MSBA7003 Decision Analytics

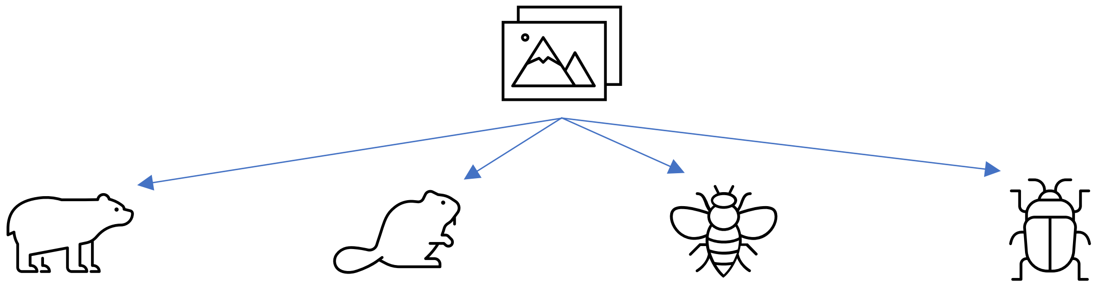
02 Probability & Bayesian Learning II

Agenda

- Bayesian Inference
  - Classification Models
  - Naïve Bayes
- Application
  - Classification: The Authorship Problem
  - Decision Support: New Product Pricing

Bayesian Inference

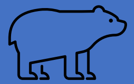
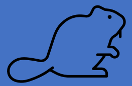
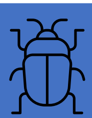
- We are interested in knowing the "state of the world Y" (e.g., demand is high or low or the true type of a subject), and there are K possible states, which we call the ***Alternative Hypotheses***.



- The alternative hypotheses are mutually exclusive and collectively exhaustive. We have a prior subjective belief on each state (i.e., a marginal distribution of Y).

Bayesian Inference

- Under each hypothesis, a random variable X will follow a known, distinct distribution.
- We wish to infer the state Y by collecting a sample of X given the unknown state.

Dominant Pic Color					Marginal
Red	0.3	0.5	0.3	0.2	0.37
Green	0.6	0.5	0.4	0.4	0.48
Blue	0.1	0.0	0.3	0.4	0.15
Prior Belief	0.2	0.4	0.3	0.1	Sum = 1

- After observing the value of X, our subjective belief (marginal distribution) of Y can be updated according to the Bayes' rule. *The posterior becomes the new belief*.
- If multiple samples of X can be obtained, the marginal distribution of X will converge to the conditional distribution given the "true" state Y with enough data points.

Bayesian Inference

- What if X follows a continuous distribution under each hypothesis and only a specific value of X is observed?
- P(X = x | Y) = 0.
- We use the conditional density to approximate the conditional probability:
- F(x | Y) = P(X ≤ x | Y); f(x | Y) = F'(x | Y).
- P(X = x | Y) ≈ f(x | Y)*d.
- d is a very small positive number.

- If Y = 0 or 1, then $\Pr(Y = 0|X = x) = \frac{\Pr(Y=0)*f(x|Y=0)*d}{\Pr(Y=0)*f(x|Y=0)*d+\Pr(Y=1)*f(x|Y=1)*d}$.

