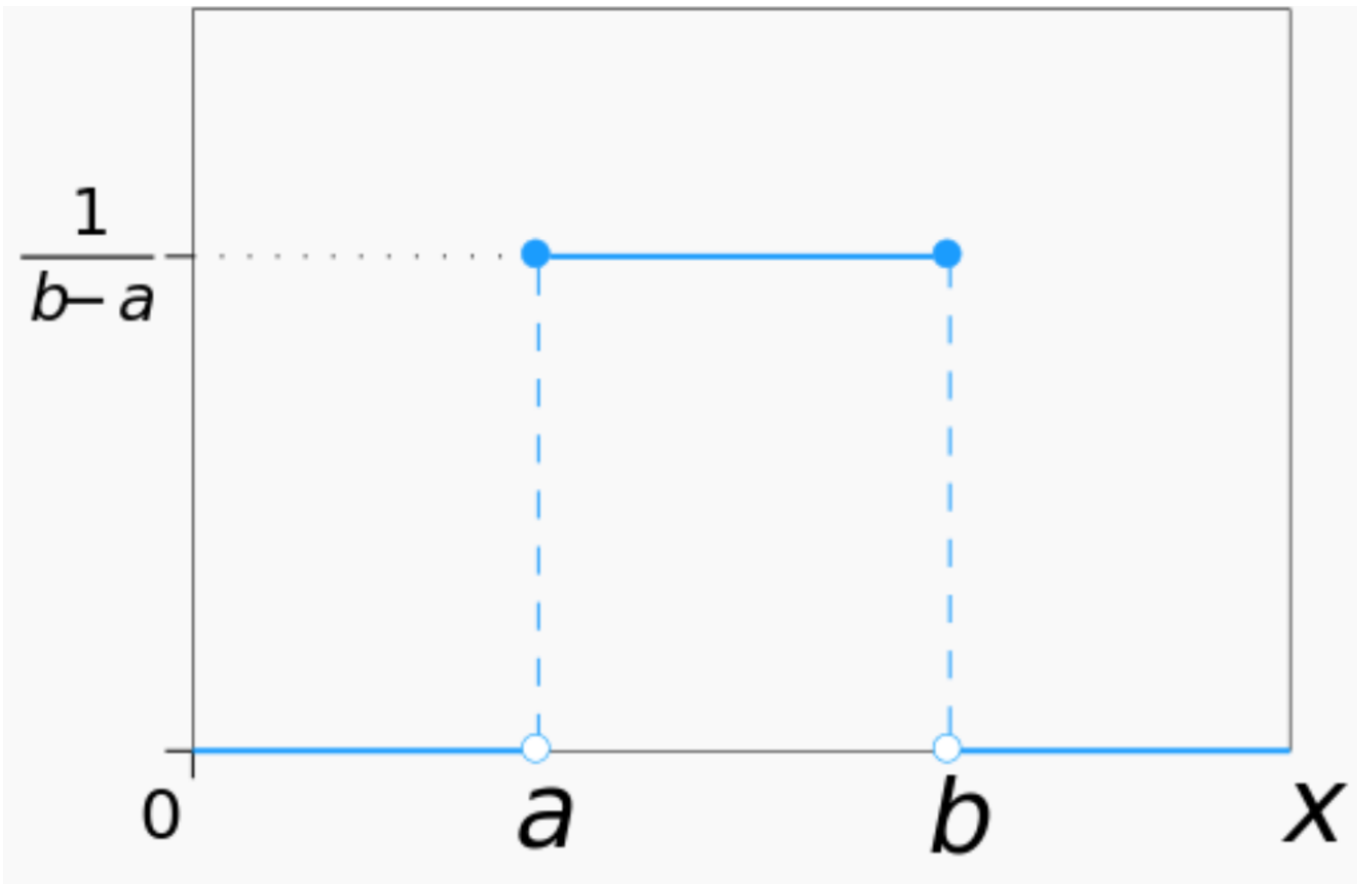
Discrete Uniform Distribution



Slide credit: Amit Datta, Carnegie Mellon University

Continuous Uniform Distribution

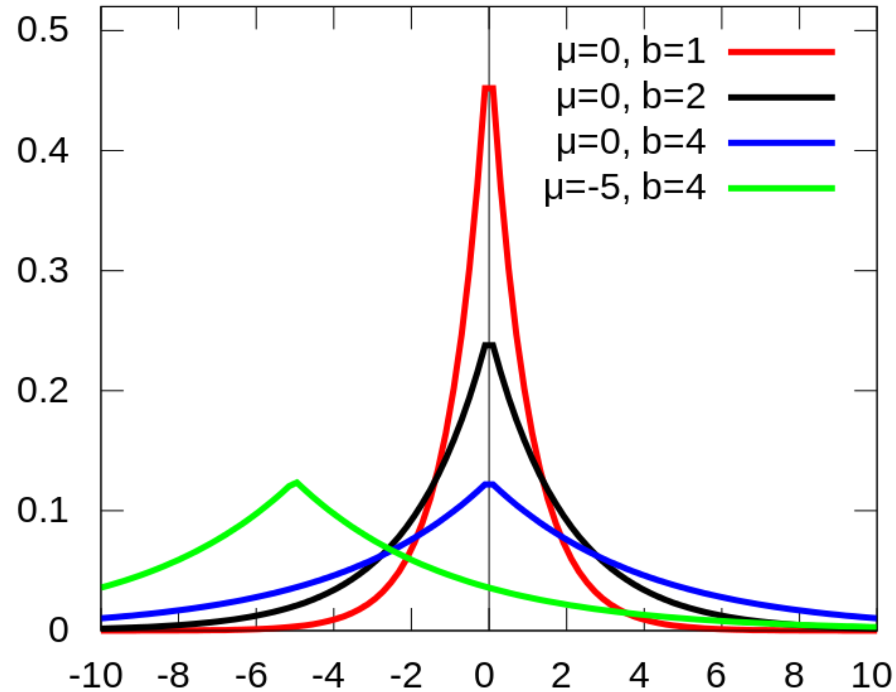
$f_X(X)$



Slide credit: Amit Datta, Carnegie Mellon University

Laplace Distribution

PDF = $\frac{1}{2b} exp(-\frac{|y-\mu|}{b})$



μ: location parameter
b: scale parameter

Similar to PDF for normal distribution

