EECS 391

WH5

Qiwen Luo qxl216

1.

Q1. Bernoulli trials and bias beliefs

Recall the binomial distribution describing the likelihood of getting y heads for n flips:

$$p(y|\theta,n)=\binom{n}{y}\theta^{y}(1-\theta)^{n-y}$$

where θ is the probability of heads.

a) Using the fact

$$\int_0^1 p(y|\theta,n)d\theta=\frac{1}{1+n}$$

derive the posterior distribution for θ assuming a uniform prior. (10 P.)

b) Plot the likelihood for $n=4$ and $\theta=3/4$. Make sure your plot includes $y=0$. (10 P.)

c) Plot the posterior distribution of θ after *each* of the following coin flips: head, head, tail, head. You should have four plots total. (10 P.)

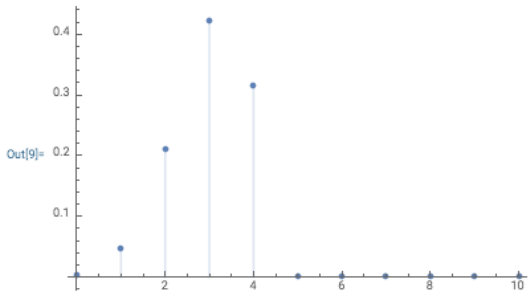
a) $P(\theta|y,n)=\frac{P(y|\theta,n)P(\theta|n)}{\int_0^1 P(y|\theta,n)P(\theta|n)d\theta}$

Since the prior is uniform, $P(\theta|n)$ is a constant.

Therefore, $P(\theta|y,n)=\frac{P(y|\theta,n)P(\theta|n)}{P(\theta|n)\int_0^1 P(y|\theta,n)d\theta}=\frac{\binom{n}{y}\theta^{y}(1-\theta)^{n-y}}{\frac{1}{1+n}}=(1+n)\binom{n}{y}\theta^{y}(1-\theta)^{n-y}$

b) $P(y|\theta,n)=\binom{n}{y}\theta^{y}(1-\theta)^{n-y}=\binom{4}{y}0.75^{y}0.25^{4-y}$

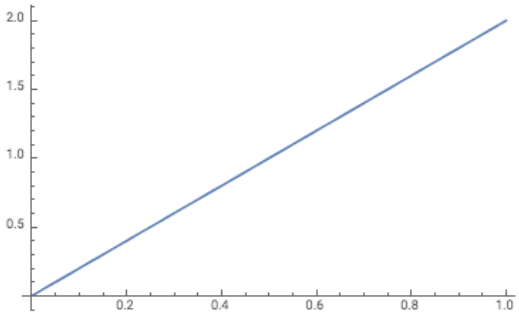
```
In[9]:= DiscretePlot[Binomial[4, y] * 0.75^y * 0.25^(4 - y), {y, 0, 10, 1}]
```



c) $n=1,y=1$

In[8]:=
```
Plot[Piecewise[{{2 * Binomial[1, 1] * θ ^ 1 * (1 - θ) ^ 0, θ < 1}, {2 * Binomial[1, 1] * θ ^ 1, θ == 1}}], {θ, 0, 1}]
```

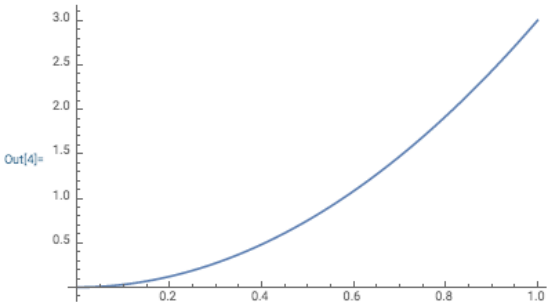
Out[8]=



$n = 2, y = 2$

In[4]:=
```
Plot[Piecewise[{{3 * Binomial[2, 2] * θ ^ 2 * (1 - θ) ^ 0, θ < 1}, {3 * Binomial[2, 2] * θ ^ 2, θ == 1}}], {θ, 0, 1}]
```

