Machine Learning
4771

Instructor: Itsik Pe'er

Reminder:
Fitting = Maximizing Likelihood

◆ Regression to fit the Poisson rate

$$\hat{Y} = \theta_1 X + \theta_0$$

#Celebrity Deaths (y-axis: 0, 5, 10, 15, 20, 25)

x-axis: 2000, 2005, 2010, 2015
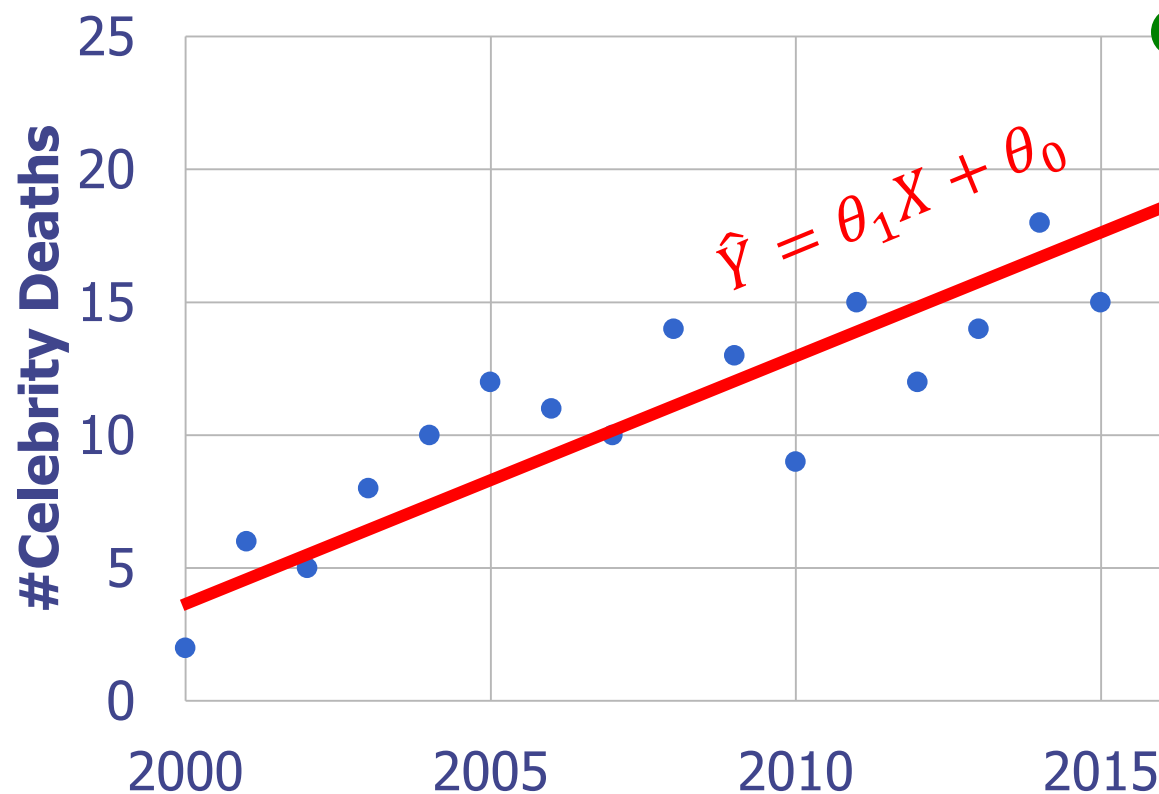
Probability Review

- Definitions
- Distributions
- Moments
- Theorems

Definition: Sample Space

◆ *Sample space* : Ω all possible outcomes

Definition: Events

◈ *Sample space* : Ω all possible outcomes
   Examples: deaths '17, weather Thu, 2 dice

◈ *Event* : subset of outcomes

Definition: Probability

◈ *Sample space* : Ω all possible outcomes
Examples: deaths '17, weather Thu, 2 dice

◈ *Event* : subset of outcomes
Examples: I die Feb, snow Thu, sum dice<10

◈ *Probability function*: $\mathrm{Prob} : \Omega \rightarrow [0,1]$,
additive, $\mathrm{Prob}(\Omega) = 1$

