# Video 10.1
Dan Lee

# Weather Prediction

Sunny

Rainy

Should I bring an umbrella?

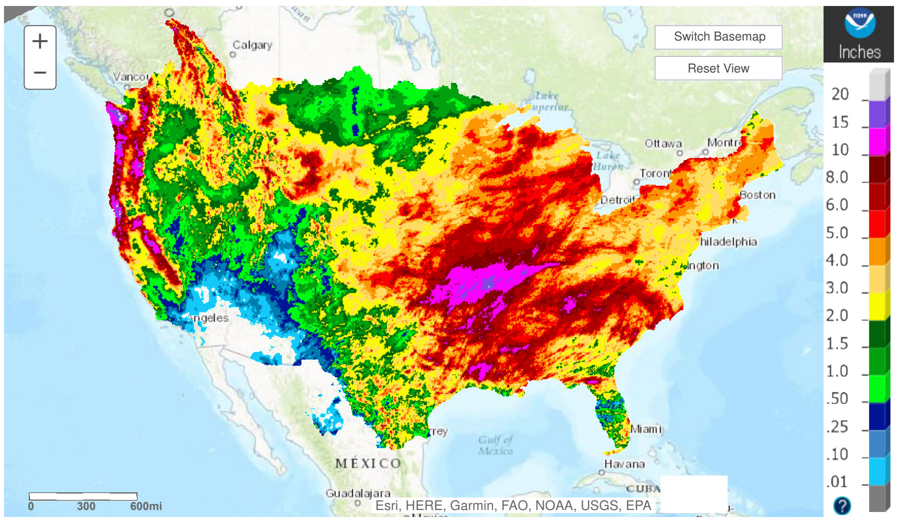
# Discrete Random Variable

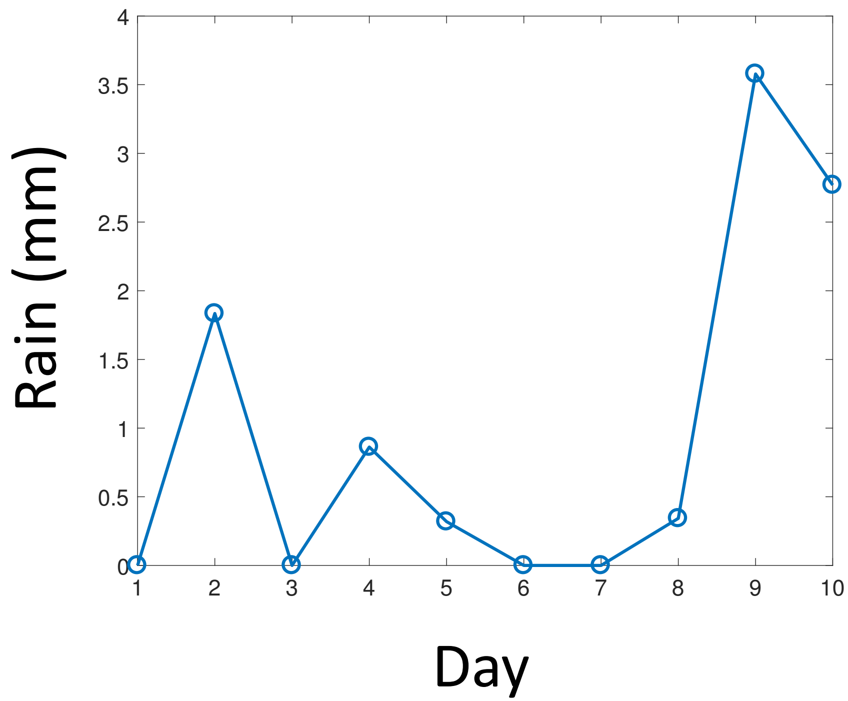
$$X \in \{0, 1\}$$

Two possible outcomes:

$$\text{sunny} \rightarrow 0$$
$$\text{rainy} \rightarrow 1$$

# Historical Data




Precipitation Records

Property of University of Pennsylvania, Dan Lee

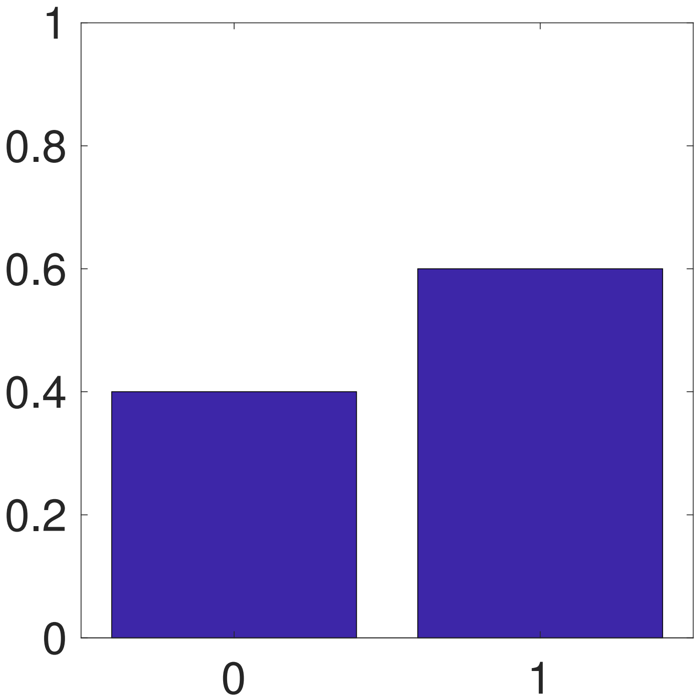
# Discrete Probabilities

$$Pr(X = 0) = \frac{4}{10} = 0.4$$

$$Pr(X = 1) = \frac{6}{10} = 0.6$$

$$\sum_{x} p(X = x) = 1.0$$

# Bernoulli Distribution




$$Pr(X = 0) = 1 - p$$

$$Pr(X = 1) = p$$

# Mean

$$E[X] = Pr(X = 1) \cdot 1 + Pr(X = 0) \cdot 0$$

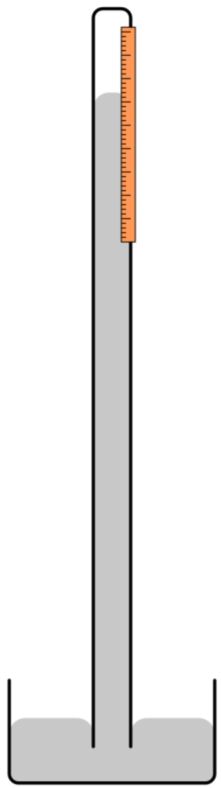
# Variance

$$E[X^2] = Pr(X = 1) \cdot 1^2 + Pr(X = 0) \cdot 0^2$$

$$Var[X] = E[X^2] - E[X]^2$$

$$Var[X] = p - p^2 = p(1 - p)$$

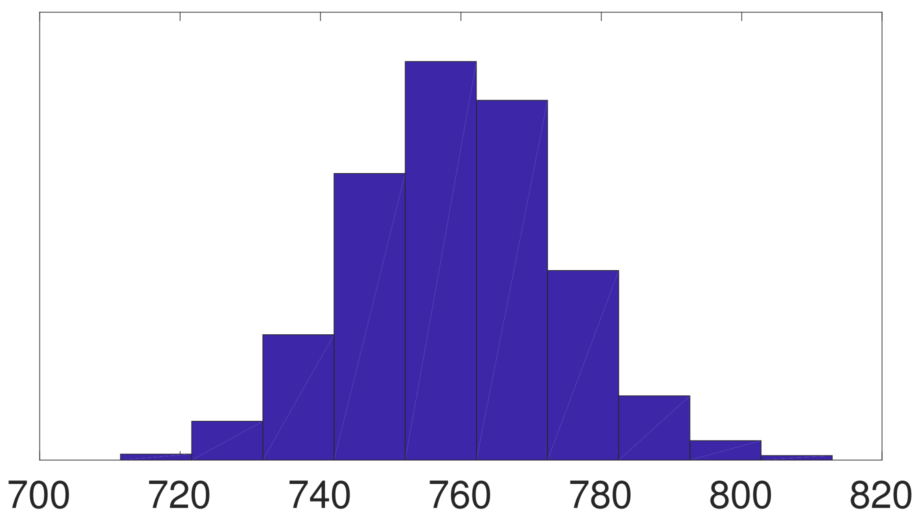
# Additional Random Variable



Barometer

Atmospheric pressure (mm)

$$Y \in \mathbb{R}$$



# Joint Distribution

$$Pr(X = x, Y = y)$$

$$\sum_{x,y} Pr(X = x, Y = y) = 1$$

$$\begin{array}{cc} & \text{Y} \\ \text{X} & \begin{bmatrix} 0.1 & 0.2 & 0.1 \\ 0.3 & 0.1 & 0.2 \end{bmatrix} \end{array}$$

# Marginal Distributions

$$Pr(X) = \sum_y Pr(X, Y = y)$$

$$Pr(Y) = \sum_x Pr(X = x, Y)$$

$$\begin{array}{cc} & Y \\ X & \begin{bmatrix} 0.1 & 0.2 & 0.1 \\ 0.3 & 0.1 & 0.2 \end{bmatrix} \end{array}$$