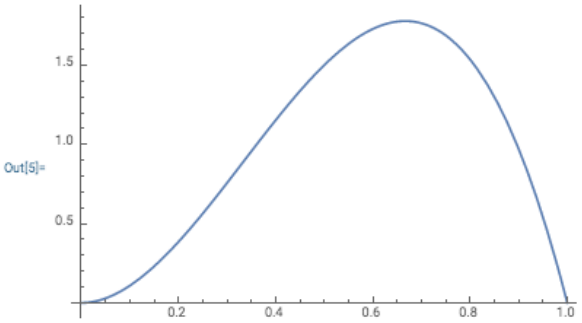
Out[4]=



$n = 3, y = 2$

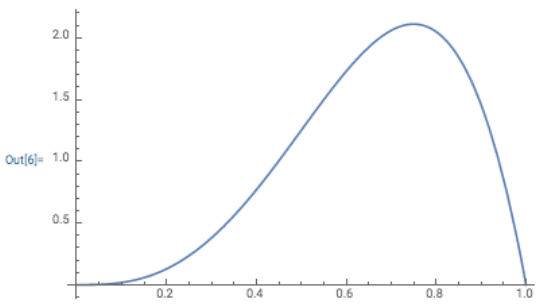
In[5]:=
```
Plot[{4 * Binomial[3, 2] * θ ^ 2 * (1 - θ) ^ 1}, {θ, 0, 1}]
```

Out[5]=



$n = 4, y = 3$

```
In[6]:= Plot[{5 * Binomial[4, 3] * θ^3 * (1 - θ)^1}, {θ, 0, 1}]
```



2.

Q2. After R&N 20.1 Bags O' Surprise

The data used for Figure 20.1 on page 804 can be viewed as being generated by h_5.

a) For each of the other four hypotheses, write code to generate a data set of length 100 and plot the corresponding graphs for $P(h_i|d_1,\ldots,d_N)$ and $P(D_{N+1} = \text{lime}|d_1,\ldots,d_N)$. The plots should follow the format of Figure 20.1. Comment on your results. (20 P.)

b) What is the mathematical expression for how many candies you need to unwrap before you are more 90% sure which type of bag you have? (10 P.)

c) Make a plot that illustrates the reduction in variabilty of the curves for the posterior probability for each type of bag by averaging each curve obtained from multiple datasets. (20 P.)

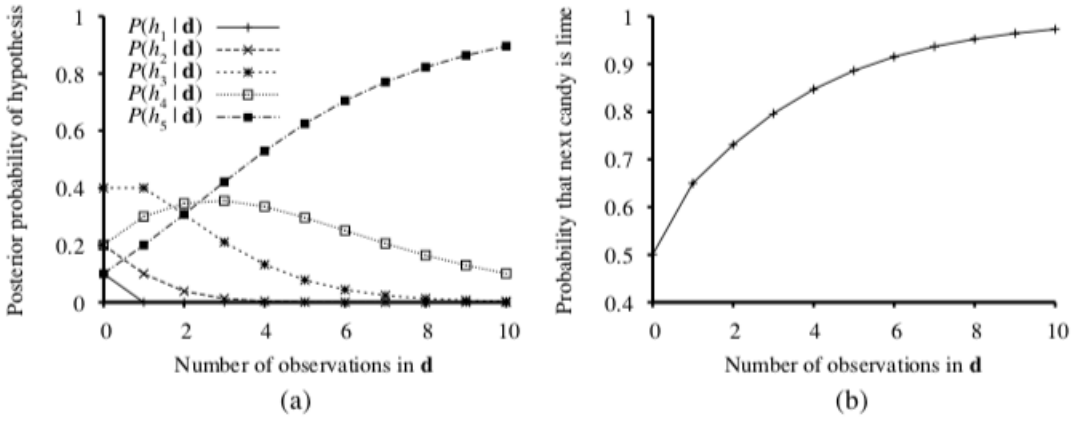


Figure 20.1 (a) Posterior probabilities $P(h_i \mid d_1, \ldots, d_N)$ from Equation (20.1). The number of observations N ranges from 1 to 10, and each observation is of a lime candy. (b) Bayesian prediction $P(d_{N+1} = lime \mid d_1, \ldots, d_N)$ from Equation (20.2).

a) For each dataset, I use

- $P(h|d) = \alpha P(d|h)P(h) = \alpha P(h) \prod_j P(d_j|h)$ to derive posterior probability.
- $P(d_{N+1} = lime|d_1 \ldots d_N) = \sum_i P(d_{N+1} = lime|h)P(h|d)$ to do the prediction.

For plots, please see the Mathematica file.

Red line indicates h1, blue line indicates h2, green line indicates h3, purple line indicates h4, yellow line indicates h5.

- For some graphs, only 4 lines are shown since red line and yellow line are totally the same.
- When evaluating graphs for $P(d_{N+1} = \text{lime} | d_1,\ldots,d_N)$ in Mathematica, if it is totally different from my graph in this document, just repeat evaluating the cell. (I don't know why, but the weird things happened several times.)