Definition: Random Variables

- *Sample space* : Ω all possible outcomes
  Examples: deaths '17, weather Thu, 2 dice

- *Event* : subset of outcomes
  Examples: I die Feb, snow Thu, sum dice<10

- *Probability function*: $\mathrm{Prob}: \Omega \rightarrow [0,1]$,
  additive, $\mathrm{Prob}(\Omega) = 1$

  Examples: forecast, $p([i,j]) = \frac{1}{36}$

- *Random variable*: $X: \Omega \rightarrow \boldsymbol{R}$ or $\boldsymbol{R}^D$

  Example: #deaths, percip.[mm], sum dice

Definition: Independence

◈ Events $A \perp B$: $\text{Prob}\,(A \cap B) = \text{Prob}\,(A)\text{Prob}\,(B)$

◈ Random variables: $X \perp Y$ if $\forall A, B: A(X) \perp B(Y)$

Definition: Independence

◈ Events $A \perp B$: $\text{Prob}\,(A \cap B) = \text{Prob}\,(A)\text{Prob}\,(B)$
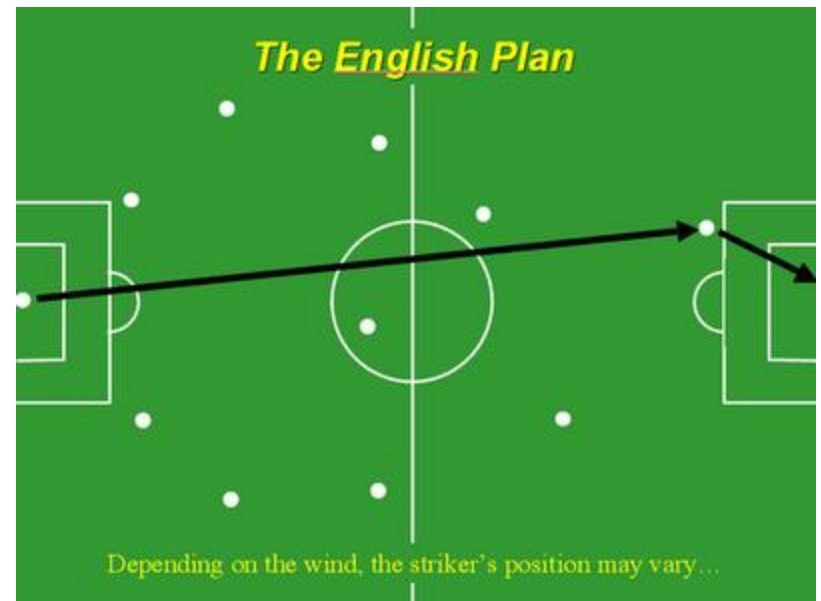   Examples: coin flips, winning MI/PA

◈ Random variables $X \perp Y$ if $\forall A, B: A(X) \perp B(Y)$
   Examples: dice, results MI/PA, height/GPA

Definition: Conditional Probability

◆ $\mathrm{Prob}\,(A \cap B) = \dfrac{\mathrm{Prob}\,(A \cap B)}{\mathrm{Prob}\,(B)}$

Examples: low pass possession & goal

The English Plan

Depending on the wind, the striker's position may vary...

Distributions

◆ Discrete

◆ Continuous

Distributions

◈ Discrete

   ■ Bernoulli, Binomial, Multinomial, Poisson Geometric

◈ Continuous

Bernoulli Distribution

- Bernoulli

	x=0	x=1
	0.95	0.05

Bernoulli Distribution

- Bernoulli(α): binary (coin flip) probability, just 1x2 table

$p(x) = \alpha^x(1-\alpha)^{1-x} \qquad \alpha \in [0,1], x \in \{0,1\}$

x=0	x=1
0.95	α=0.05

- Multidimensional Bernoulli:

Bernoulli Distribution

- Bernoulli(α): binary (coin flip) probability, just 1x2 table

$$p(x) = \alpha^x (1-\alpha)^{1-x} \qquad \alpha \in [0,1], x \in \{0,1\}$$

x=0	x=1
0.95	α=0.05

- Multidimensional Bernoulli: multiple binary events

$p(x_1, x_2)$

	x_2=0	x_2=1
x_1=0	0.4	0.1
x_1=1	0.3	0.2

Bernoulli Distribution

- Bernoulli(α): binary (coin flip) probability, just 1x2 table

$$p(x) = \alpha^x (1-\alpha)^{1-x} \qquad \alpha \in [0,1], x \in \{0,1\}$$

x=0	x=1
0.95	α=0.05

- Multidimensional Bernoulli: multiple binary events

$p(x_1, x_2)$

	x_2=0	x_2=1
x_1=0	0.4	0.1
x_1=1	0.3	0.2

$p(x_1, x_2, x_3)$
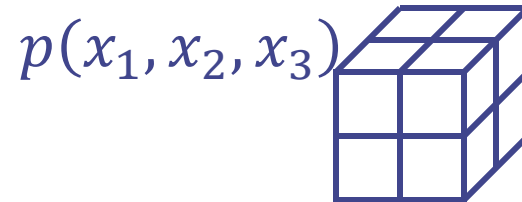
Binomial Distribution

- Bernoulli(α): recall binary (coin flip) probability, 1x2 table

$p(x) = \alpha^x(1-\alpha)^{1-x} \qquad \alpha \in [0,1], x \in \{0,1\}$

	x=0	x=1
	$1-\alpha$	α

- Binomial(n, α): sum of n identical, independent coin flips

$p(x) = \binom{n}{x} \alpha^x(1-\alpha)^{1-x} \qquad \alpha \in [0,1], x \in \{0,1\}$

Poisson Distribution

- Bernoulli(α): recall binary (coin flip) probability, 1x2 table

$p(x) = \alpha^x(1-\alpha)^{1-x}$ $\alpha \in [0,1], x \in \{0,1\}$

	x=0	x=1
	$1-\alpha$	α

- Binomial(n,α): sum of n identical, independent coin flips

$$p(x) = \binom{n}{x}\alpha^x(1-\alpha)^{1-x}   \alpha \in [0,1], x \in \{0,\ldots,n\}$$

- Poisson(λ): $\lim_{n\to\infty} Binomial\left(n,\frac{\lambda}{n}\right)$ sum of many rare iid coins

$p(x) = \frac{\lambda^x e^{-\lambda}}{x!}$ $\lambda \in \mathbf{R}^+, x \in \mathbf{Z}^+$

Geometric Distribution

- Bernoulli(α): recall binary (coin flip) probability, 1x2 table

$p(x) = \alpha^x (1-\alpha)^{1-x}$ $\alpha \in [0,1], x \in \{0,1\}$

	x=0	x=1
	$1-\alpha$	α

- Binomial(n, α): sum of n identical, independent coin flips

$$p(x) = \binom{n}{x} \alpha^x (1-\alpha)^{1-x} \qquad \alpha \in [0,1], x \in \{0, \dots, n\}$$

- Poisson(λ): $\lim_{n \to \infty} Binomial\left(n, \frac{\lambda}{n}\right)$ sum of many rare iid coins

$p(x) = \dfrac{\lambda^x e^{-\lambda}}{x!}$ $\lambda \in \boldsymbol{R}^+, x \in \boldsymbol{Z}^+$

- Geometric(α): number of iid flips till first success

$p(x) = (1-\alpha)^{x-1}$ $\alpha \in [0,1], x \in \boldsymbol{Z}^+$

Multinomial Distribution

- Multinomial($\vec{\alpha}$) : beyond binary multi-category event (dice)

	1	2	3	4	5	6
	$\vec{\alpha}(1)$	$\vec{\alpha}(2)$	$\vec{\alpha}(3)$	$\vec{\alpha}(4)$	$\vec{\alpha}(5)$	$\vec{\alpha}(6)$

$$p(x) = \prod_{m=1}^{M} \vec{\alpha}(m)^{\vec{x}(m)} \qquad \sum_{m} \vec{\alpha}(m) = 1$$