# Conditional Distributions

$$Pr(X|Y) = \frac{Pr(X,Y)}{Pr(Y)}$$

$$Pr(Y|X) = \frac{Pr(X,Y)}{Pr(X)}$$

$$\begin{array}{cc} & Y \\ X & \begin{bmatrix} 0.1 & 0.2 & 0.1 \\ 0.3 & 0.1 & 0.2 \end{bmatrix} \end{array}$$

# Bayes Rule

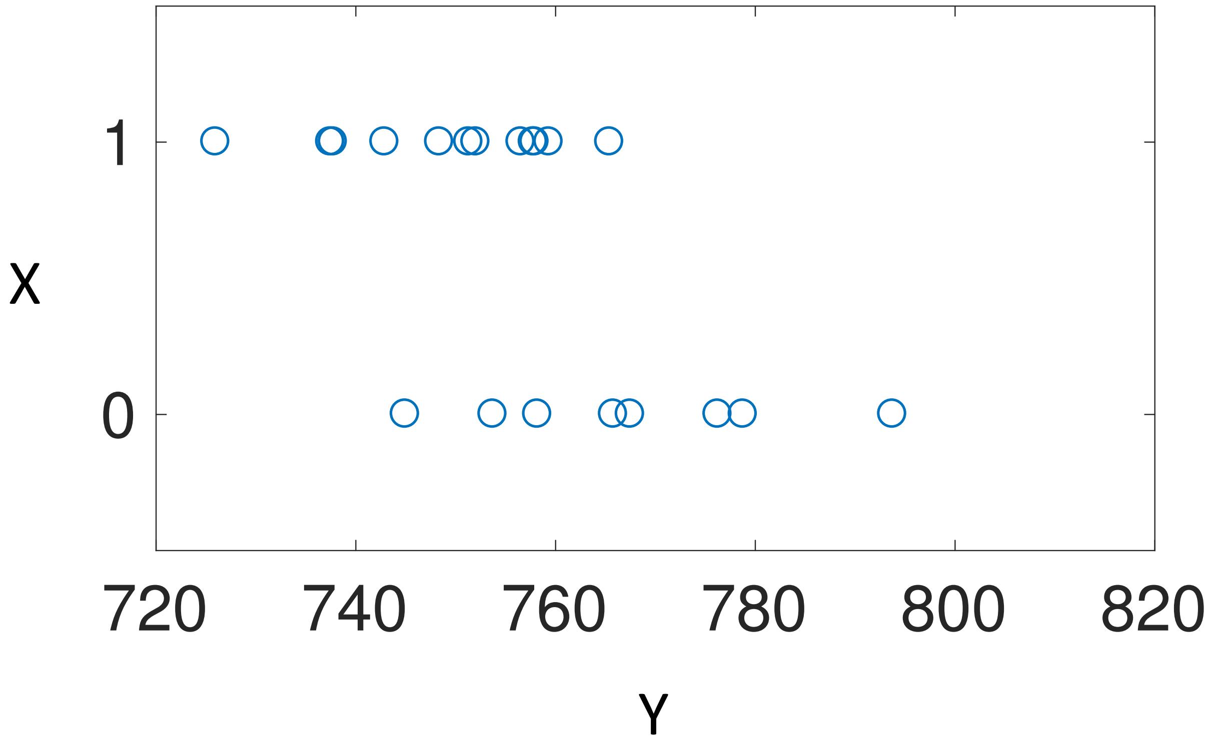
$$Pr(X|Y) = \frac{Pr(Y|X)Pr(X)}{Pr(Y)}$$

$$\begin{array}{cc} & Y \\ X & \begin{bmatrix} 0.1 & 0.2 & 0.1 \\ 0.3 & 0.1 & 0.2 \end{bmatrix} \end{array}$$