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\sigma^2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Slide credit: Amit Datta, Carnegie Mellon University

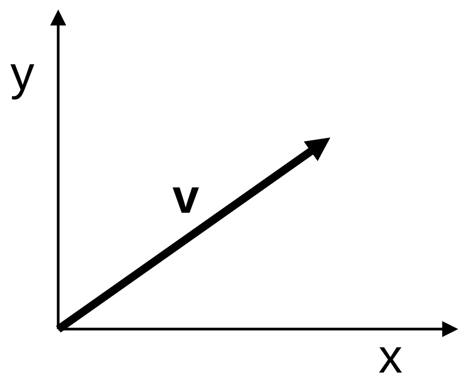
Linear Algebra Review

Slide credit: Miao Tang, University of Delaware

What is a Vector ?

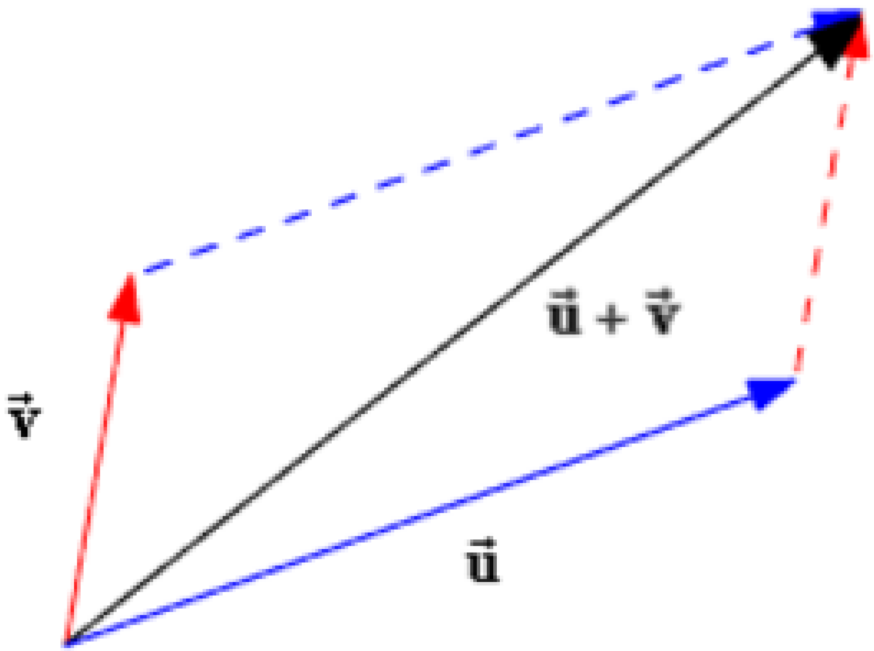
- Directed line segment in N-dimensions
  - Has "length" and "direction"
- $\mathbf{v} = [a\ \ b\ \ c]^T$
  - Geometry becomes linear algebra on vectors like **v**