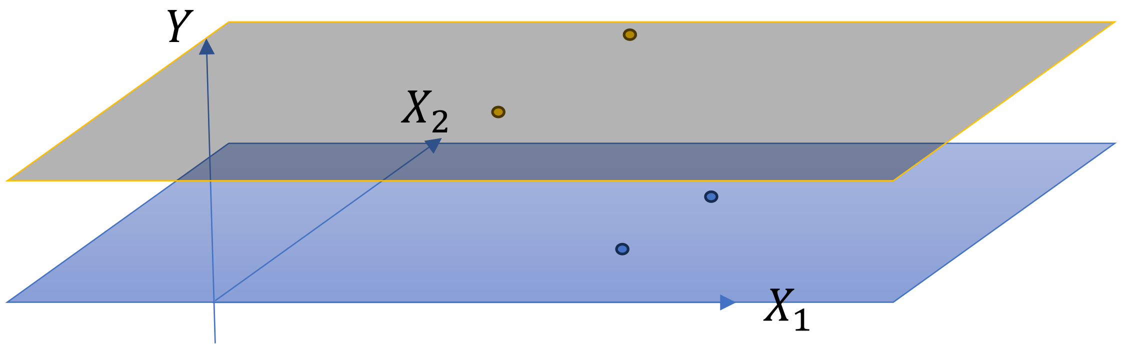
Binary Classification

- Consider a binary classification problem.
- Suppose we have data as shown on the right.
  - The data has correct labels
  - There are two features
- We compare three different classification models.
  - Linear regression
  - Decision tree
  - Bayesian inference
- For a new subject with features (3,2), what should be the label?

Y	X1	X2
0	6	2
1	3	4
1	4	8
0	7	5
…	…	…

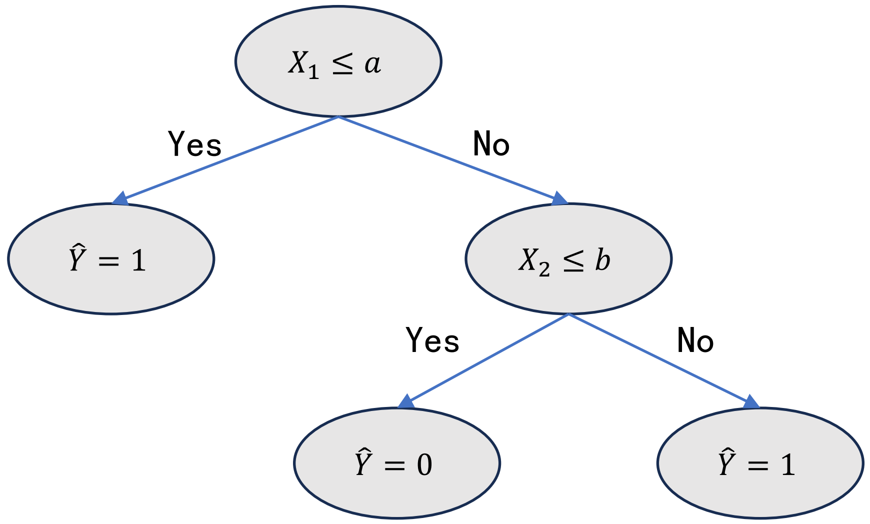
Binary Classification

- Linear Regression
  - Assume the model structure: $Y = \alpha + \beta X_1 + \gamma X_2 + \epsilon$
  - Prediction/classification: $\hat{Y} = \alpha + \beta X_1 + \gamma X_2$
  - Given data set $\{Y_i, X_{1i}, X_{2i},\}_{i=1}^{n}$, generate predictions $\widehat{Y_i} = \alpha + \beta X_{1i} + \gamma X_{2i}$ for all i
  - Optimize $\{\alpha, \beta, \gamma\}$ to minimize $\sum_i \left(\hat{Y}_i - Y_i\right)^2$



Binary Classification

- Decision Tree
  - The algorithm finds a progressive classification rule that maximizes the information gain (or reduction of entropy) at each classification step.
  - Some other rules may be used.
  - Step 1:
    - Try thresholds $a_1, a_2, \ldots, a_K$ for X_1 and compute the corresponding entropy values.
    - Try thresholds $b_1, b_2, \ldots, b_K$ for X_2 and compute the corresponding entropy values.
    - Find the division that minimizes the entropy.
  - Step 2:
    - Try thresholds for the feature left and compute the corresponding entropy values.
    - Find the division that minimizes the total entropy.



Binary Classification

- Bayesian Inference
  - Define Y = 0 and 1 as two hypotheses.
  - Obtain the prior: Compute the proportion of each hypothesis in the data set.
  - Obtain the conditional distributions of X: Compute the relative frequency of observing each possible combination (X_1,X_2) in the data under each hypothesis.
  - Given the feature values (3,2) of the new subject, compute the posterior probability of each hypothesis.
  - Compare and predict the value of Y.