Penn Engineering
ONLINE LEARNING

# Video 10.2
# Dan Lee

# Correlations




$$Pr(X,Y) \neq Pr(X)Pr(Y)$$

# Logistic Model

$$Pr(X = 1|y) = \frac{\exp(\theta \cdot y)}{Z}$$

$$Pr(X = 0|y) = \frac{1}{Z}$$

$$Z = 1 + \exp(\theta \cdot y)$$

# Logistic Function

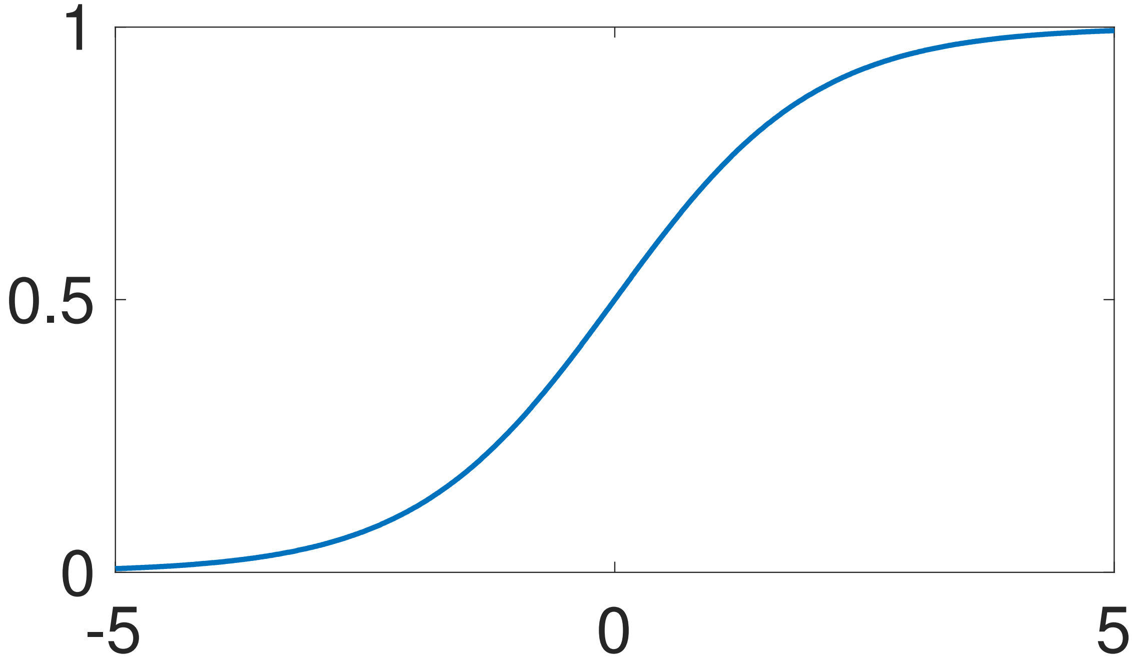
$$Pr(X = 1|y) = \frac{\exp(\theta \cdot y)}{1 + \exp(\theta \cdot y)}$$

$$Pr(X = 1|y) = \frac{1}{1 + \exp(-\theta \cdot y)}$$

$$\equiv \sigma\left(\theta \cdot y\right)$$

# Sigmoid

$$\sigma\left(\theta \cdot y\right)$$



# Maximum Likelihood

Given training data: $\{x^\mu, y^\mu\}$

$$Pr(X|Y = y^\mu) = \begin{cases} \sigma(\theta \cdot y^\mu), & X = 1 \\ 1 - \sigma(\theta \cdot y^\mu), & X = 0 \end{cases}$$

$$\max_\theta \prod_{x^\mu = 1} \sigma(\theta \cdot y^\mu) \prod_{x^\mu = 0} [1 - \sigma(\theta \cdot y^\mu)]$$

# Log Likelihood

$$\max_{\theta} \log \prod_{x^{\mu}=1} \sigma(\theta \cdot y^{\mu}) \prod_{x^{\mu}=0} [1 - \sigma(\theta \cdot y^{\mu})]$$

$$\max_{\theta} \sum_{x^{\mu}=1} \log \sigma(\theta \cdot y^{\mu}) + \sum_{x^{\mu}=0} \log[1 - \sigma(\theta \cdot y^{\mu})]$$

$$\max_{\theta} \sum_{\mu} x^{\mu} \log \sigma(\theta \cdot y^{\mu}) + (1 - x^{\mu}) \log[1 - \sigma(\theta \cdot y^{\mu})]$$

# Gradient descent

$$J(\theta) = -\sum_{\mu} x^{\mu} \log \sigma(\theta \cdot y^{\mu}) + (1 - x^{\mu}) \log[1 - \sigma(\theta \cdot y^{\mu})]$$

