6.s077 — INTRODUCTION TO DATA SCIENCE
EECS, MIT, Spring 2018

Lecture 5

Multidimensional data, $\mathbb{E}[Y \mid X]$
Bayesian Inference: posterior distribution

Today's agenda — Part I

- Multidimensional data
  - visualization
  - i.i.d. models
  - estimation (plugin, feature matching, ML)

- Predicting one variable from another
  - model-based, $\mathbb{E}[Y \mid X]$: today
  - based on data (X_i, Y_i): come back during regression lectures

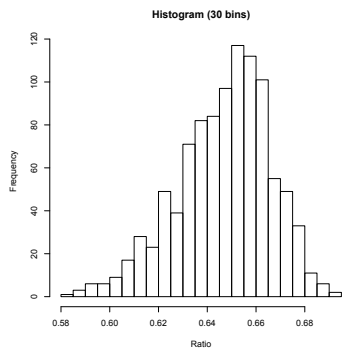
Multidimensional data

- ith sample (x_i, y_i, z_i)

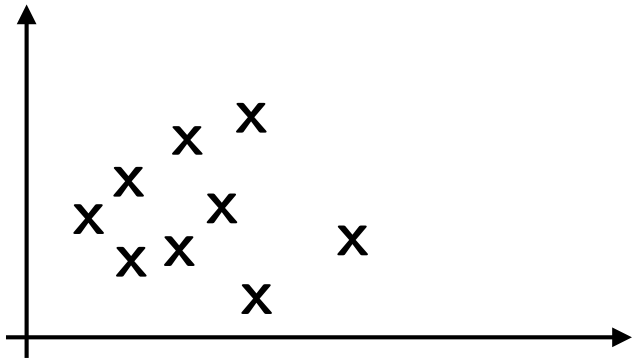
  e.g., (gender, height, weight) of ith person

- Visualize:

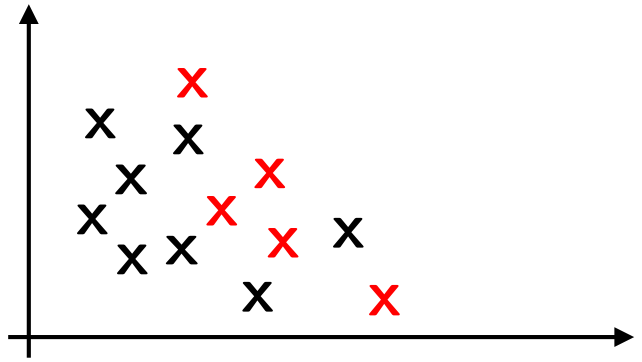
one at a time



two at a time



three, one categorical



"Standard model:" Independent random sampling

- Assume two-dimensional data (only for exposition purposes)
- $\mathbb{P}$: distribution on $\mathbb{R}^2$ (discrete; continuous; mixed)

$$p_{X,Y}(x,y) \quad f_{X,Y}(x,y)$$

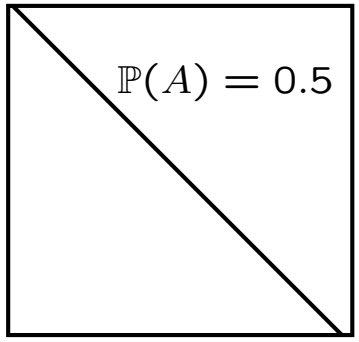
- $(X_i, Y_i) \sim \mathbb{P}$, independent for different i

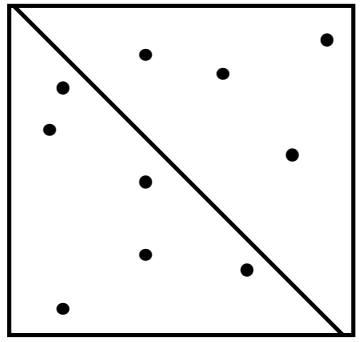
  X_i and Y_i usually dependent

- $\widehat{\mathbb{P}}$: empirical distribution on $\mathbb{R}^2$
  - weight $1/n$ on each data point

    $\widehat{\mathbb{P}}(A) =$ fraction of data that belong to A

uniform $\mathbb{P}$





$\widehat{\mathbb{P}}(A) = 4/10$

Parameter estimation with multidimensional data

- Nothing new!