(X_1, X_2)	$Y = 0$	$Y = 1$
6, 2	0.70	0.10
3, 4	0.05	0.20
4, 8	0.00	0.20
7, 5	0.25	0.10
1, 2	0.00	0.10
2, 3	0.00	0.15
…	…	…
Prior	0.35	0.65

Advantage over linear regression: it can obtain the probability of correct/wrong prediction.
Advantage over decision tree: it requires less computation.

Naïve Bayes

- When there are multiple features, Naïve Bayes method assumes that the features are independent under a given hypothesis.
- The features may be correlated without a condition.

- $\Pr\{Y = a | X_1 = b, X_2 = c\} = \frac{\Pr\{X_1=b, X_2=c | Y=a\} \times \Pr\{Y=a\}}{\Pr\{X_1=b, X_2=c\}} = \frac{\Pr\{X_1=b | Y=a\} \times \Pr\{X_2=c | Y=a\} \times \Pr\{Y=a\}}{\Pr\{X_1=b, X_2=c\}}$.

- Without knowing $\Pr\{X_1 = b, X_2 = c\}$, we can compare the posteriors of different hypotheses by computing the ratio:
- $\frac{\Pr\{Y=a_1 | X_1=b, X_2=c\}}{\Pr\{Y=a_2 | X_1=b, X_2=c\}} = \frac{\Pr\{X_1=b | Y=a_1\} \times \Pr\{X_2=c | Y=a_1\} \times \Pr\{Y=a_1\}}{\Pr\{X_1=b | Y=a_2\} \times \Pr\{X_2=c | Y=a_2\} \times \Pr\{Y=a_2\}}$

Naïve Bayes

- For the binary classification problem, we create two separate distribution tables:

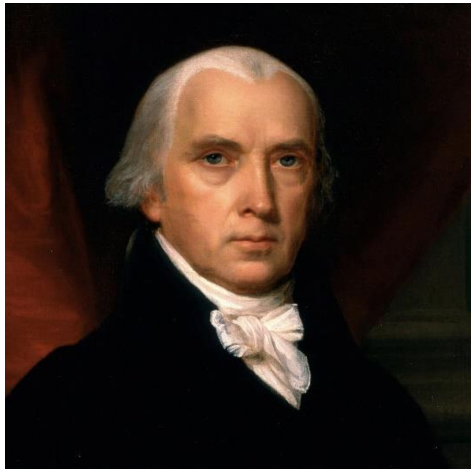
X_1	$Y = 0$	$Y = 1$
6	0.45	0.10
3	0.05	0.20
4	0.05	0.20
7	0.25	0.10
1	0.05	0.10
2	0.05	0.15
…	…	…
Prior	0.35	0.65

X_2	$Y = 0$	$Y = 1$
2	0.60	0.15
4	0.05	0.20
8	0.00	0.15
5	0.15	0.10
1	0.00	0.10
3	0.10	0.15
…	…	…
Prior	0.35	0.65

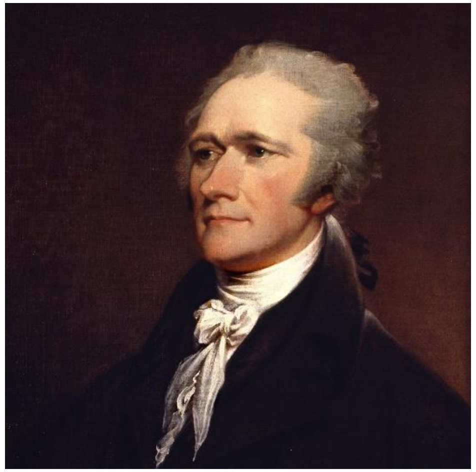
- $\frac{\Pr\{Y=0|X_1=3,X_2=2\}}{\Pr\{Y=1|X_1=3,X_2=2\}} = \frac{0.05\times0.60\times0.35}{0.20\times0.15\times0.65} = 0.54 < 1$. Hence, predict Y = 1.