$$\min_{\theta} J(\theta)$$

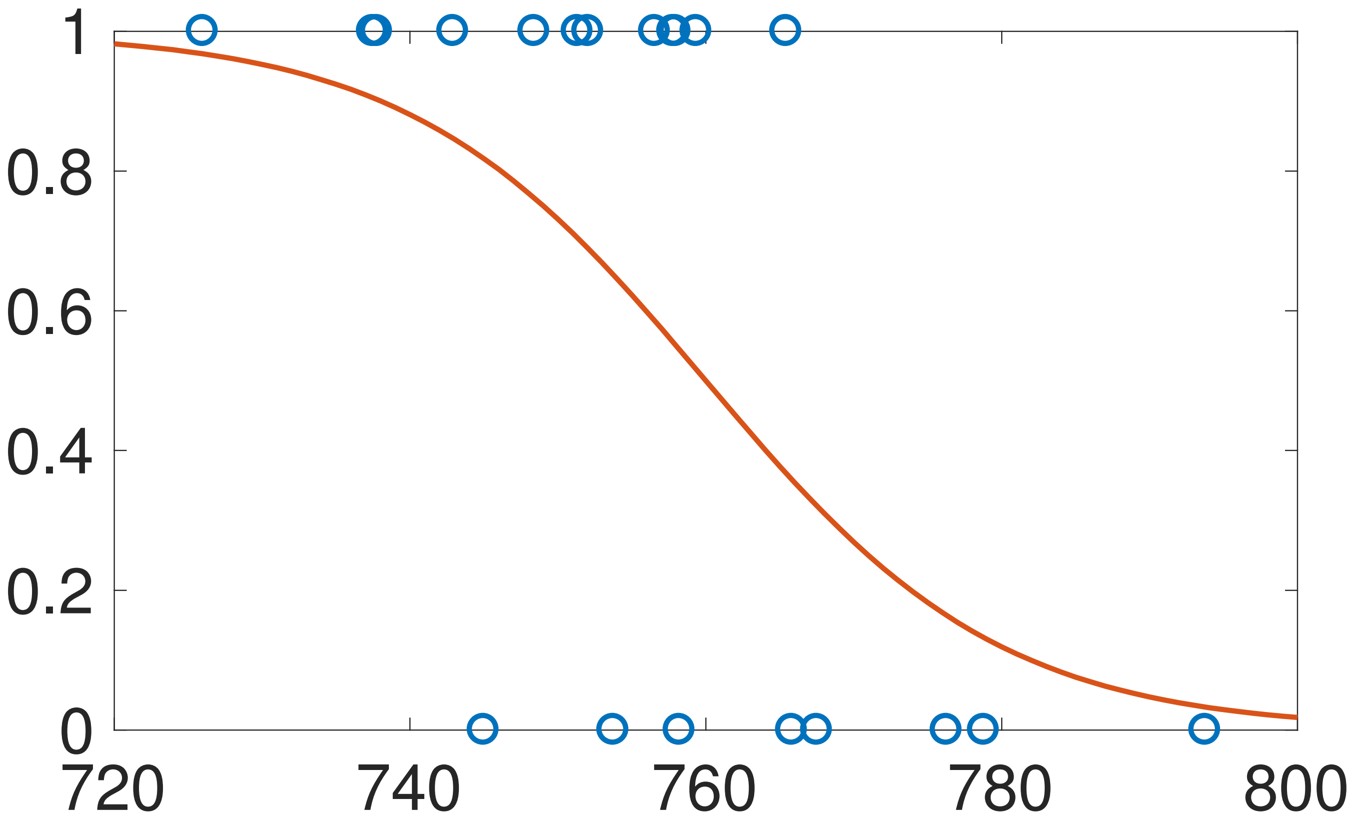
$$\theta' \leftarrow \theta - \eta \frac{\partial}{\partial \theta} J(\theta)$$

# Derivatives

$$\frac{\partial}{\partial\theta}\sigma(\theta\cdot y) = \frac{\partial}{\partial\theta}\frac{1}{1+e^{-\theta\cdot y}}$$

$$= \sigma(\theta\cdot y)[1-\sigma(\theta\cdot y)]y$$

# Results



Video 10.3

Dan Lee

# Logistic Regression

$$Pr(Y = 1|\vec{x}) = \frac{\exp(\vec{\theta} \cdot \vec{x})}{1 + \exp(\vec{\theta} \cdot \vec{x})}$$

Feature vector: $\vec{x}$

Parameters: $\vec{\theta}$

# Linear Bias

$$\exp\left(\vec{\theta} \cdot \vec{x} + b\right)$$

$$\vec{\theta} \leftarrow \begin{bmatrix} \vec{\theta} \\ b \end{bmatrix} \qquad \vec{x} \leftarrow \begin{bmatrix} \vec{x} \\ 1 \end{bmatrix}$$

