STAT 391
Homework 5
Out Tuesday May 1, 2018
Due Tuesday May 8, 2018

©Marina Meilă
mmp@stat.washington.edu

Problem 1 – Statistical decision making

Rob (who is a robot roaming in the basement of the CSE building) has lost his graphics card. He is now unable to amuse himself playing video games at the times when he is not picking up empty coke cans or escorting lost students out of the basement.

He has lost the card in either room A (with probability 1/3) or room B (with probability 2/3). These rooms are humid and rat infested and the probability that the card (which was good and working just before it was lost) will survive a night in room A is 4/5. Room B is even worse: the probability that the graphics card is still good after a night in room B given that it was good the day before is 3/5. Call the event of the card being in good working condition G.

If the card is in room A and Rob spends a day searching for it in A then the probability of finding it (call this event F) that day is 1/2. The similar probability for finding the card on a day of search in room B is 2/5. Assume that the card surviving (or not surviving) the night and the card being found (not found) are independent events, e.g $P(F, G|A) = P(F|A)P(G|A)$, etc.

Rob can only search one room in a given day and he has at most two days to search.

a. Rob, not being very good with probabilities, isn't capable of figuring it all out; so he decides that he will start the same day (the entire day) by looking in room B, then looking in room A on the next day, if necessary. Make a neatly labeled table of the outcome space for this problem.

b. What is the probability that Rob finds the graphics card on the second day?

c. What is the probability that Rob finds the graphics card?

d. What is the probability that he finds the card and the card is still good?

e. Rob has established the following values for the outcomes of his search:

Finding card in good condition	+60	(this is the cost of buying a new card)
Each day (or part thereof) spent searching	-10	(energy costs)
Additional cost if he searches in both rooms	-5	(more energy)
Finding broken card	0	
Not finding card	-10	

What is the expected value of Rob's search policy?

f. Help Rob compare the following policies:

Plan 1: Search B on 1st day, search A on 2nd day if necessary.
Plan 2: Search A on 1st day, search B on 2nd day if necessary.
Plan 3: Search B on 1st day, don't search on 2nd day.
Plan 4: Search A on 1st day, don't search on 2nd day.
Plan 5: Don't search at all!

Rob would still have to choose his own decision criteria. However, to help him quantify his thinking choose one of the criteria below and determine which of the above policies optimize it.

i) Maximize his expected gain (graduate student)
ii) Maximize the probability that he will find his card in good state (idealist)

Justify your answers by showing your reasoning or your calculations in all cases.

[Problem 2 – Least squares – Not graded]

The number of requests per second for a cluster of servers follows the Poisson distribution

$$P(n) \;=\; e^{-\lambda}\frac{\lambda^n}{n!}$$

with $\lambda = 5$. The cluster has m servers and each of them can serve exactly 1 request in any given second. Therefore the number of requests served in a second is $s = \min(n, m)$. The cost of operating the cluster is quadratic i.e

$$C(n) \;=\; (n-m)^2$$

a. What are the mean and variance of the number of requests n ? What is the value $E[n^2]$ (as a function of λ and numerical value)? *Use the results in the notes, sections 8.4, 8.6, 8.7, to answer this question.*

b. What is the expected value of the cost of operating the cluster as a function of m the number of servers?

c. What is the value of m that minimizes the expected cost obtained in **b** and what is the resulting cost?

Problem 3 – Bayesian Inference

Nisqually River, Inc. sells books A,B,C on line. Each customer buys 0 or 1 copy of each title.

a. Mrs. Independence Day, the company's data mining expert, makes the assumptions that: i) a customer decides to buy a book independently of what other books (s)he buys and independently of other customers; ii) all customers buy according to the same *joint probability distribution* $\bar{P}_{ABC} = P_A P_B P_C$, with $P_A(1) = 0.6, P_B(1) = 0.3, P_C(1) = 0.4$. [For example, the probability that a customer buys A and B but not C is $\bar{P}_{ABC}(1,1,0) = 0.6 \times 0.3 \times (1-0.4)$.

Compute the probability that a customer buys all three books under $\bar{P}_{ABC}$, i.e., compute

$$\bar{P}_{ABC}(1,1,1)$$

b. Mr. Mean Variance, her aide, insists that Mrs. Day's model is not correct. He assumes that ii) all customers buy according to the same probability distribution; but iii) Book C is bought independently from A, B but A, B are not independent of each other.

$$C \;\perp\; A, B \tag{1}$$
$$A \;\not\perp\; B \tag{2}$$

Denote Mr. Variance's model by $\tilde{P}_{ABC} = \tilde{P}_{AB}\tilde{P}_C$ with $\tilde{P}_C(1) = 0.4$, and $\tilde{P}_{AB}$:

		A	
		0	1
B	0	0.2	0.5
	1	0.2	0.1

[For example, $\tilde{P}_{ABC}(1,0,0) = 0.5 \times (1-0.4)$.] Compute the probability that a customer buys all three books under $\tilde{P}$:

$$\tilde{P}_{ABC}(1,1,1)$$

c. Al Bayes, the guy who fills out the orders, notices that in fact women buy according to model $\bar{P}$ while men buy according to model $\tilde{P}$.