The Authorship Problem

- Federalist Papers (Published anonymously during 1787 - 1788)
  - Total: 77 papers
  - John Jay: 5
  - Alexander Hamilton: 43
  - James Madison: 14
  - Unknown: 12 + 3
- Bayesian Inference
  - Establish hypotheses: H_h vs. H_m
  - Determine the prior belief: 0.5 vs. 0.5 (or 0.75 vs. 0.25)
  - Collect data (the wording pattern in each paper)
  - Compute the probability of observing the data under each hypothesis
    - Using the papers with a known author
  - Compute the posterior of each hypothesis (for the papers in question)



James Madison (1751 - 1836)



Alexander Hamilton (1757 - 1840)

The Authorship Problem

- Focus on non-contextual words
    - Rate of use is nearly invariant under change of topic.
    - Focus on the word [upon]
    - In paper 54, occurrence rate PTW = 0.996
- $$\frac{\Pr(H\_h|data)}{\Pr(H\_m|data)} = \frac{\Pr(data|H\_h)\cdot\Pr(H\_h)}{\Pr(data|H\_m)\cdot\Pr(H\_m)}$$

TABLE 2.3. FREQUENCY DISTRIBUTION FOR *upon*

Rate/1000	H	M
0 (exactly)	—	41
0+–1	1	7
1 –2	10	2
2 –3	11	
3 –4	11	
4 –5	10	
5 –6	3	
6 –7	1	
7 –8	1	
Totals	48	50

TABLE 2.5. FUNCTION WORDS AND THEIR CODE NUMBERS FOR THE FEDERALIST STUDY

1 a	8 as	15 do	22 has	29 is	36 no	43 or	50 than	57 this	64 when
2 all	9 at	16 down	23 have	30 it	37 not	44 our	51 that	58 to	65 which
3 also	10 be	17 even	24 her	31 its	38 now	45 shall	52 the	59 up	66 who
4 an	11 been	18 every	25 his	32 may	39 of	46 should	53 their	60 upon	67 will
5 and	12 but	19 for	26 if	33 more	40 on	47 so	54 then	61 was	68 with
6 any	13 by	20 from	27 in	34 must	41 one	48 some	55 there	62 were	69 would
7 are	14 can	21 had	28 into	35 my	42 only	49 such	56 thing	63 what	70 your

TABLE 2.6. ADDITIONAL WORDS AND CODE NUMBERS FOR THE FEDERALIST STUDY

*71 affect	*79 city	*87 direction	*94 innovation	102 perhaps	*110 vigor
*72 again	*80 commonly	*88 disgracing	*95 join	*103 rapid	*111 violate
*73 although	*81 consequently	89 either	*96 language	104 same	*112 violence
74 among	*82 considerable	*90 enough (and in sample of 20)	97 most	105 second	*113 voice
75 another	*83 contribute		98 nor	106 still	114 where
76 because	*84 defensive	*91 fortune	*99 offensive	107 those	115 whether
77 between	*85 destruction	*92 function	100 often	*108 throughout	*116 while
78 both	86 did	93 himself	*101 pass	109 under	*117 whilst

TABLE 2.7. NEW WORDS FROM THE WORD INDEX STUDY TOGETHER WITH THEIR CODE NUMBERS

118 about	130 choice	142 intrust+s+ed+ing	154 proper
119 according	131 common	143 kind	155 propriety
120 adversaries	132 danger	144 large	156 provision+s
121 after	133 decide+s+ed+ing	145 likely	157 requisite
122 aid	134 degree	146 matter+s	158 substance
123 always	135 during	147 moreover	159 they
124 apt	136 expence+s	148 necessary	160 though
125 asserted	137 expense+s	149 necessity+ies	161 truth+s
126 before	138 extent	150 others	162 us
127 being	139 follow+s+ed+ing	151 particularly	163 usage+s
128 better	140 I	152 principle	164 we
129 care	141 imagine+s+ed+ing	153 probability	165 work+s