- **Plugin:**

$$\theta = \mathbb{E}\Big[g(X,Y)\Big] \qquad \widehat{\Theta} = \widehat{\mathbb{E}}\Big[g(X,Y)\Big] = \frac{1}{n}\sum_{i=1}^{n} g(X_i, Y_i)$$

$$\theta = \mathsf{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$$

$$\widehat{\Theta} = \widehat{\mathbb{E}}[XY] - \widehat{\mathbb{E}}[X]\,\widehat{\mathbb{E}}[Y]$$

Parameter estimation with multidimensional data (ctd.)

- **Feature matching and the method of moments**
- Example:

  $\mathbb{P}^\theta \qquad \theta = (\theta_1, \theta_2, \theta_3)$

  Choose $\mathbb{E}^\theta[X],\ \mathbb{E}^\theta[Y],\ \mathbb{E}^\theta[XY]$ as features

$$\begin{aligned} \mathbb{E}^\theta[X] &= \widehat{\mathbb{E}}[X] \\ \mathbb{E}^\theta[Y] &= \widehat{\mathbb{E}}[Y] \\ \mathbb{E}^\theta[XY] &= \widehat{\mathbb{E}}[XY] \end{aligned} \qquad \text{solve for } \theta = (\theta_1, \theta_2, \theta_3)$$

  - easiest when we have closed-form expressions (as a function of θ) for left-hand side

Parameter estimation with multidimensional data (ctd.)

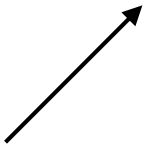
- **Maximum Likelihood**

$$\max_{\theta} \quad \prod_{i=1}^{n} \mathbb{P}^{\theta}(X_i = x_i) \qquad \prod_{i=1}^{n} f_X^{\theta}(x_i)$$

X_i, x_i are now vectors $\qquad$ f_X is a joint PDF

  - Nothing new
    - nice theoretical properties remain

- Sampling distribution, bootstrap, confidence intervals: again nothing new



one parameter at a time

What is new? New types of prediction problems

θ?	$\mathbb{P}_Y^\theta$
i.i.d.	Y_1
	$\vdots$
	Y_n
	Y?

$$\min_{\widehat{y}} \ \mathbb{E}^\theta\left[(Y-\widehat{y})^2\right]$$

θ?	$\mathbb{P}_{X,Y}^\theta$
X_1	Y_1
$\vdots$	$\vdots$
X_n	Y_n
X	Y?

$$\min_{g(\cdot)} \ \mathbb{E}^\theta\left[(Y-g(X))^2\right]$$

$\widehat{Y}=g(X)$ estimator

A. Have $\mathbb{P}_Y^\theta$, know θ

$\widehat{Y}=\mathbb{E}^\theta[Y]$

Have $\mathbb{P}_{X,Y}^\theta$, know θ

$\widehat{Y}=\mathbb{E}^\theta[Y\,|\,X]$

data $X_1, Y_1, \ldots, X_n, Y_n$ are irrelevant

Some details

$$\mathbb{E}\Big[\big(Y - \mathbb{E}[Y]\big)^2\Big] \le \mathbb{E}\Big[\big(Y - c\big)^2\Big] \qquad \text{for all } c$$

apply to conditional universe, where $X = x$ has occurred

$$\mathbb{E}\Big[\big(Y - \mathbb{E}[Y \mid X = x]\big)^2 \;\Big|\; X = x\Big] \le \mathbb{E}\Big[\big(Y - c\big)^2 \;\Big|\; X = x\Big], \qquad \text{for all } c$$