$\vec{x}(1)$	$\vec{x}(2)$	$\vec{x}(3)$	$\vec{x}(4)$	$\vec{x}(5)$	$\vec{x}(6)$

Expectation

◆ $E(X) = \sum_x x p(X = x)$

What is your best guess for X?
Example:

E(Bernoulli(α))

E(dice)

Expectation

- $E(X) = \sum_x x p(X = x)$
- Important thms:
  - Linearity: $E(X + Y) = E(X) + E(Y)$
    $$E(aX) = aE(X)$$
  - Law of large numbers:
    $$\{X_1, \dots\} \text{ i. i. d.}, \text{ then } S_n = \frac{\sum_{i=1}^{n} x_i}{n} \xrightarrow[n \to \infty]{} E(X)$$

Variance

◈ $Var(X) = E\left(\left(X - E(X)\right)^2\right)$

How wide is X 's distribution around $E(X)$?

Variance

◆ $Var(X) = E\left((X - E(X))^2\right)$

◆ Quadratic scaling: $Var(aX) = a^2 Var(X)$

◆ Standard deviation: $Std(X) = \sqrt{Var(X)}$

◆ Covariance

$Cov(X, Y) = E\left((X - E(X))(Y - E(Y))\right)$

Continuous Probability Models

- Probabilities can have both discrete & continuous variables

- We will discuss:
  1) discrete probability tables

x=T	x=H
0.4	0.6

x=1	x=2	x=3	x=4	x=5	x=6
0.1	0.1	0.1	0.1	0.1	0.5

  2) continuous probability distributions

$p(x) =$ probability density function, not probability mass function

$cdf(x) = \int_{-\infty}^{x} p(t)dt$ gives actual probabilities
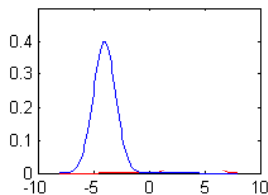
$$\int_{-\infty}^{\infty} p(x)dx = 1$$
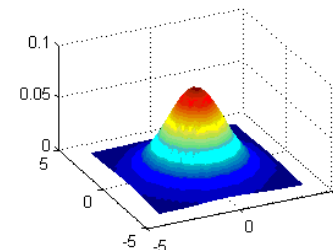
Continuous Distributions: Uniform

◆ Uniform(a, b) :

$$p(x) = \frac{1}{b-a} \qquad a < b \in \boldsymbol{R}, x \in [a, b]$$

Exponential Distribution

◆ Exponential(λ) : Time till next Poisson

arrival, $\lim\limits_{n \to \infty} \dfrac{Geometric\left(\frac{\lambda}{n}\right)}{n}$

$p(x) = \lambda e^{-\lambda} \qquad \lambda \in \boldsymbol{R}^+ , x \in \boldsymbol{R}^+$

Std. Gaussian (Normal) Distribution

- Bell shape curve

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$
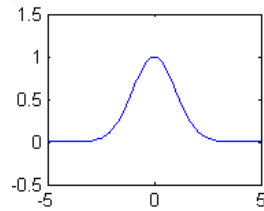
Central Limit Theorem

- $\{X_1, \dots\}$ i.i.d., $S_n = \dfrac{\sum_{i=1}^n x_i}{n}$

  $E(X) = \mu, Var(X) = \sigma^2$

  then $\sqrt{n}\dfrac{S_n - \mu}{\sigma} \xrightarrow[n \to \infty]{}$ std. normal