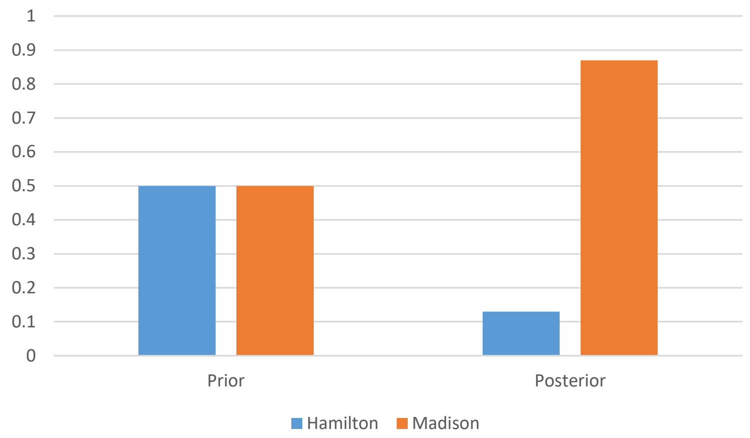
The Authorship Problem

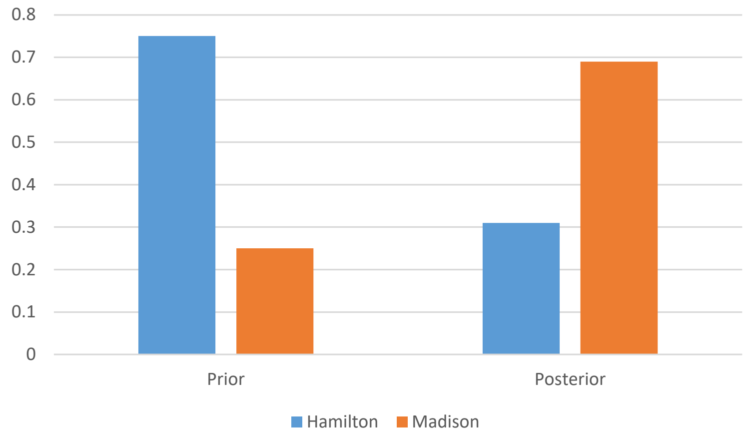
- $\Pr(data|H\_h) = 1/48$; $\Pr(data|H\_m) = 7/50$

- Hence, if $\Pr(H\_h) : \Pr(H\_m) = 1{:}1$, then
  - $\Pr(H\_h|data) / \Pr(H\_m|data) = (1/48)/(7/50) = 0.15 < 1$
  - Because $\Pr(H\_h|data) + \Pr(H\_m|data) = 1$, we get
  - $\Pr(H\_h|data) = 0.13$; $\Pr(H\_m|data) = 0.87$.

- If $\Pr(H\_h) : \Pr(H\_m) = 0.75{:}0.25$, then
  - $\Pr(H\_h|data) / \Pr(H\_m|data) = 3*50/7/48 = 0.4464 < 1$
  - Because $\Pr(H\_h|data) + \Pr(H\_m|data) = 1$, we get
  - $\Pr(H\_h|data) = 0.31$; $\Pr(H\_m|data) = 0.69$.

- Conclusion: the author is more likely to be Madison.





The Authorship Problem

- Assume the use of different words are independent for a given author.
- Consider three more non-contextual words: *by*, *from*, and *to*. Their conditional distributions (in terms of rate PTW) are given in the below table:

Rate	*by*		Rate	*from*		Rate	*to*	
	H	M		H	M		H	M
1– 3	2		1– 3	3	3	20–25		3
3– 5	7		3– 5	15	19	25–30	2	5
5– 7	12	5	5– 7	21	17	30–35	6	19
7– 9	18	7	7– 9	9	6	35–40	14	12
9–11	4	8	9–11		1	40–45	15	9
11–13	5	16	11–13		3	45–50	8	2
13–15		6	13–15		1	50–55	2	
15–17		5				55–60	1	
17–19		3	Totals	48	50			
						Totals	48	50
Totals	48	50						

- Suppose, in paper 54, occurrence rates PTW for them are: 5.5, 3.7, and 46.