$$\mathbb{E}\Big[\big(Y - \mathbb{E}[Y \mid X = x]\big)^2 \;\Big|\; X = x\Big] \le \mathbb{E}\Big[\big(Y - g(x)\big)^2 \;\Big|\; X = x\Big], \qquad \text{for all } g(\cdot)$$

$$\mathbb{E}\Big[\big(Y - \mathbb{E}[Y \mid X]\big)^2 \;\Big|\; X\Big] \le \mathbb{E}\Big[\big(Y - g(X)\big)^2 \;\Big|\; X\Big], \qquad \text{for all } g(\cdot)$$

take expectation of both sides; use law of iterated expectations, and the property $X \le Y \;\Rightarrow \mathbb{E}[X] \le \mathbb{E}[Y]$

$$\mathbb{E}\Big[\big(Y - \mathbb{E}[Y \mid X]\big)^2\Big] \le \mathbb{E}\Big[\big(Y - g(X)\big)^2\;\Big], \qquad \text{for all } g(\cdot)$$

Prediction via a model versus directly from data

B. Have $\mathbb{P}^{\theta}_{X,Y}$, do not know θ

θ?	$\mathbb{P}^{\theta}_{X,Y}$
X_1	Y_1
$\vdots$	$\vdots$
$\vdots$	$\vdots$
X_n	Y_n
X	Y?

$\longrightarrow$ regression

Today's agenda — Part II

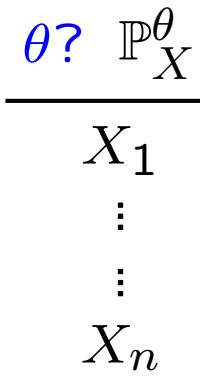
- The Bayesian framework $\longrightarrow$ posterior distributions
- Four versions of Bayes' rule
- Bayesian confidence intervals
- Coin parameter example

Today's agenda — Part II

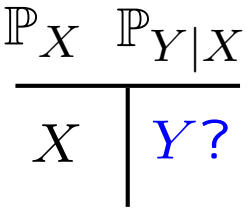
θ? $\mathbb{P}_X^\theta$	
X_1 $\vdots$ $\vdots$ X_n	• Classical/frequentist statistics: θ: unknown constant

- Bayesian philosophy
  - parameter treated as (unobserved) realized value of random variable Θ
  - prior distribution $\mathbb{P}_\Theta$: $p_\Theta(x)$ or $f_\Theta(\theta)$ joint PMF/PDF if θ is a vector
- Observation model $\mathbb{P}_{X|\Theta}$: $p_{X|\Theta}(x \,|\, \theta)$ or $f_{X|\Theta}(x \,|\, \theta)$ replace $p^\theta(x)$, $f^\theta(x)$
- Prior comes from: symmetry; known range; earlier studies; subjective beliefs

The big picture

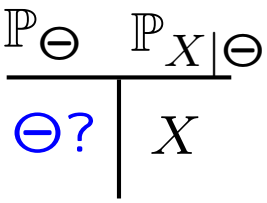


Classical inference



Model-based prediction

$\mathbb{E}[Y \mid X]$

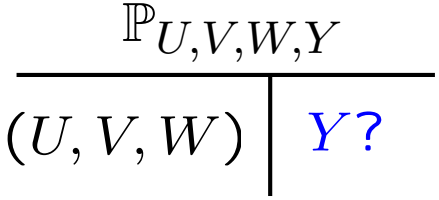


Bayesian inference

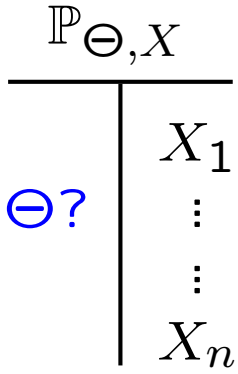
$\mathbb{E}[\Theta \mid X]$

$$\mathbb{P}_{\Theta|X} = \frac{\mathbb{P}_\Theta \cdot \mathbb{P}_{X|\Theta}}{\mathbb{P}_X}$$

