Bayesian Data Analysis - Assignment 1

September 16, 2018

1. **Basic probability theory notation and terms**

   a) Explanations:

      – *probability* is the estimation of the possibility that an event will occur.

      – *probability mass* refers to the probability of samples on an interval; eg. the entire probability sample space is equal to 1.

      – *probability density* is the probability of mass divided by unit of the sample space.

      – *probability mass function (pmf)* gives the probabilities of the possible values for a discrete random variable.

      – *probability density function (pdf)* is a function of a continuous random variable, whose integral across an interval gives the probability that the value of the variable lies within the same interval.

      – *probability distribution* is a function of a discrete variable whose integral over any interval is the probability that the variate specified by it will lie within that interval.

      – *discrete probability distribution* refers to the probability of occurrence of each value of a discrete random variable.

      – *continuous probability distribution* refers to the probabilities of the possible values of a continuous random variable.

      – *cumulative distribution function (cdf)* calculates the cumulative probability for a given x-value and it can be used to determine the probability that a random observation that is taken from the sample space will be less than or equal to a certain value.

      – *likelihood* is a function of the parameters of a statistical model, given specific observed data.

   b) Answers to the questions:

      – *What is observation model?*
        Observation model refers to the expression that relates the parameters of the model to the observations.

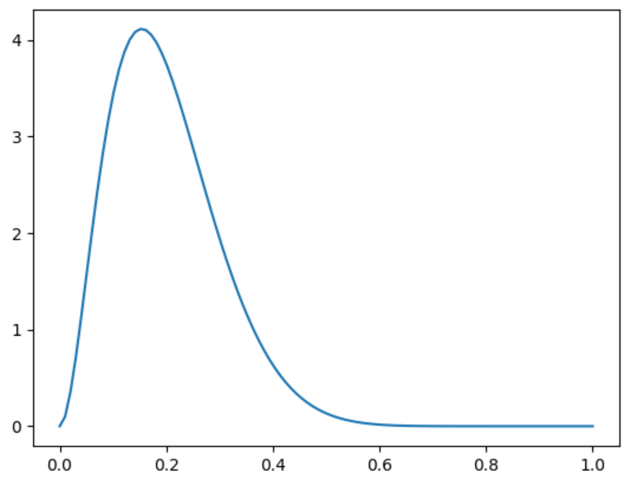
      – *What is statistical model?*
        Statistical modeling is a mathematically simplified way to approximate reality and optionally to make predictions from this approximation.

– *What is the difference between mass and density?*
Mass refers to the entire sample space, whereas density is the fraction from that sample space.

2. **Basic computer skills**

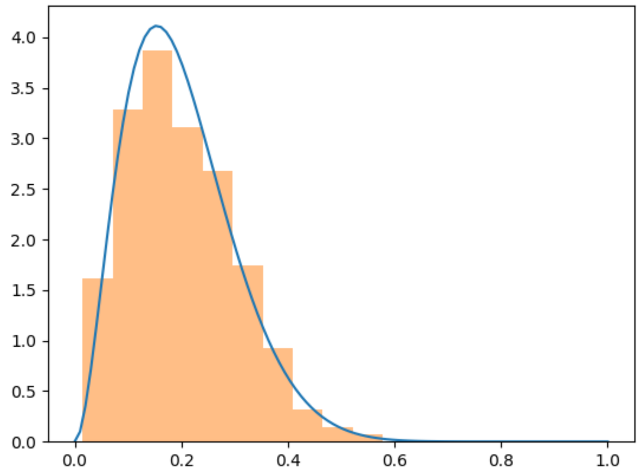
a) Plot the density function

```
from scipy import stats
import numpy
import matplotlib
import matplotlib.pyplot as plt

MEAN = 0.2
VARIANCE = 0.01

fig, axis = plt.subplots(1, 1)

alfa = MEAN * ( (MEAN * (1 - MEAN) / VARIANCE) - 1 )
beta = alfa * (1 - MEAN) / MEAN

x_range = numpy.linspace(0, 1, 100)
y_range = stats.beta.pdf(x_range, alfa, beta)

axis.plot(x_range, y_range)
```



b) Take a sample of 1000 random numbers and plot a histogram of the results

```
random_samples = stats.beta.rvs(alfa, beta, size=1000)
axis.hist(random_samples, density=True, alpha=0.5)
```



c) Compute the sample mean and variance from the drawn sample

```
sample_mean = numpy.mean(random_samples)
sample_variance = numpy.var(random_samples)
print('sample mean: ', sample_mean)
print('sample variance: ', sample_variance)
```

```
$ sample mean:  0.19997418955672838
$ sample variance:  0.01045597802573812
```

d) Estimate the central 95%-interval of the distribution

```
sample_percentile = numpy.percentile(random_samples, q=97.5)
print('sample central percentile 95%: ', sample_percentile)
```

```
$ sample central percentile 95%:  0.4188436088624379
```