The Authorship Problem

- How should we update the posterior probabilities of the two hypotheses?
- $\frac{\Pr(H\_h|data)}{\Pr(H\_m|data)} = \frac{\Pr(data|H\_h)\cdot\Pr(H\_h)}{\Pr(data|H\_m)\cdot\Pr(H\_m)}$
- $= \frac{\Pr(upon|H\_h)\cdot\Pr(by|H\_h)\cdot\Pr(from|H\_h)\cdot\Pr(to|H\_h)\cdot\Pr(H\_h)}{\Pr(upon|H\_m)\cdot\Pr(by|H\_m)\cdot\Pr(from|H\_m)\cdot\Pr(to|H\_m)\cdot\Pr(H\_m)}$
- If $\Pr(H\_h) : \Pr(H\_m)$ = 1:1, then
- $\frac{\Pr(H\_h|data)}{\Pr(H\_m|data)} = \frac{\left(\frac{1}{48}\right)\left(\frac{12}{48}\right)\left(\frac{15}{48}\right)\left(\frac{8}{48}\right)}{\left(\frac{7}{50}\right)\left(\frac{5}{50}\right)\left(\frac{19}{50}\right)\left(\frac{2}{50}\right)}$ = 1.275 > 1; $\Pr(H\_h|data)$ = 0.56
- if $\Pr(H\_h) : \Pr(H\_m)$ = 3:1, then
- $\frac{\Pr(H\_h|data)}{\Pr(H\_m|data)}$ = 1.275*3 = 3.82 > 1; $\Pr(H\_h|data)$ = 0.79

Pricing with Unknown Demand

- Suppose Chow Tai Fook introduced a new gold ring. Historical sales data of similar rings suggest that customers' willingness to pay (WTP) follows a normal distribution with a standard deviation of $1,000. However, the mean WTP of this new ring is uncertain. It could be $2,000, $3,500, or $5,000. They are equally likely.
- The introductory price for this ring is $4,000. The cost for this ring is $2,000. There is sufficient inventory.
- If the first five customers who showed interested in this ring did not buy it, how should the price be adjusted afterwards?

Pricing with Unknown Demand

Mathematical foundations:

- w: the willingness to pay for a randomly sampled customer
- p: the price of the product
- A customer will purchase the product if and only if $w \geq p$
- If w follows cumulative distribution function F, then $\Pr\{w \geq p\} = 1 - F(p)$
- If there are three equally likely distributions: F_1, F_2, and F_3, then for a given price p we can build the following probability table:

	H1: F_1	**H2: F_2**	**H3: F_3**
Purchase	$1 - F_1(p)$	$1 - F_2(p)$	$1 - F_3(p)$
Walk away	$F_1(p)$	$F_2(p)$	$F_3(p)$
Prior	1/3	1/3	1/3

Pricing with Unknown Demand

Mathematical foundations (Cont'd)

- Given the posterior probabilities: β_1, β_2, and β_3, the estimated WTP CDF becomes

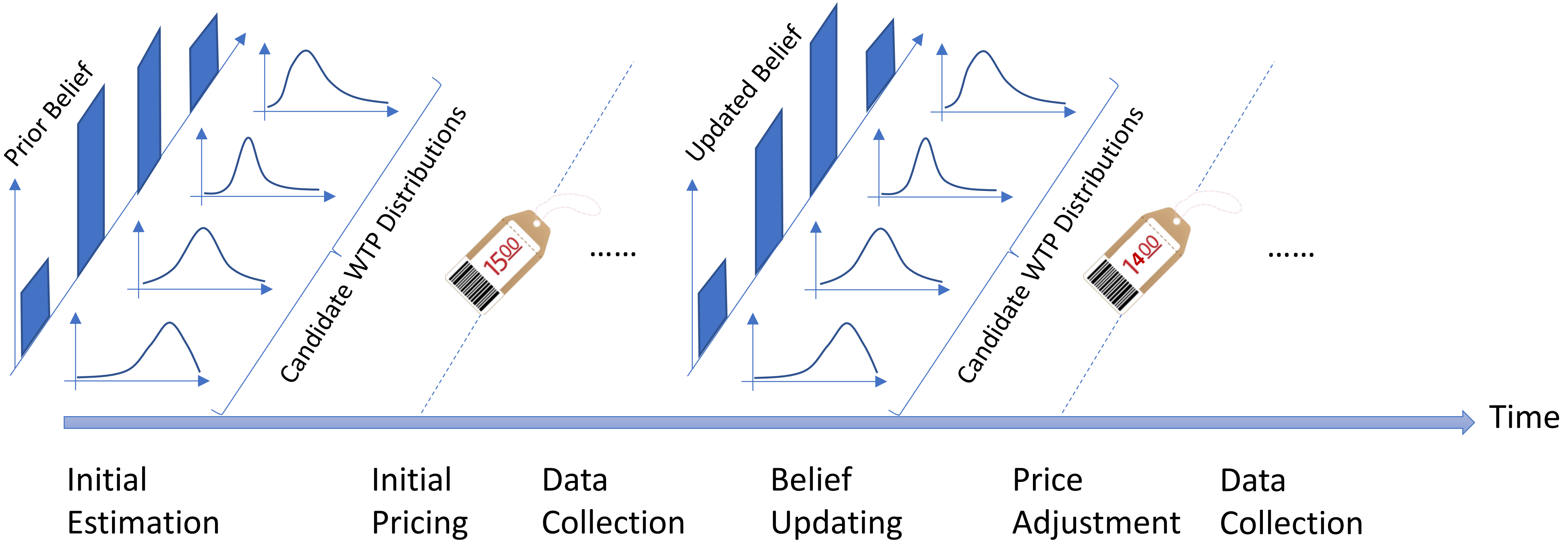
$$\hat{F}(v) = \Pr\{w \leq v\} = \beta_1 \cdot F_1(v) + \beta_2 \cdot F_2(v) + \beta_3 \cdot F_3(v)$$