$$\vec{v} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$



Slide credit: Miao Tang, University of Delaware

Vector Addition



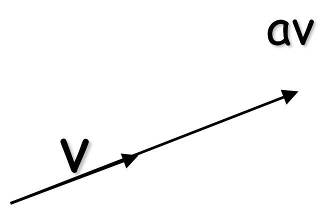
u = (u1, u2)
v = (v1, v2)

u + **v** = (u1+v1, u2+v2)

Slide credit: Miao Tang, University of Delaware

Scalar Product: *a*v

$$a\mathbf{v} = a(x_1, x_2) = (ax_1, ax_2)$$



Changes only the length ("scaling"), but keep ***direction fixed***.

Slide credit: Miao Tang, University of Delaware

Vectors: Dot Product

$$A \cdot B = A^T B \;=\begin{bmatrix} a & b & c \end{bmatrix}\begin{bmatrix} d \\ e \\ f \end{bmatrix} = ad + be + cf$$

Think of the dot product as a matrix multiplication

$$\|A\|^2 = A^T A = aa + bb + cc$$

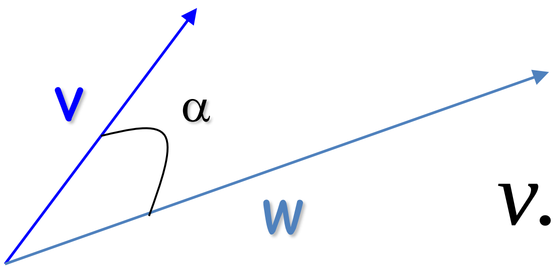
The magnitude is the dot product of a vector with itself

$$A \cdot B = \|A\| \; \|B\| \cos(\theta)$$

The dot product is also related to the angle between the two vectors

Slide credit: Miao Tang, University of Delaware

Inner (dot) Product: v.w or $w^T v$



$$v.w = (x_1, x_2).(y_1, y_2) = x_1 y_1 + x_2.y_2$$

The inner product is a **SCALAR**

$$v.w = (x_1, x_2).(y_1, y_2) = \| v \| \cdot \| w \| \cos\alpha$$

$$v.w = 0 \Leftrightarrow v \perp w$$

If vectors **v, w** are "columns", then dot product is $\mathbf{w}^T\mathbf{v}$

Slide credit: Miao Tang, University of Delaware

Matrix

- A matrix is a set of elements, organized into rows and columns

$$\begin{array}{cc} & \text{rows} \\ \text{columns} & \begin{bmatrix} a & b \\ c & d \end{bmatrix} \end{array}$$

Slide credit: Miao Tang, University of Delaware

Basic Matrix Operations

Addition, Subtraction, Multiplication: creating new matrices (or functions)

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} + \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} a+e & b+f \\ c+g & d+h \end{bmatrix}$$

Add elements

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} - \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} a-e & b-f \\ c-g & d-h \end{bmatrix}$$

Subtract elements

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{bmatrix}$$

Multiply each row by each column

Slide credit: Miao Tang, University of Delaware

Matrix Times Matrix

$$\mathbf{L} = \mathbf{M} \cdot \mathbf{N}$$

$$\begin{bmatrix} l_{11} & l_{12} & l_{13} \\ l_{21} & l_{22} & l_{23} \\ l_{31} & l_{32} & l_{33} \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{bmatrix} \cdot \begin{bmatrix} n_{11} & n_{12} & n_{13} \\ n_{21} & n_{22} & n_{23} \\ n_{31} & n_{32} & n_{33} \end{bmatrix}$$

$$l_{12} = m_{11}n_{12} + m_{12}n_{22} + m_{13}n_{32}$$

Slide credit: Miao Tang, University of Delaware

Multiplication

- Is AB = BA?

Slide credit: Miao Tang, University of Delaware

Multiplication

- Is AB = BA?