Gaussian Distribution

- 1-dimensional Gaussian with mean parameter μ translates Gaussian left & right

$$p(x|\mu) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}(x-\mu)^2\right)$$
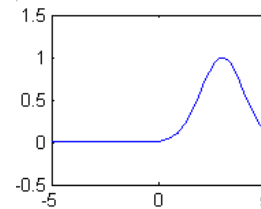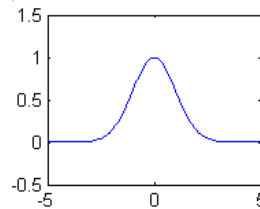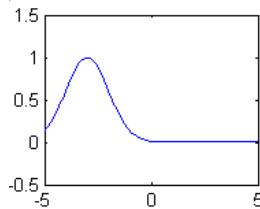
- Variance parameter σ^2 controls the width of the Gaussian

$$p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$
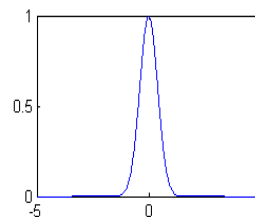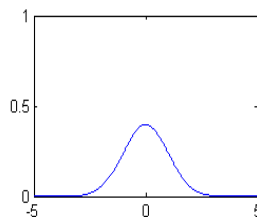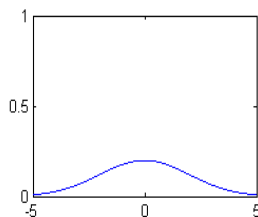
Note: $\int_{-\infty}^{\infty} p(x)dx = 1$

Multivariate Gaussian

- Gaussian can extend to D-dimensions

- Gaussian mean parameter μ vector, it translates the bump
- Covariance matrix Σ stretches and rotates bump

$$p(\vec{x}|\vec{\mu},\Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}}\sqrt{|\Sigma|}}\, exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})\Sigma^{-1}(\vec{x}-\vec{\mu})\right)$$

- Mean is any real vector
- Max and expectation = μ
- Variance parameter is now Σ matrix
- Covariance matrix is positive definite
- Covariance matrix is symmetric
- Need matrix inverse (inv)
- Need matrix determinant (det)
- Need matrix trace operator (trace)
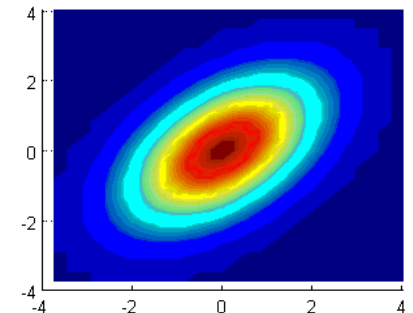
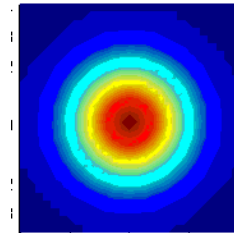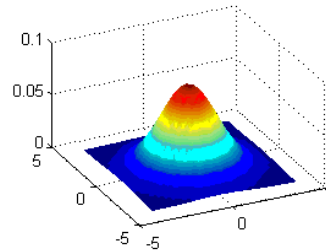# Multivariate Gaussian

- Spherical:

$$\Sigma = \sigma^2 I = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix}$$
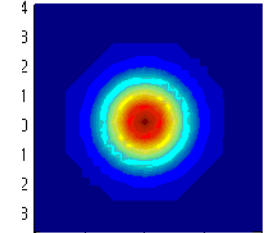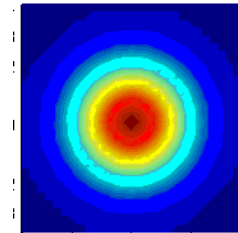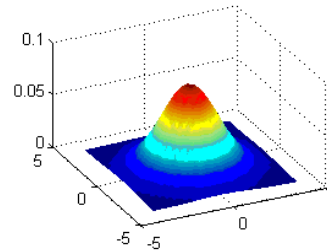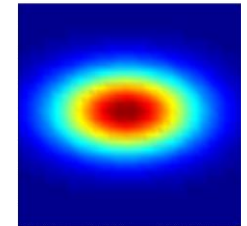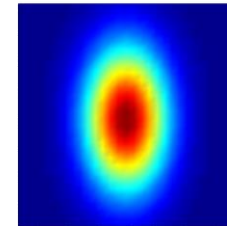
Multivariate Gaussian