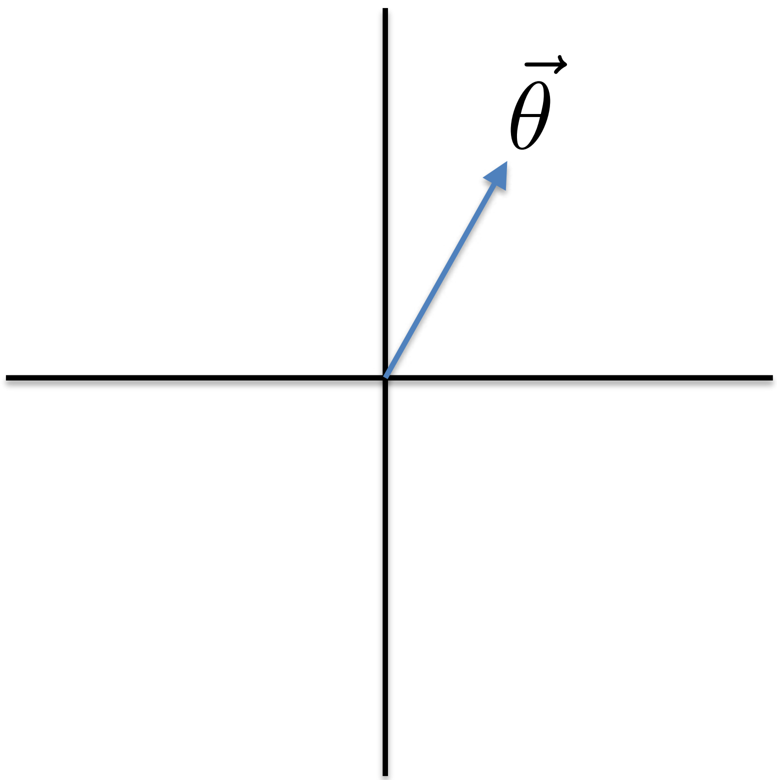
# Decision boundary

$$Pr(Y = 1|\vec{x}) = \frac{\exp(\vec{\theta} \cdot \vec{x})}{1 + \exp(\vec{\theta} \cdot \vec{x})} = \frac{1}{2}$$

$$\exp\left(\vec{\theta} \cdot \vec{x}\right) = 1$$

$$\vec{\theta} \cdot \vec{x} = 0$$

# Geometrical Interpretation



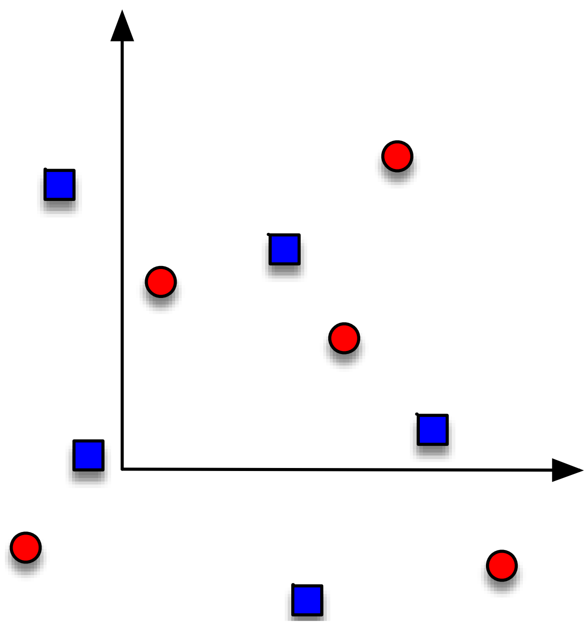

$$\vec{\theta} \cdot \vec{x} = 0$$

# Perceptron

$$f(\vec{x}) = \begin{cases} +1 : & \vec{\theta} \cdot \vec{x} > 0 \\ -1 : & \vec{\theta} \cdot \vec{x} < 0 \end{cases}$$