3. **Bayes' theorem**: How would you advice a group of researchers who designed a new test for detecting lung cancer?

$P(Person_{randomly\ selected\ and\ has\ cancer}|Positive\ result) =$
$= \frac{P(P_{has\ cancer})*P(Cancer)}{P(Positive\ result)} = \frac{0.98*0.001}{0.98*0.001+0.04*0.999} = 0.023$

The result of test is not very satisfying for a randomly selected person who has cancer. Thus, I would tell the researchers to make their test prediction better.

4. **Bayes' theorem**: Find the probability of the selected ball being red and the box it came from.
$A = 2_{red}\ 5_{white};\ selected\ 40\%\ of\ the\ time$
$B = 4_{red}\ 1_{white};\ selected\ 10\%\ of\ the\ time$
$C = 1_{red}\ 3_{white};\ selected\ 50\%\ of\ the\ time$

$P(red) = 0.4 * \frac{2}{7} + 0.1 * \frac{4}{5} + 0.5 * \frac{1}{4} = 0.319$

$P(A|red) = \frac{P(red\ from\ A)*P(A)}{P(red)} = \frac{\frac{2}{7}*0.4}{0.319} = 0.358$

$P(B|red) = \frac{P(red\ from\ B)*P(B)}{P(red)} = \frac{\frac{4}{5}*0.1}{0.319} = 0.250$

$P(C|red) = \frac{P(red\ from\ C)*P(C)}{P(red)} = \frac{\frac{1}{4}*0.5}{0.319} = 0.391$

The probability of a red ball being picked is 31.9% and there is 39.1% chance that it came from Box C.

5. **Bayes' theorem**: What is the probability that Elvis was an identical twin?
   Let's assume that: S_{gt} is a notation of same gender twins, I_t stands for identical twins and F_t means fraternal twins. We need to find $P(I_t|S_{gt})$.

   $P(S_{gt}) = P(I_t) + P(F_t\ for\ the\ same\ gender) = \frac{1}{300} + 0.5 * \frac{1}{125} = 0.0073$

   $P(I_t|S_{gt}) = \frac{P(Elvis\ being\ twin)*P(I_t)}{S_{gt}} = \frac{1*\frac{1}{300}}{0.0073} = 0.45$

   There is 45% chance that Elvis was an identical twin.

Appendix A Source code

```python
from scipy import stats
import numpy
import matplotlib
matplotlib.use('TkAgg')
import matplotlib.pyplot as plt

MEAN = 0.2
VARIANCE = 0.01
fig, axis = plt.subplots(1, 1)

alfa = MEAN * ( (MEAN * (1 - MEAN) / VARIANCE) - 1 )
beta = alfa * (1 - MEAN) / MEAN

x_range = numpy.linspace(0, 1, 100)
y_range = stats.beta.pdf(x_range, alfa, beta)

# a) Plot the density function of Beta-distribution
axis.plot(x_range, y_range)
fig.savefig('./ex1/prob_distribution.png')

# b) Take a sample of 1000 random numbers and plot a histogram
random_samples = stats.beta.rvs(alfa, beta, size=1000)
axis.hist(random_samples, density=True, alpha=0.5)
fig.savefig('./ex1/prob_distribution_hist.png')

# c) Compute the sample mean and variance from the drawn sample
sample_mean = numpy.mean(random_samples)
sample_variance = numpy.var(random_samples)
print('sample mean: ', sample_mean)
print('sample variance: ', sample_variance)

```

```python
# d) Estimate the central 95%-interval from the drawn samples
sample_percentile = numpy.percentile(random_samples, q=97.5)
print('sample central percentile 95%: ', sample_percentile)
```