- Spherical:

$$\Sigma = \sigma^2 I = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix}$$
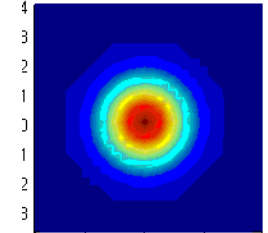
- Diagonal Covariance: $\hat{Y}$
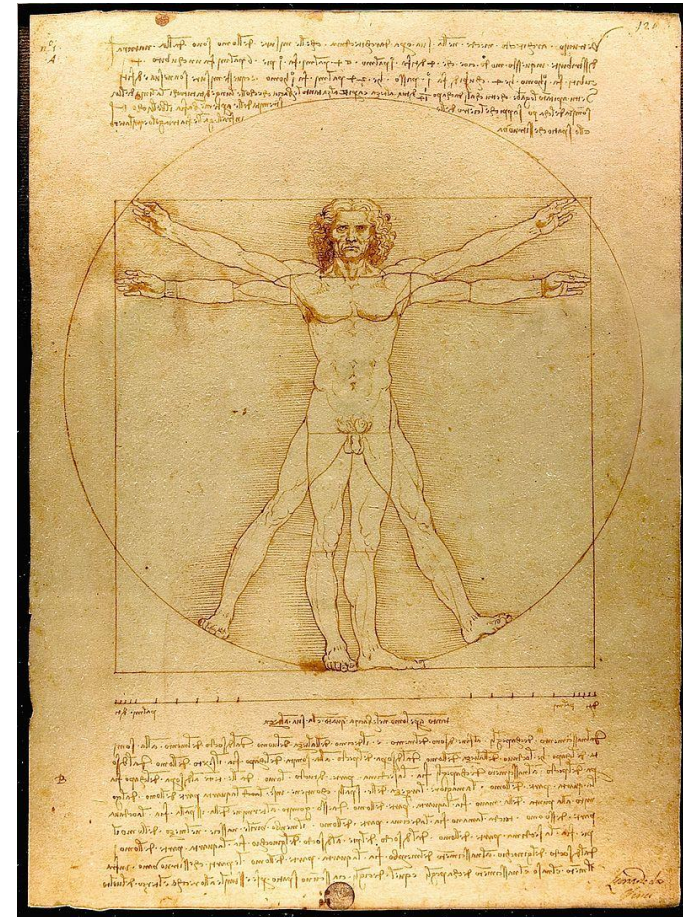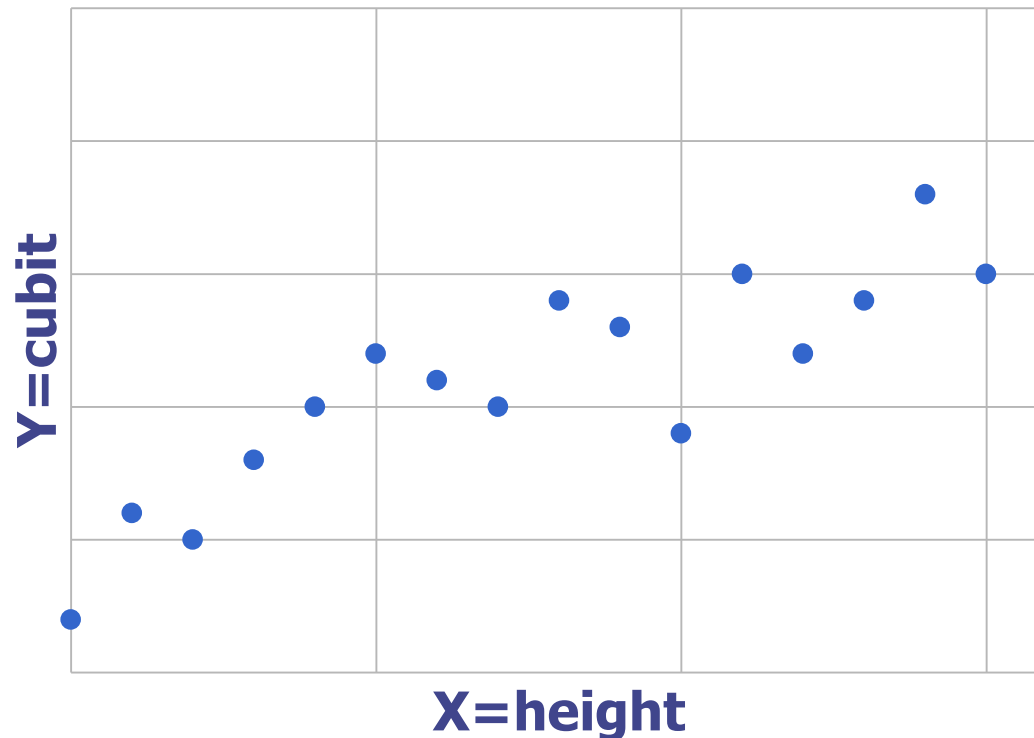 dimensions of x are independent
  product of multiple 1d Gaussians