$$\sigma(\vec{\theta} \cdot \vec{x})$$

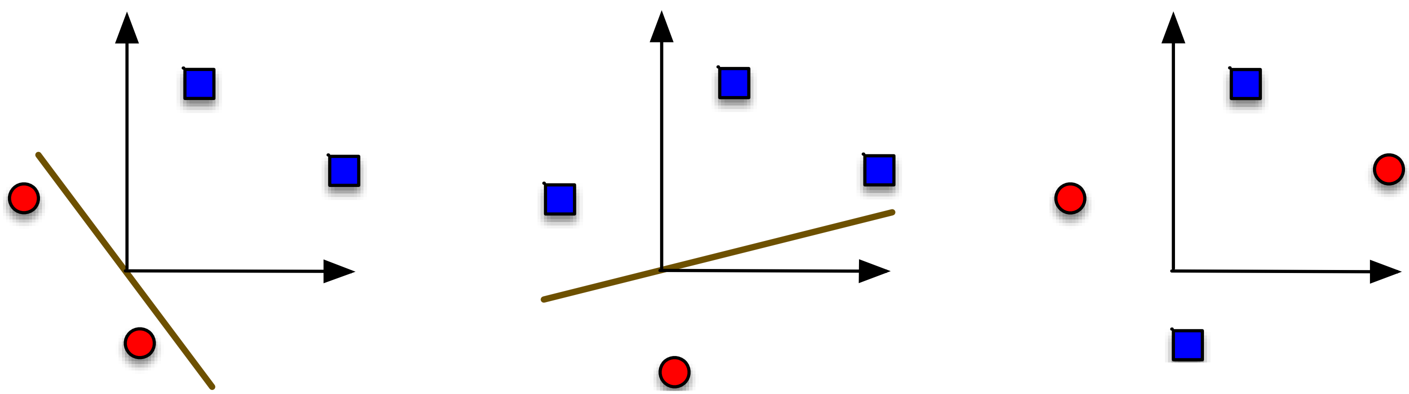
# Dichotomies



$$(\vec{x}^\mu \in \mathbb{R}^N, y^\mu = \pm 1)$$

$$\mu = \{1, \ldots, P\}$$

# Linear Separability



Video 10.3
Dan Lee

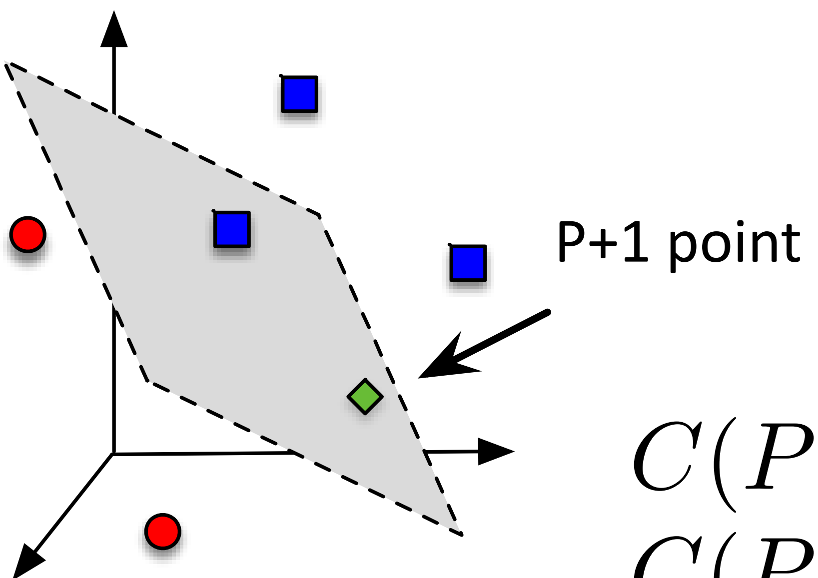
# Cover Counting Theorem

Number of linearly separable dichotomies
of P points in N dimensions:

$$C(P,N) = 2\sum_{k=0}^{N-1} \binom{P-1}{k} \leq 2^P$$

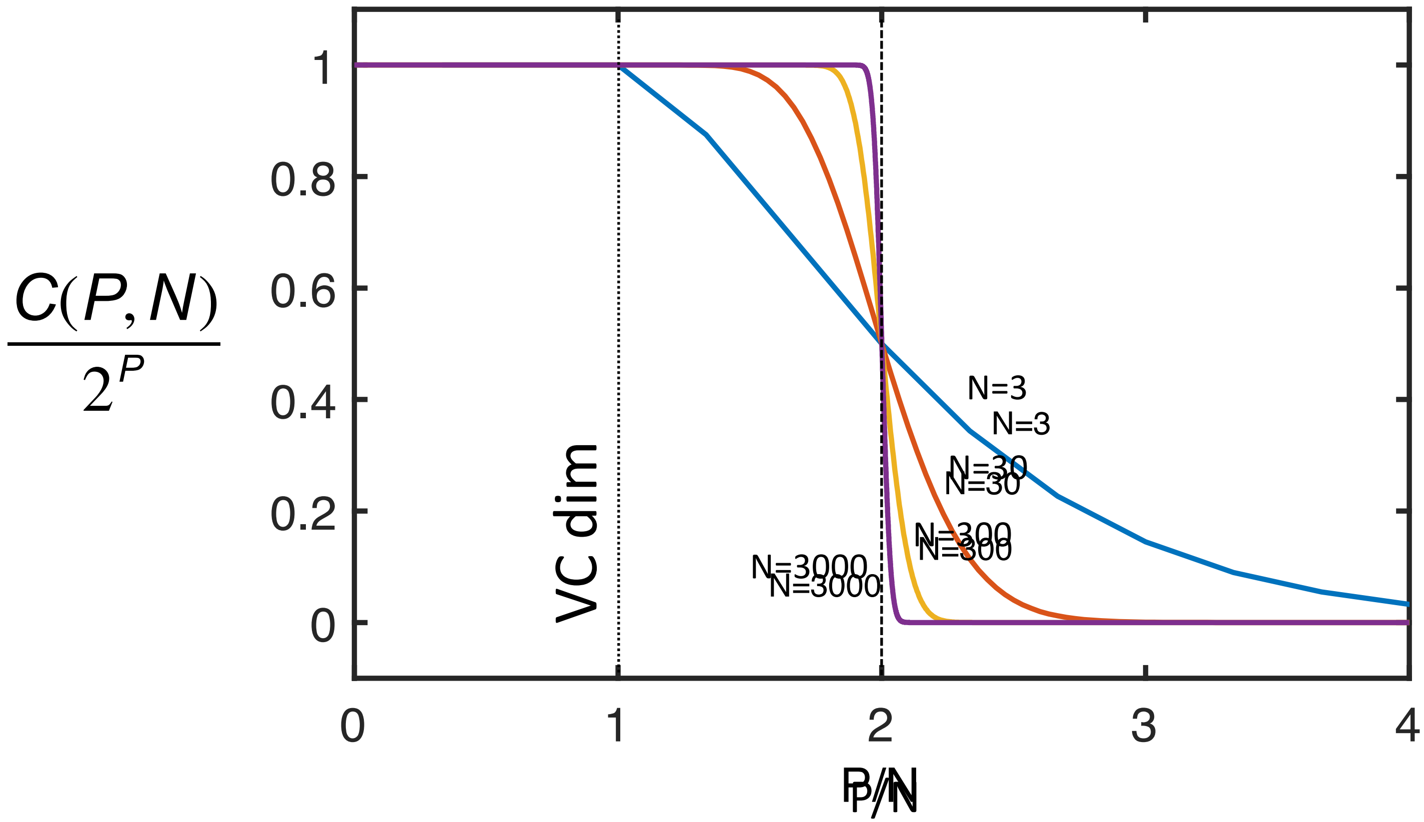
$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