Comment on results

- For each dataset, the posterior probability is always consistent with the true hypothesis. And the posterior probability of any false hypothesis will finally goes to 0. For example, for dataset of hypothesis 1, $P(h_1|d) = 1$ and $P(h_2|d) = P(h_3|d) = P(h_4|d) = P(h_5|d) = 0$ for sufficiently large number of samples.
- For each dataset, the prediction of next sweet is always consistent with $P(d = \text{lime} | d_1,\ldots,d_N)$ for the true hypothesis. Namely, the Bayesian prediction finally agrees with the true hypothesis. For hypothesis 1, the probability approaches to 0. For hypothesis 2, the probability approaches to 0.25. For hypothesis 3, the probability approaches to 0.5. For hypothesis 4, the probability approaches to 0.75. For hypothesis 5, the probability approaches to 1.

b) More than 90% sure means that for a dataset, there is a bag type which is more than 90% to be the bag type of the dataset. Namely, max posterior probability of hypothesis is greater than 90%.

$$\max\{P(h_1|d), P(h_2|d), P(h_3|d), P(h_4|d), P(h_5|d)\} > 0.9$$

⇨ $max\{\prod_j P(d_j|h_1)P(h_1), \prod_j P(d_j|h_2)P(h_2), \prod_j P(d_j|h_3)P(h_3), \prod_j P(d_j|h_4)P(h_4), \prod_j P(d_j|h_5)P(h_5)\} > 0.9$

c) For plots, please see the Mathematica file.

Red line indicates h1, blue line indicates h2, green line indicates h3, purple line indicates h4, yellow line indicates h5.

3.

Q3. k-means Clustering

In k-means clustering, $\boldsymbol{\mu}_k$ is the vector mean of the k^{th} cluster. Assume the data vectors have I dimensions, so $\boldsymbol{\mu}_k$ is a (column) vector $[\mu_1,\ldots,\mu_I]_k^T$, where the symbol ^{T} indicates vector transpose.

a) Derive update rule for $\boldsymbol{\mu}_k$ using the objective function

$$D = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{n,k}\|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2$$

where $\boldsymbol{x}_n$ is the n^{th} data vector, $r_{n,k}$ is 1 if $\boldsymbol{x}_n$ is in the k^{th} class and 0 otherwise, and $\|\boldsymbol{x}\|^2 = \boldsymbol{x}^T\boldsymbol{x} = \sum_i x_i x_i = \sum_i x_i^2$. The update rule is derived by computing the gradient for each element of the k^{th} mean and solving for the value where the gradient is zero. Express your answer first in scalar form for $\mu_{k,i}$ and in vector form for $\boldsymbol{\mu}_k$. (20 P.)

a) $D = \sum_{n=1}^{N}\sum_{x=1}^{K} r_{nx}\sum_{y=1}^{I}(x_{n,y} - \mu_{x,y})^2$

$\Rightarrow \frac{\partial D}{\partial \mu_{k,i}} = \frac{\partial(\sum_{n=1}^{N} r_{nk}\sum_{y=1}^{I}(x_{n,y}-\mu_{k,y})^2)}{\partial \mu_{k,i}}$

$\Rightarrow \frac{\partial D}{\partial \mu_{k,i}} = \frac{\partial(\sum_{n=1}^{N} r_{nk}(x_{n,i}-\mu_{k,i})^2)}{\partial \mu_{k,i}}$

$\Rightarrow \frac{\partial D}{\partial \mu_{k,i}} = -2\sum_{n=1}^{N} r_{nk}\,(x_{n,i} - \mu_{k,i})$

$\frac{\partial D}{\partial \mu_{k,i}} = 0$

$\Rightarrow \mu_{k,i} = \frac{\sum_{n=1}^{N} r_{nk}x_{n,i}}{\sum_{n=1}^{N} r_{nk}}$

$\Rightarrow \mu_k = [\mu_{k,1},\ldots,\mu_{k,I}]^T = [\frac{\sum_{n=1}^{N} r_{nk}x_{n,1}}{\sum_{n=1}^{N} r_{nk}},\ldots,\frac{\sum_{n=1}^{N} r_{nk}x_{n,I}}{\sum_{n=1}^{N} r_{nk}}]^T$

$\Rightarrow \mu_k = \frac{\sum_{n=1}^{N} r_{nk}x_n}{\sum_{n=1}^{N} r_{nk}}$