$$p(\vec{x}|\vec{\mu}, \Sigma) = \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi}\vec{\sigma}(d)} exp\left(-\frac{(\vec{x}(d) - \vec{\mu}(d))^2}{2\vec{\sigma}(d)^2}\right)$$

$$\Sigma = \begin{bmatrix} \vec{\sigma}(1)^2 & 0 & 0 & 0 \\ 0 & \vec{\sigma}(2)^2 & 0 & 0 \\ 0 & 0 & \vec{\sigma}(3)^2 & 0 \\ 0 & 0 & 0 & \vec{\sigma}(4)^2 \end{bmatrix}$$
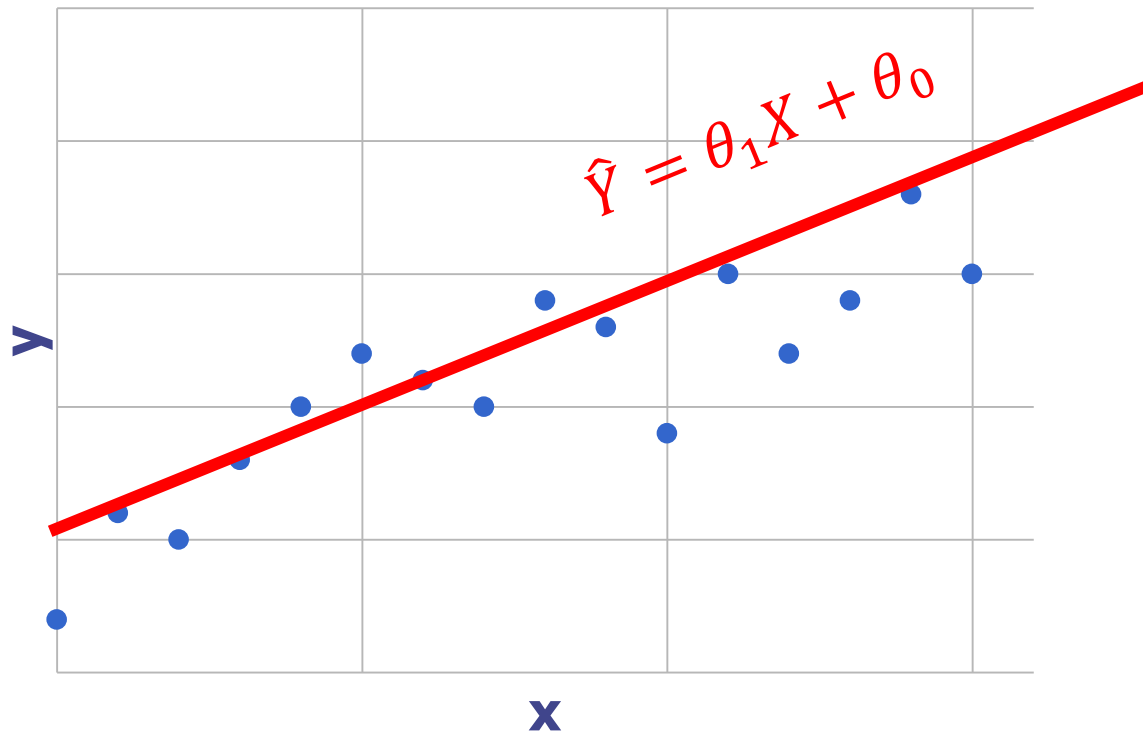
Regression and Gaussians