The next customer, Robin Hood, has ordered books A, B but not C. Using $\bar{P}$ and $\tilde{P}$ from above, determine if the likelihood that Robin Hood is a man is higher than the likelihood that (s)he's a woman.

e. Al also knows from experience that there are twice as many men customers than there are women. (Hence the prior probability that a customer is a man is 2/3.)

Determine the posterior probability that Robin Hood is a man, i.e

$$P(\text{man}|A=1,B=1,C=0)$$

f. Determine the the posterior probability that Robin Hood is a man, if Al, being forgetful, doesn't recall whether Robin ordered book C or not, but he is sure that (s)he ordered A and B.

$$P(\text{man}|A=1,B=1)$$

g. Follow-up on **c.**. The *Likelihood Ratio (LR)* of the two models given data $\mathcal{D}$ is $LR(\mathcal{D}) = \frac{\bar{P}(\mathcal{D})}{\tilde{P}(\mathcal{D})}$.

Compute the value of the LR for the data $A=1, B=1, C=0$.

Compute the value of the LR if the data consists of 3 customers $A_1=1, B_1=0, C_1=0$, $A_2=0, B_2=1, C_1=0$, $A_3=1, B_3=0, C_3=1$. The customers make their purchases independently of each other.

h. Give an example of a data set where $LR > 1$.

Problem 4 – Testing a hypothesis

[**a. Not graded**] Prove: If integer numbers with at most d digits are drawn uniformly,[1] then the distribution of the first digit is uniform over $S = \{0, 1, \dots 9\}$.

b. Denote by P_0 the uniform distribution over S. Consider the event E_t="in a data set of n integers, at least t/n of them start with 1". Write an expression of $p_t = P_0(E_t)$. This should be a function of t and n.

c. Read the first $n = 60$ data from file `hw6_digit.dat`, [2] then calculate the numerical value of p_t for these data. *You will find that $p_t \ll 1$!!. In other words, it is extremely unlikely for this data to be generated by P_0.* To convince yourself that not all p_t values are small, calculate also $p_{t'}$ for $t' = 0.11n$, by plugging t' into the expression from **b**.

[**d. Extra credit**] The formula in **b.** is impractical for large n. Compute the mean and variance of n_1/n for a given n, then use the CDF, normal approximation, and Φ^{-1} (CDF of standard normal) to obtain an approximate expression $\tilde{p}_t$ for p_t. Literal answer only for this part.

Now read the whole data in `hw6_digit.dat` and compute $\tilde{p}_t$ for the whole data set.

Problem 5 – Dirichlet/Beta distribution (Read: Ch 11 from textbook)

For $m = 2$, $S = \{1, 2\}$, the Dirichlet distribution is known as the Beta distribution, that is $Dir(\theta_1, \theta_2; \alpha_1, \alpha_2) = Beta(\theta_1, \theta_2; \alpha_1, \alpha_2) = \frac{\Gamma(\alpha)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1}$, with $\alpha_{1,2} > 0$ and $\alpha = \alpha_1 + \alpha_2$.

[**a. Change of variable – Not graded**] The likelihood of a sample of size n from a Bernoulli(θ_1, θ_2) distribution is given by $L = P(\theta_1, \theta_2) = \theta_1^{n_1} \theta_2^{n_2}$. Change the variables θ_j to $\xi_j = \ln \theta_j$ for $j = 1 : m$ and express L as a function of $\xi_{1:m}$.

Now change the variables in the Beta density to $\xi_{1:m}$. Remember the change of variable formula in densities! Do you need to apply it to L as well?

Calculate the expression of the posterior in the variables $\xi_{1:m}$ and show that it is also a Dirichlet/Beta distribution. The parameters $\xi_{1:m}$ are called the *natural parameters* of the multinomial/Bernoulli distribution.

b. Let $m = 2$, $S = \{1, 2\}$, and $\mathcal{D} = \{1, 1, 2, 1, 1\}$. For the following 3 Dirichlet priors, give the numerical values of the *fictitious sample size* α, and the posterior parameters $\alpha'_{1,2}$.

[1] This means that any sequence of d digits from S is equally likely to appear. If the sequence starts with 0, we call it an integer with less than d digits.

[2] This data contains the population of towns and cities in the US.

- $Diri(\theta_1, \theta_2; 0.9, 0.1)$
- $Diri(\theta_1, \theta_2; 2, 3)$
- $Diri(\theta_1, \theta_2; 20, 20)$

c. Let $\theta = \theta_2$. For each of the 3 cases above, make a plot showing the prior, posterior as functions of θ, as well as the location of the ML estimate θ_2^{ML} on the θ axis.

d. Assume now the prior is uniform, that is $Diri(\theta_1, \theta_2; 1, 1)$. Show that the posterior of (θ_1, θ_2) is a Beta distribution and calculate its parameters for the data in **b**.

e. Same as **c.**, plot the prior, posterior and locaion of ML estimate for θ_2^{ML} for the uniform prior.