- Given the estimated WTP CDF, the expected profit from a random customer under price p and unit cost c is

$$\pi(p) = (p - c) \times \left[1 - \hat{F}(p)\right]$$

- The firm's problem is to find optimal price p that maximizes $\pi(p)$.

Pricing with Unknown Demand



Pricing with Unknown Demand

- For Chow Tai Fook's new gold ring, the three possible WTP distributions are
- $N(2000, 1000^2)$, $N(3500, 1000^2)$, and $N(5000, 1000^2)$
- We can build the probability table in Excel and perform the belief updating:

	A	B	C	D	E
1	St. Dev.	1000			
2	Price	4000			
3	Hypothesis	1	2	3	
4	Mean	2000	3500	5000	Marginal
5	Purchase	0.02275	0.308538	0.841345	0.080447
6	Walk Away	0.97725	0.691462	0.158655	0.919553
7	Prior	0.799146	0.200299	0.000555	
8	Posterior	0.849289	0.150616	9.58E-05	

- The final posterior probabilities are 0.85, 0.15, and 0.

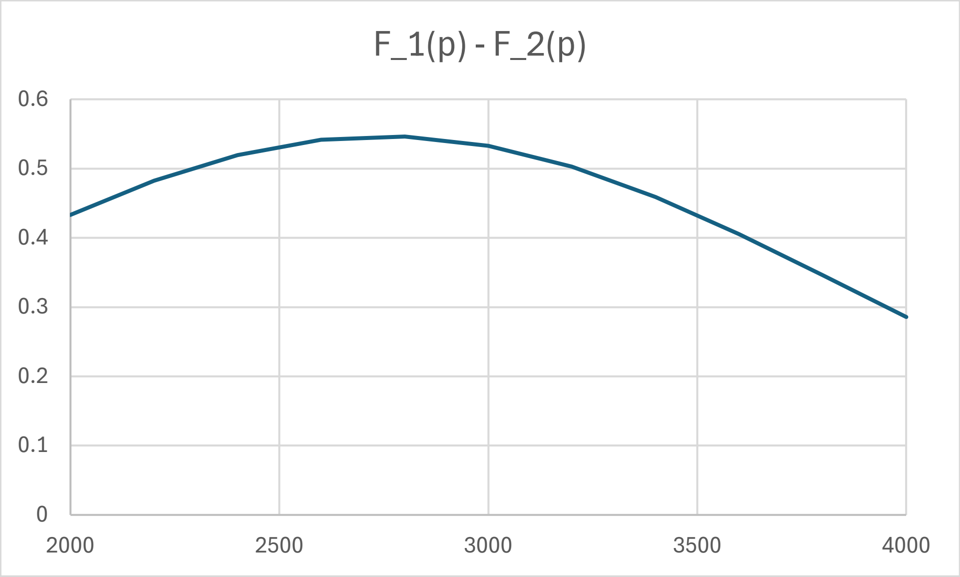
Pricing with Unknown Demand

- Use numerical method to search for the optimal price given the current belief:

cost	2000		
price	4000		
F(p)	0.97725	0.691462	0.158655
F_hat(p)	0.934127		
profit	131.7452		
price \ profit	131.7452		
2000	0		
2200	98.69229		
2400	169.1696		
2600	213.545		
2800	235.3544		
3000	238.9828		
3200	229.0642		
3400	209.9759		
3600	185.5002		
3800	158.6673		
4000	131.7452		

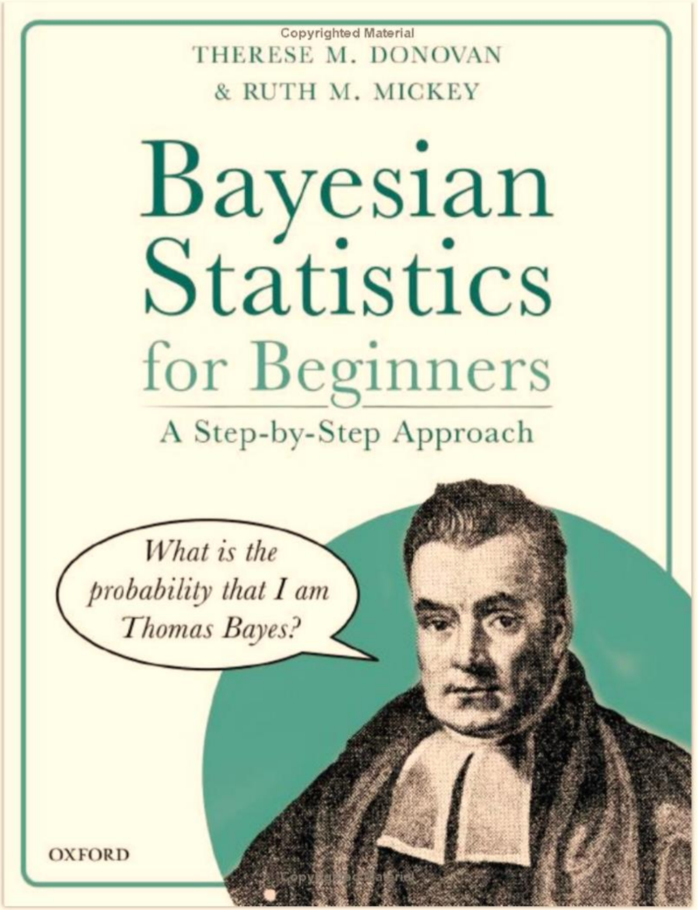
Pricing with Unknown Demand

- Is $3,000 really the optimal price? Why?
- Exploration vs. Exploitation
- Reinforcement Learning



More on Bayesian Inference & Machine Learning

- Bayesian Conjugates
- Markov Chain Monte Carlo (MCMC)
    - Metropolis algorithm
    - Metropolis-Hastings algorithm
    - Gibbs Sampling
- Bayesian Network
- Applications



Takeaways

- To do Bayesian inference, you need to
  - 1) construct MECE hypotheses,
  - 2) collect data to compute conditional probabilities about a feature under each hypothesis,
  - 3) form a prior belief, and
  - 4) sample the feature to get posterior beliefs.

- Your decision may affect the effectiveness of learning and payoff at the same time. In this case, you have a trade-off between exploitation and exploration.

Quiz

- This picture was taken from either Thailand or southern Vietnam. Based on the following information, we can conclude that it was taken from Thailand. True or False?
  - Motorbike ownership rate:
    - Southern Vietnam: 70%
    - Thailand: 35%
  - Sunny day rate:
    - Southern Vietnam: 53%
    - Thailand: 75%
  - Palm tree rate in vegetation:
    - Southern Vietnam: 7.5%
    - Thailand: 12.5%