Vitruvian Man: cubit = ¼ height

Reality : cubit = ¼ height + noise

Y=cubit

X=height

Regression and Gaussians

Assume y_i is supposed to be $\hat{y}_i$ but many iid, small sources of error
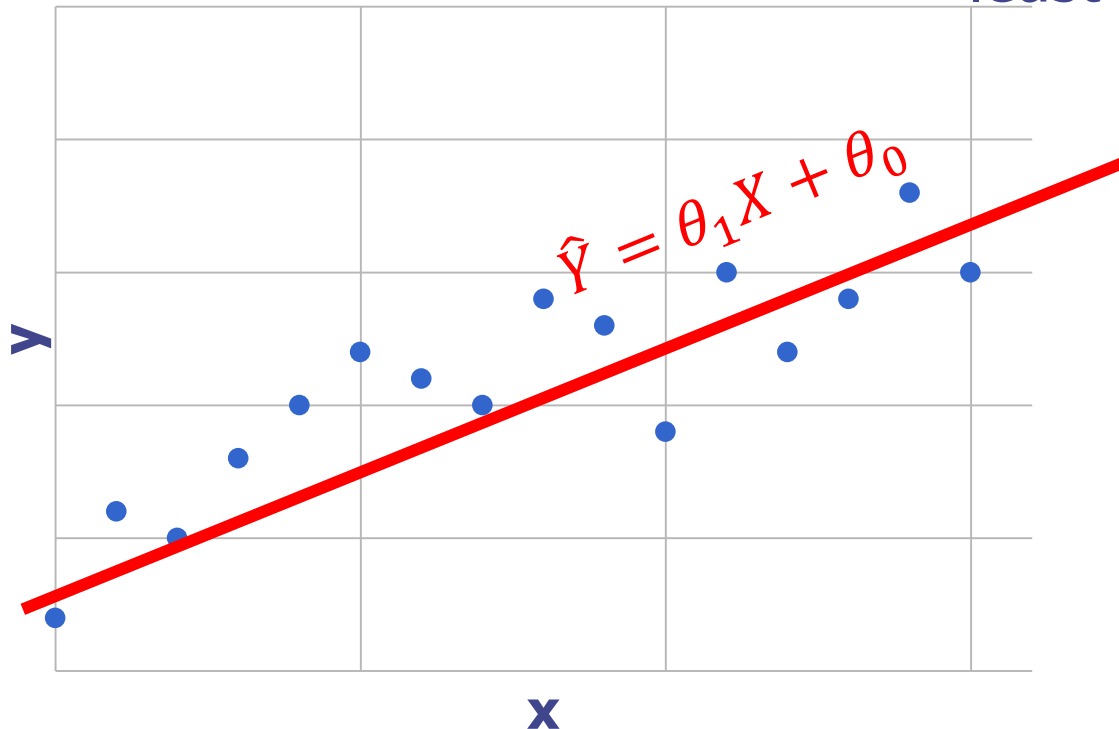

$\hat{Y} = \theta_1 X + \theta_0$

Regression and Gaussians

Assume y_i is supposed to be $\hat{y}_i$ but many iid, small sources of error $y_i \sim Normal(\hat{y}_i, \sigma^2)$

log-likelihood:

$$l(Y) = \log \prod_i \text{Prob}(y_i | \hat{y}_i, \sigma^2) = C - \frac{1}{2\sigma^2} \sum_i (y_i - \hat{y}_i)^2$$

least squares=max likelihood

$\hat{Y} = \theta_1 X + \theta_0$

y

x

Summary

◆ Probability definitions, distributions, moments, theorems

◆ Gaussians motivate least squares