Two different flavors



prediction

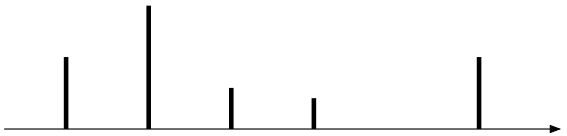


inference

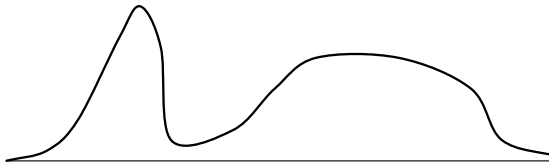
The general answer to Bayesian inference problems

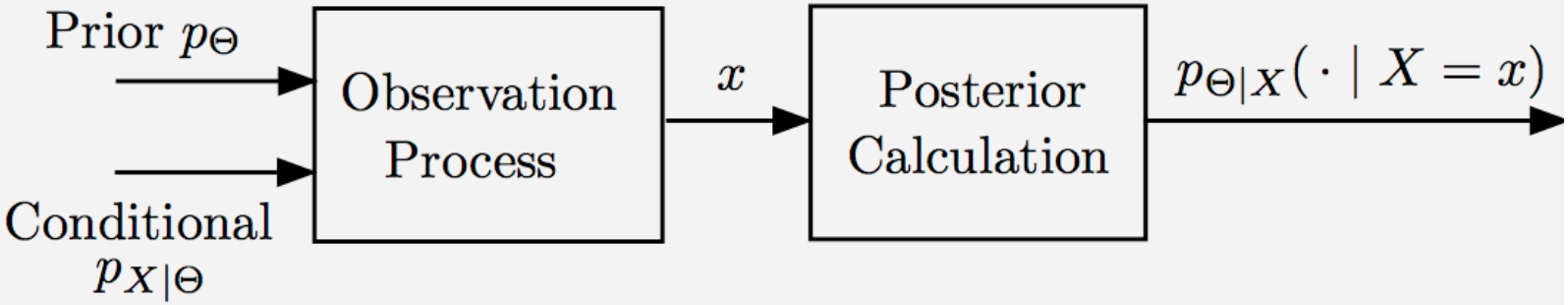
$\mathbb{P}_\Theta$	$\mathbb{P}_{X\mid\Theta}$
Θ?	X

$p_{\Theta\mid X}(\cdot \mid x)$



$f_{\Theta\mid X}(\cdot \mid x)$





- Use appropriate version of the Bayes rule to find $p_{\Theta\mid X}(\cdot \mid X = x)$ or $f_{\Theta\mid X}(\cdot \mid X = x)$

$\mathbb{P}_\Theta,\ \mathbb{P}_{X\mid\Theta} \ \rightarrow\ \mathbb{P}_{X,\Theta} \ \rightarrow\ \mathbb{P}_{\Theta\mid X}$