# Mathematical Induction




$$C(P+1, N) =$$
$$C(P, N) + C(P, N-1)$$

# Capacity



# Learning Algorithm

Training Data: $(\vec{x}^\mu \in \mathbb{R}^N, y^\mu = \pm 1)$

Initialize: $\vec{\theta}$

Check: $y^\mu \left( \vec{\theta} \cdot \vec{x}^\mu \right) > 0$

$$\vec{\theta} \leftarrow \vec{\theta} + \eta \, y^\mu \vec{x}^\mu$$

# Guaranteed convergence

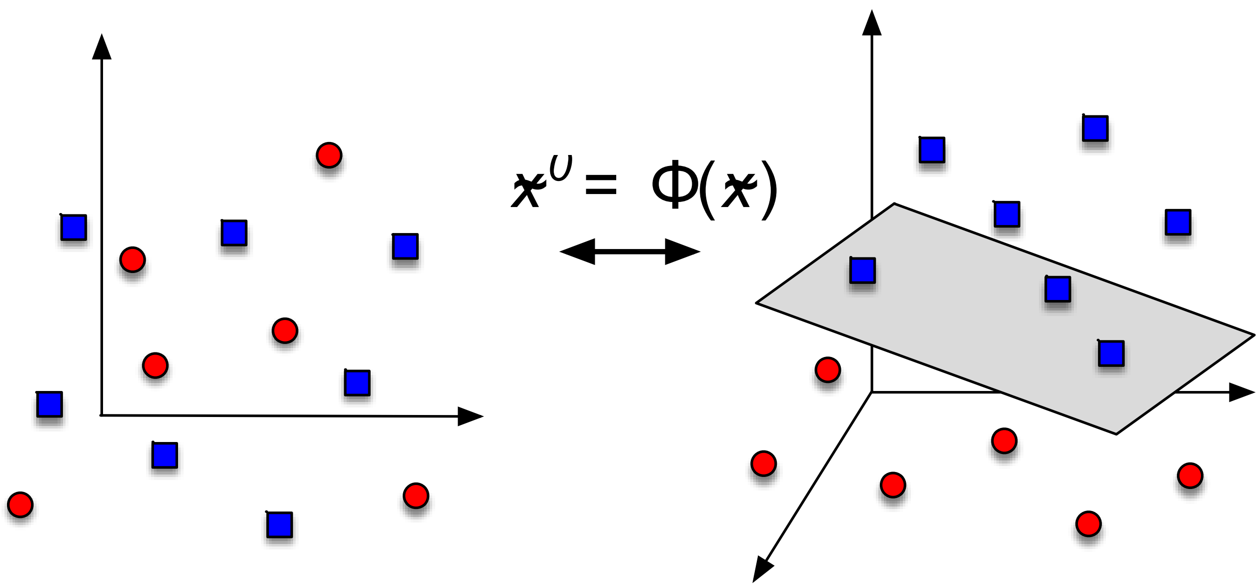
$$\vec{\theta} \leftarrow \vec{\theta} + \eta\, y^{\mu} \vec{x}^{\mu}$$

# Cost function

$$J = \sum_{\mu} \max\left[-y^{\mu}\sigma(\vec{\theta}\cdot\vec{x}^{\mu}), 0\right]$$

$$\vec{\theta} \leftarrow \vec{\theta} - \eta\frac{\partial J}{\partial\vec{\theta}}$$

# Support Vector Machines



# Multilayer Perceptrons