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} ae+bg & \dots \\ \dots & \dots \end{bmatrix} \quad \begin{bmatrix} e & f \\ g & h \end{bmatrix}\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} ea+fc & \dots \\ \dots & \dots \end{bmatrix}$$

- Matrix multiplication AB:
  - Apply transformation B first, then transform using A
- Not commutative

Slide credit: Miao Tang, University of Delaware

Matrix operating on vectors

- Matrix is like a <u>function</u> that <u>transforms the vectors on a plane</u>
- Matrix operating on a general point => transforms x- and y-components
- *System of linear equations*: matrix holds the coefficients

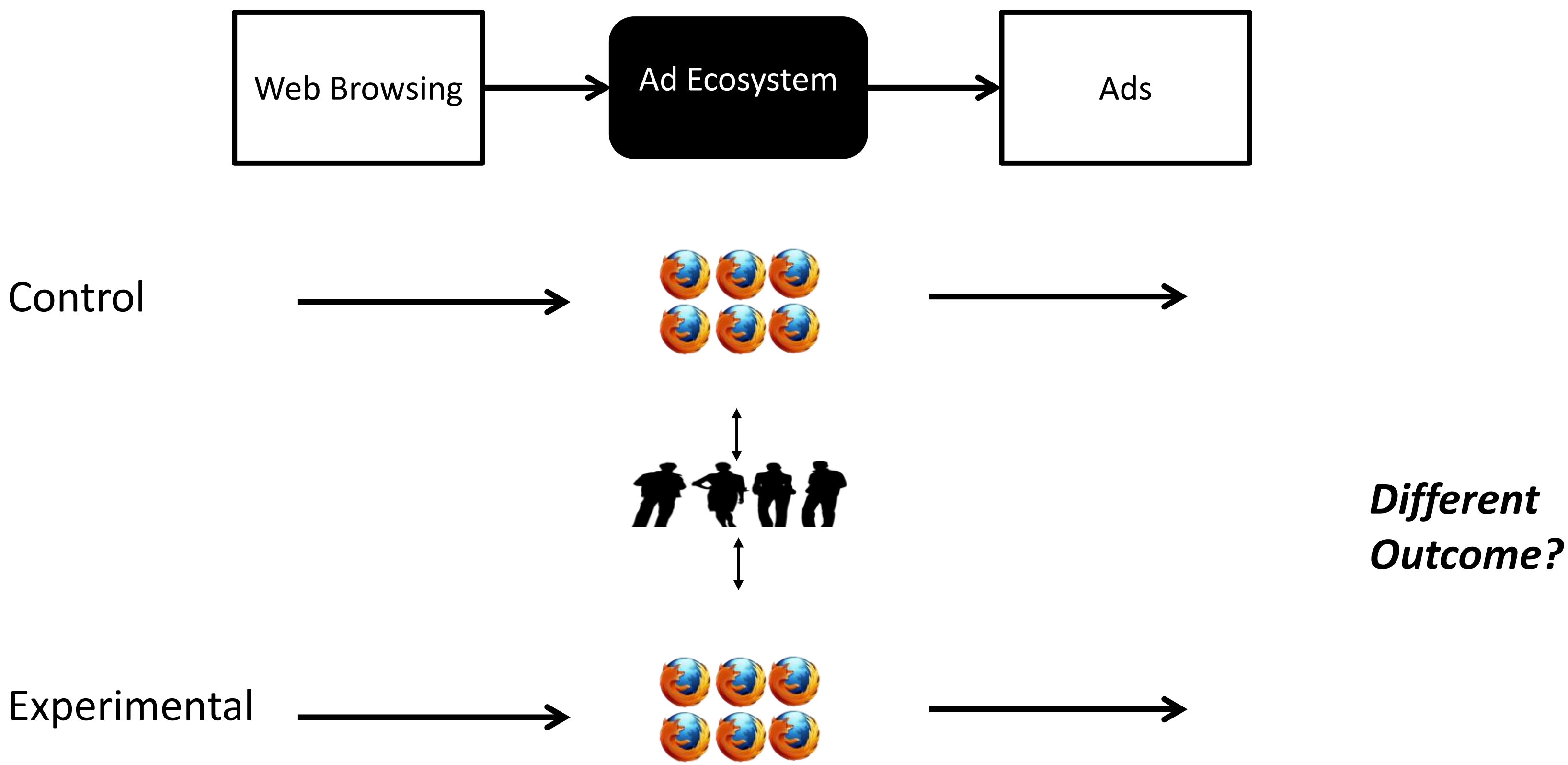
- $x' = ax + by$
- $y' = cx + dy$

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x' \\ y' \end{bmatrix}$$

Slide credit: Miao Tang, University of Delaware

Logistic Regression and AdFisher

AdFisher



Slide Credit: Amit Datta

Determining whether the difference in outcomes is statistically significant

- AdFisher splits the measurements collected into training and testing subsets.
- Examines the training subset to select a classifier that distinguishes between the measurements taken from each group.
- Uses ***logistic regression*** for classification.