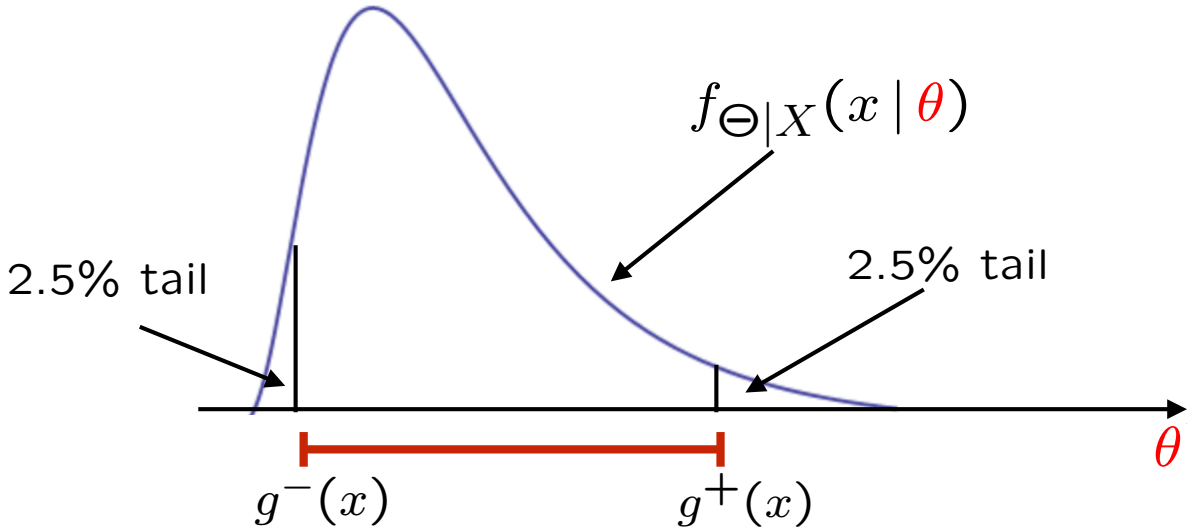
$$\mathbb{P}_{\Theta\mid X} = \frac{\mathbb{P}_\Theta \cdot \mathbb{P}_{X\mid\Theta}}{\mathbb{P}_X}$$

Four versions of the Bayes' rule

$$\mathbb{P}_{\Theta|X} = \frac{\mathbb{P}_\Theta \cdot \mathbb{P}_{X|\Theta}}{\mathbb{P}_X}$$

Variation	Canonical example	Bayes' rule
Θ discrete X discrete	ill/healthy test positive/negative	$p_{\Theta\|X}(\theta\,\|\,x) = \dfrac{p_\Theta(\theta)\cdot p_{X\|\Theta}(x\,\|\,\theta)}{p_X(x)}$
Θ discrete X continuous	signal Θ=0/1 $X = \Theta$+(independent noise)	$p_{\Theta\|X}(\theta\,\|\,x) = \dfrac{p_\Theta(\theta)\cdot f_{X\|\Theta}(x\,\|\,\theta)}{f_X(x)}$
Θ continuous X discrete	Θ: coin bias $X \sim$Binomial(n, Θ)	$f_{\Theta\|X}(\theta\,\|\,x) = \dfrac{f_\Theta(\theta)\cdot p_{X\|\Theta}(x\,\|\,\theta)}{p_X(x)}$
Θ continuous X continuous	$X = \Theta + W$ Θ, W: independent, normal	$f_{\Theta\|X}(\theta\,\|\,x) = \dfrac{f_\Theta(\theta)\cdot f_{X\|\Theta}(x\,\|\,\theta)}{f_X(x)}$

Bayesian confidence intervals



95% coinfidence interval

$$\mathbb{P}\Big(g^-(x) \leq \Theta \leq g^+(x) \,|\, X = x\Big) = 0.95$$

- Contrast with classical CI:

$$\mathbb{P}^\theta\Big(g^-(X) \leq \theta \leq g^+(X)\Big) = 0.95$$

Inferring the unknown bias of a coin

- Coin with bias Θ
  - prior: uniform on $[0,1]$: $f_\Theta(\theta) = 1$, for $0 \le \theta \le 1$
- n independent tosses
  - given $\Theta = \theta$, $K \sim$ Binomial(n,θ): $p_{K|\Theta}(k \,|\, \theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$

$$f_{\Theta|K}(\theta \mid k) = \frac{f_\Theta(\theta)\, p_{K|\Theta}(k \mid \theta)}{p_K(k)}$$

$$p_K(k) = \int f_\Theta(\theta')p_{K|\Theta}(k \mid \theta')\, d\theta'$$

$$= \frac{1}{d(n,k)}\,\theta^k(1-\theta)^{n-k}$$

dependence on θ

normalizing constant

- Animation: https://av2076.herokuapp.com/bayesian