Slide credit: J. Jeffry Howbert, University of Washington

Logistic Regression

- Technique for classification
  - Know as “regression” because a linear model is fit to the feature space
  - Probabilistic method of classification
- Models relationship between set of variables
  - Binary variables: Allergic to peanuts
  - Categorical: types of cancer such as brain cancer / leukemia / lymphoma / melanoma / etc
  - Continuous: weight / height

Slide credit: J. Jeffry Howbert, University of Washington

Ways to express probability

- Pr(E1) = p
- Pr(E2) = 1 – p = q

- Express Pr(E1) as:

	Notation	Range		
standard	p	0	0.5	1
odds	p/q	0	1	+∞
Log(odds)	log (p/q)	-∞	0	+∞

Slide credit: J. Jeffry Howbert, University of Washington

Log(odds)

- If neither event is favored:
  - log(odds) = log(0.5/0.5) = log(1) = 0
- If event E1 is favored over event E2:
  - Log(odds of E1) = log(p/q) = log(0.8/0.2) = log(4)
  - Log(odds of E2) = log(q/p) = log(0.8/0.2) = -log(4)

- Useful in domains where relative probabilities are small

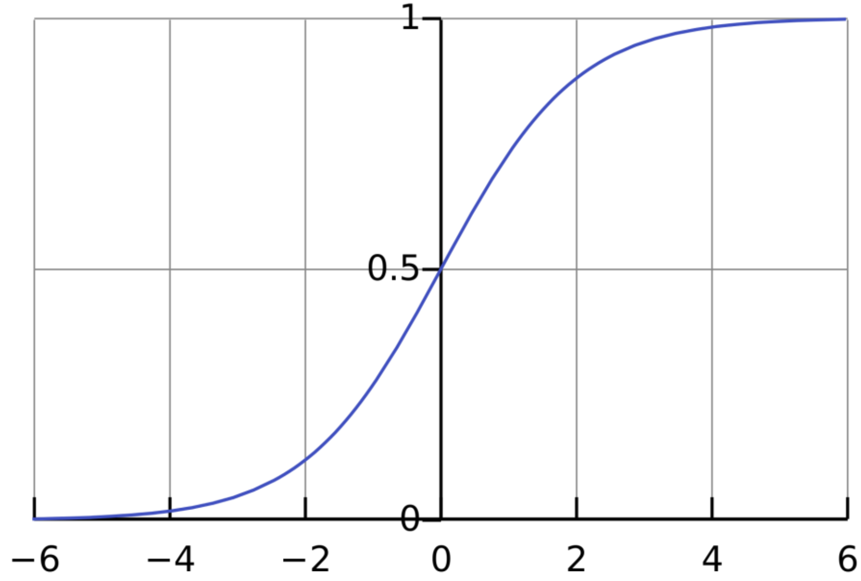
Slide credit: J. Jeffry Howbert, University of Washington

Log(odds) to logistic functions

$$z = \log\left(\frac{p}{1-p}\right)$$

$$\frac{p}{1-p} = e^{z}$$

$$p = \frac{e^{z}}{1+e^{z}} = \frac{1}{1+e^{-z}}$$



Slide credit: J. Jeffry Howbert, University of Washington

Using a logistic regression model

- Model a vector **B** in d-dim features space
- For a point **x** in feature space, project it onto **B** to convert it into a real number it into a real number z in the range in the range - ∞ to + ∞
- Map z to range [0,1] using logistic function

$$p = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$$

- Prediction from a logistic regression model can be viewed as a probability of class membership

Slide credit: J. Jeffry Howbert, University of Washington

Training a logistic regression model

- Optimize vector B
- Ensures the model gives the best possible reproduction of training set labels
- Usually done by numerical approximation of maximum likelihood

Slide credit: J. Jeffry Howbert, University of Washington